Title : A visual analytics approach for visualisation and knowledge discovery from time-varying personal life data

Name Farzad Parvinzamir

A Visual Analytics Approach for Visualisation and Knowledge Discovery from Time-varying Personal Life Data

by

Farzad Parvinzamir

in the

Centre for Visualisation and Data Analytics

A thesis submitted to the University of Bedfordshire, in fulfilment of the requirements for the degree of Doctor of Philosophy

January 2018

"The greatest value of a picture is when it forces us to notice what we never expected to see."

John Tukey

A Visual Analytics Approach for Visualisation and Knowledge Discovery from Time-varying Personal Life Data

Farzad Parvinzamir

Abstract

Today, the importance of big data from lifestyles and work activities has been the focus of much research. At the same time, advances in modern sensor technologies have enabled self-logging of a significant number of daily activities and movements. Lifestyle logging produces a wide variety of personal data along the lifespan of individuals, including locations, movements, travel distance, step counts and the like, and can be useful in many areas such as healthcare, personal life management, memory recall, and socialisation.

However, the amount of obtainable personal life logging data has enormously increased and stands in need of effective processing, analysis, and visualisation to provide hidden insights owing to the lack of semantic information (particularly in spatiotemporal data), complexity, large volume of trivial records, and absence of effective information visualisation on a large scale. Meanwhile, new technologies such as visual analytics have emerged with great potential in data mining and visualisation to overcome the challenges in handling such data and to support individuals in many aspects of their life.

Thus, this thesis contemplates the importance of scalability and conducts a comprehensive investigation into visual analytics and its impact on the process of knowledge discovery from the European Commission project MyHealthAvatar at the Centre for Visualisation and Data Analytics by actively involving individuals in order to establish a credible reasoning and effectual interactive visualisation of such multivariate data with particular focus on lifestyle and personal events.

To this end, this work widely reviews the foremost existing work on data mining (with the particular focus on semantic enrichment and ranking), data visualisation (of time-oriented, personal, and spatiotemporal data), and methodical evaluations of such approaches. Subsequently, a novel automated place annotation is introduced with multi-level probabilistic latent semantic analysis to automatically attach relevant information to the collected personal spatiotemporal data with low or no semantic information in order to address the inadequate information, which is essential for the process of knowledge discovery. Correspondingly, a multi-significance event ranking model is introduced by involving a number of factors as well as individuals' preferences, which can influence the result within the process of analysis towards credible and high-quality knowledge discovery. The data mining models are assessed in terms of accurateness and performance. The results showed that both models are highly capable of enriching the raw data and providing significant events based on user preferences.

An interactive visualisation is also designed and implemented including a set of novel visual components significantly based upon human perception and attentiveness to

visualise the extracted knowledge. Each visual component is evaluated iteratively based on usability and perceptibility in order to enhance the visualisation towards reaching the goal of this thesis.

Lastly, three integrated visual analytics tools (platforms) are designed and implemented in order to demonstrate how the data mining models and interactive visualisation can be exploited to support different aspects of personal life, such as lifestyle, life pattern, and memory recall (reminiscence). The result of the evaluation for the three integrated visual analytics tools showed that this visual analytics approach can deliver a remarkable experience in gaining knowledge and supporting the users' life in certain aspects.

Contents

List of Figures

List of Tables

List of Algorithms

List of Snippets

List of Publications

[1] Deng, Z., Zhao, Y., Parvinzamir, F., Zhao, X., Wei, H., Liu, M., Zhang, X., Dong, F., Liu, E. and Clapworthy, G. [2016], *MyHealthAvatar: A Lifetime Visual Analytics Companion for Citizen Well-being*, Springer International Publishing, Cham, pp. 345–356.

[2] Parvinzamir, F., Zhang, X. and Dong, F. [2017], 'An automated place annotation with multi-level probabilistic latent semantic analysis to support knowledge discovery of personal spatio-temporal data', *Intelligent Systems and Technology (TIST)* – submitted.

[3] Parvinzamir, F., Zhao, Y., Deng, Z., Zhao, X., Ersotelos, N., Dong, F., Liu, E. and Clapworthy, G. [2015], MyHealthAvatar: A case study of web-based interactive visual analytics of lifestyle data, *in* '2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing', pp. 2335–2339.

[4] Parvinzamir, F., Zhao, Y. and Dong, F. [2017], 'Myevents: A personal visual analytics approach for mining key events in support of personal reminiscence', *Computer Graphics Forum* – submitted.

[5] Zhao, Y., Parvinzamir, F., Deng, Z., Wei, H., Zhao, X., Liu, E., Dong, F., Clapworthy, G., Lukoševičius, A., Marozas, V. and Kaldoudi, E. [2016], 'MyHealthAvatar and CARRE: case studies of interactive visualisation for internet-enabled sensor-assisted health monitoring and risk analysis', *IET Networks* 5(5), 114–121.

[6] Zhao, Y., Parvinzamir, F., Wei, H., Liu, E., Deng, Z., Dong, F., Third, A., Lukoševičius, A., Marozas, V., Kaldoudi, E. and Clapworthy, G. [2016], *Visual Analytics for Health Monitoring and Risk Management in CARRE*, Springer International Publishing, Cham, pp. 380–391.

[7] Zhao, Y., Parvinzamir, F., Wilson, S., Wei, H., Deng, Z., Portokallidis, N., Third, A., Drosatos, G., Liu, E., Dong, F., Marozas, V., Lukoševičius, A., Kaldoudi, E. and Clapworthy, G. [2017], 'Integrated visualisation of wearable sensor data and risk models for individualised health monitoring and risk assessment to promote patient empowerment', *Journal of Visualization* 20(2), 405–413.

[8] Zhao, Y., Parvinzamir, F., Zhao, X., Deng, Z., Dong, F., Ersotelos, N. and Clapworthy, G. [2015], Web-based visual analytics of lifestyle data in myhealthavatar, *in* 'Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare', MOBIHEALTH'15, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, pp. 139–143.

Declaration

I, Farzad Parvinzamir, declare that this thesis is my own unaided work. It is being submitted for the degree of Doctor of Philosophy (PhD) at the University of Bedfordshire.

It has not been submitted before for any degree or examination in any other University.

Name of candidate:
Farzad Parvinzamir

Signature:

Date:

Acknowledgements

I would like to express my sincere gratitude to my Director of Studies, Professor Feng Dong, for his endless support, patience, and constructive advice during my PhD research. His methodical guidance has allowed me to complete this challenging journey.

I would also like to thank my second supervisor, Dr Enjie Liu, for her support and encouragement. I should also thank my friend and colleague Dr Youbin Zhao for many technical and valuable discussions at different stages of my research work. I would also like to thank my friends and colleagues Dr Nigel Mcfarlane and Dr Peter Norrington for all the practical discussions and proofreading my thesis.

I would also like to thank Caroline Aird from the Research Graduate School for her kind assistance with the official progress of the PhD programme at the University.

Above all, I would like to thank my parents and particularly my partner Eirini for their limitless support and love in all aspects over the recent years, which has enabled me to stay strong and complete my research work successfully.

To my parents Ali & Fereshteh

and my other half
Eirini

CHAPTER 1

Introduction

Today, the importance of big data from lifestyles activities has been the focus of much research with the aim of enhancing quality of life through capturing and analysing users' everyday activities. Concurrently, new technologies have emerged with great potential to overcome the challenges in systematically acquiring and handling such data, and to help individuals empower their health and lifestyle [6, 73, 118].

Contemporarily, the amount of the practical personal data related to an individual's fitness, movements, health, and lifestyle have become extensively available at a large scale as a result of widely available sensors in mobile and wearable technology. The data acquired is entirely affiliated with individual lives and contain valuable details for exploring significant and engaging behavioural information in space and time [96, 119, 242]. In order to gain knowledge and empower an individual's life, this data, given its wide availability, needs to be reachable, comprehensible, and explicable [96]. Nevertheless, given the data size and complexity, gaining insights has until now constituted inevitable challenges which results in the need for novel personal visual analytics to support effective utilisation of such data and related information.

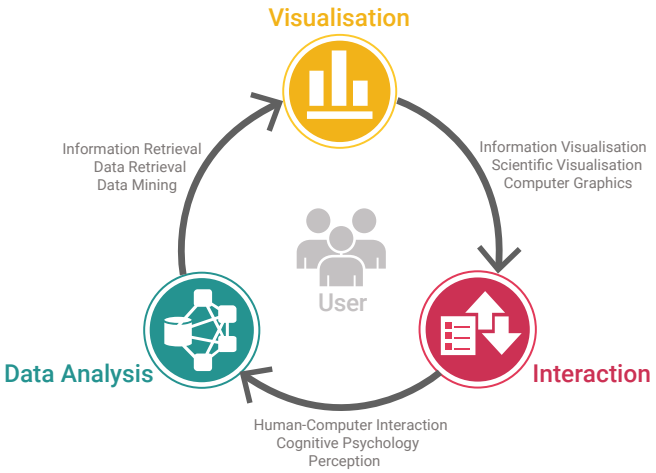


FIGURE 1.1: The seminal Visual Analytics framework by Keim et al. [113]

Personal visual analytics implies the use of analytical reasoning together with visual representation in making sense of personal life data [96]. In principle, visual analytics embodies: 1) insightful data mining to address scalability and achieve intelligible understanding, 2) human-computer interaction to explore the visual representation further, and 3) visualisation to encode the acquired information into sensible graphical content (Figure 1.1). An effective visual analytic approach is able to take advantage of its arithmetical models, information visualisation, and human perception via expressive and flexible interactions to deliver a best practice [68]. However, there has only been a small amount of visual analytics research on personal life logging data owing to several challenges, such as inefficient visualisation caused by over-plotting and issues of accommodating data from compound sources that need to be considered in order to provide robust interpretation of the data [16, 190, 203].

This thesis strives to conduct a deep investigation into the fundamental visual analytics approaches in order to set out a novel knowledge discovery and effectual information visualisation of multivariate and temporal personal life data based on the research project at the Centre for Visualisation and Data Analytics – MyHealthAvatar – funded by the European Commission. The thesis's research, correspondingly, contemplates the importance of scalability, rational data mining, human perception, interaction, and effectiveness within its investigation to provide

a user-friendly and walk-up interface and minimise the learning procedure.

This chapter provides a brief background on visual analytics, the motivation that drives this work, the research aim and objectives, the scope and limitations, the main research questions, the key contributions, the definition of a number of terms used in this work, and the outline of this thesis.

1.1 Background

Visual analytics, in the main, focuses on turning information overload into an opportunity and providing a transparent data processing [73, 113]. This technology points towards incorporating human perception plus analysis capabilities with automatic data exploration to increase the quality of knowledge discovery at different scales, strengthen the data exploration, and provide comprehensible graphical presentation (Figure 1.2). There are a number of key features that visual analytics approaches ought to make allowances for, such as:

- enabling real-time data analysis
- supporting real-time creation of dynamic and interactive presentation
- versatile interaction options
- holding data in-memory for the processes of data mining and visualisation.

Furthermore, much research investigates scalability within their visual analytics techniques to deal with large-scale datasets. Scalable techniques involve the ability to analyse and visualise large-scale data with the properties of growing variety, velocity, volume, and veracity within the display limitation [8, 10, 13, 48, 185]. Many visual analytics techniques have been developed to address different real-world problems by considering the type of dataset, the given tasks, and the users [30, 48, 86, 110, 235]. Typically, these techniques include three main components of visual analytics: data mining, data visualisation, and human computer interaction.

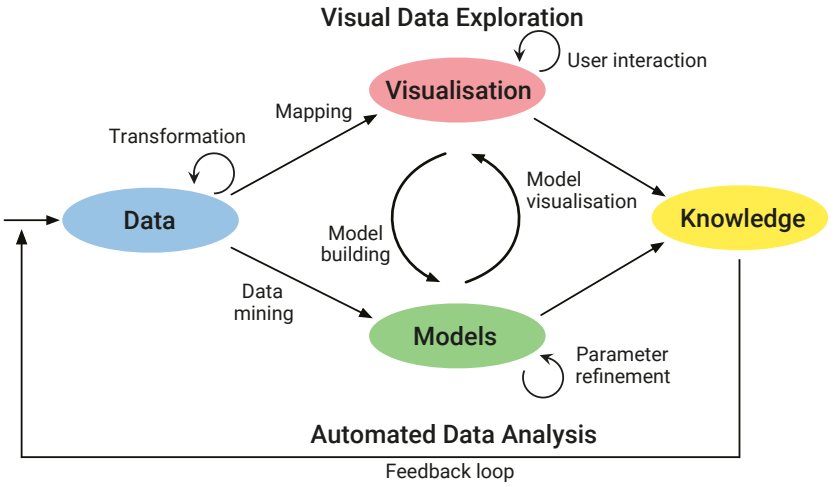


FIGURE 1.2: Visual Analytics process for knowledge discovery by Kohlhammer et al. [118, pg. 118]

Each component performs a set of functions that can be exploited interchangeably by the other components. In the following paragraphs a brief definition of each component is provided.

1.1.1 Data mining

Data mining is a fundamental key to reduce the complexity of large-scale data, significantly improve system performance, and extract beneficial knowledge. Data mining incorporates a wide range of processing and analytical approaches, such as filtering, dimension reduction, feature extraction, clustering, semantic enrichment, and significance ranking. These methods can be applied to the raw data – personal life logging data in this thesis – in order to eliminate the errors, reducing the size and complexity, attaching additional information, and uncover the important points accordingly towards providing a more concrete environment for scalable data visualisation [98, 102, 111, 112, 152]. There are a number of commonly used data mining methods in visual analytics, particularly, within the field of personal data such as:

- **Temporal data abstraction**: reduces unneeded value ranges of multivariate data, which are excessively complex and large to manage, to make data much easier to manipulate [21, 55, 115].
- **Dimension reduction**: decreases the number of dimensions by focusing on major trends to obtain a feasible analysis of massive, high-dimensional data (e.g. Principal Components Analysis (PCA), Kernel PCA, Multidimensional Scaling (MDS) [43, 48, 54, 166].
- **Clustering algorithms** (e.g. K-means and Density-based algorithm): partitions the data into the different categories based on their similarities or relations by employing efficient heuristic algorithms [26, 54, 99].
- **Semantic enrichment**: aims at allocating additional contextual information to the unlabelled data points (e.g location coordinates) in order to turn such data into a rich asset that can be used in the process of knowledge discovery in personal data visualisation [119, 120, 156].
- **significance ranking**: supports the process of visual analytics by computing the significance score for data points based on the variety of parameters such as frequency in order to address scalability of large-scale datasets with a large amount of trivial detail that cannot be visualised as a whole due to visual clutter and limited visual properties [13, 25, 60, 85].

1.1.2 Data visualisation

Data visualisation, in principle, is the process of encoding the data in a graphical format to show the extracted information meaningfully and allow for gaining knowledge. It amplifies human understanding through using the capability of interaction and different visual encoding such as interactive depiction, sensory representation, and typically, visual illustration [25, 107, 203, 229]. The successful visualisation leads to identifying patterns and potent knowledge discovery by considering scalability and employing suitable visual properties which notably increase the effectiveness.

However, visualising the information extracted from large-scale data is consistently challenging due to limitation of the display's size and lack of effective visual components. The most common factors that can partially cover the visual representation are [8, 229]:

- **Levels of detail:** render the visualisation to an optimum scale by modifying the resolution of visualised objects, which are not identical on the display;
- **Call outs:** use a set of annotations in a limited display to show the global context together with the sub-scale detail;
- **Visual properties:** employ effective visual encodings within the designed components in visualisation to facilitate the process of understanding.

Numerous data visualisation techniques are introduced in this regard for different purposes. This thesis strives to present related techniques within the literature review and then demonstrates its own visualisation method by using several inventive visual components.

1.1.3 Human interaction

Interaction is essential in visualisation as it can let users interact with the visual result of the analytical methods to hand. In general, human interaction is denoted as sense-making that allows for exploring possible connections, gaining meaningful insight, and investigating hypotheses [161, 189, 203, 210, 219].

Human interaction includes two fundamental stages, namely, foraging and synthesis. Foraging is related to filtering and collecting relevant but interesting information by an individual, whilst synthesis is associated with creating and testing hypotheses about the foraged information. Normally, foraging deals with computational processing, whereas synthesis hinges on human perception for setting out relationships among the collected information. Therefore, this requires

a rigorous integration between the data mining model and visualisation to enable the user to carry out the interaction [107]. In addition, human interaction can inherently reduce uncertainty; facilitate the data navigation and comparison process; and hence amplify the process of gaining knowledge [68].

1.2 Motivation of this Work

The motivation of this research has three underlying factors:

- the growth of mobile and wearable technology adoption;
- the availability of valuable personal data; and
- the lack of well-designed visual analytics approaches for use in the domain of personal daily life.

According to the Fitbit official report [1], the number of commercial devices that capture, monitor, and store personal life activities has been dramatically increased. Fitbit – manufacturing a wide range of wearables – sold 21.4 million devices during 2015 compared to 9.2 million in 2014. Another well-known company is Withings [2] with a high selling rate of wearables and mobile monitoring devices. Withings provides more than 10 different devices that record individuals' activities and vital signs, e.g. heart rate, weight, blood pressure.

Furthermore, smartphone applications such as *Moves*[3], *Map My Run*[4], and *Endomondo*[5] have become widely popular for tracking and logging activity. These applications are able to accurately recognise user activity and log relevant details such as date, time, geographical coordinates, activity type, and calories automatically or on demand. What is more, the recorded data can be obtained by third

[1] https://investor.fitbit.com/press/press-releases/press-release-details/2016
[2] https://www.withings.com/uk/en/
[3] https://moves-app.com/
[4] http://www.mapmyrun.com/
[5] https://www.endomondo.com/

parties by using provided APIs. This opens up a path for research programmes on acquiring raw data and investigating real-world problems.

The amount of personal daily data has been equally elevated owing to the growing number of wearable and smartphone applications and their adoption. Thus, this results in more demand from individuals to explore and improve their lifestyle. To this end, manufacturers provide services such as standalone desktop software, web services, mobile applications, and the like to make sense of the recorded data. However, they do not use any means of visual analytics including data mining approaches in order to allow individuals to explore and gain compelling knowledge interactively. Instead, they depend on manual annotation from users and use basic visual data exploration for a limited part of this large-scale data.

Meanwhile, a small amount of research makes use of these personal data to deliver beneficial visual analytics solutions to individuals. Although the provided solutions are valuable, most of them are complex, coming with a series of weaknesses (in terms of scalability and effectiveness) and cannot be generalised. This makes the process of gaining meaningful knowledge convoluted and results in decreasing user engagement. The complexity makes the approaches burdensome to use, which leads to users' refusal. Obstacles such as scalability and effectiveness prevent delivering a meaningful understanding towards knowledge discovery while lack of generalisation results in an incoherent outcome. These challenges mainly hinge on the fact that individuals have different preferences while exploring such data. This has been manifested by a large amount of research showing that lack of consideration of the human during design and implementation leads to an ineffective visual analytics approach. Thus, visual analytics approaches require determining the factors that comprise meaningful and informative outcomes based upon users' understanding and preference, with the scalability in mind. Hence, in a nutshell, visual analytics needs to consider the following factors:

- Focusing on users' viewpoints to improve end-user understanding;
- Considering user preferences such as motivation, priorities, expectations, and goals within its underlying data mining components;

- Providing perceptible, coherent, and easy-to-understand results;
- Involving users in the process of designing the visual components and interface (user-centred design).

By considering the aforementioned facts, this thesis endeavours to introduce an effective approach to explore and gain hidden insights from such data via:

- considering users' sentiments in the process of design and implementation;
- rational data processing (e.g. data cleaning, filtering, etc.);
- the strength of data mining to enrich unlabelled data and determine significant points to deal with large personal life logging data with numerous trivial points;
- an uncomplicated interface to explore and grasp meaningful understanding of data for non-expert users with no need for additional knowledge in programming or visual analytics;
- robust algorithms to achieve real-time analysis;
- innovative visual exploration and visual properties to encode the extracted information;
- concrete human-computer interaction to allow individuals to find and narrow down their own answers by considering users' preferences.

1.3 Research Questions

This work introduces a visual analytics approach including three main methods to automatically annotate unknown GPS data points, determine the significant GPS data points followed by extracting information, and present the information interactively with an effective aesthetic for personal daily life data towards gaining

meaningful knowledge. The research tries to answer the following questions during an extensive investigation.

Q1. What is the impact of semantic enrichment in personal spatiotemporal data and how does it contribute to mining meaningful knowledge?

Q2. How important is significant event ranking to the visual analytics of personal data and what is the impact of effective event ranking in knowledge discovery? Also, what are the most influential factors on the result of event ranking in persuasive knowledge discovery?

Q3. How important is the visual representation of extracted knowledge in personal visual analytics? And how effective does eloquent visualisation support the process of knowledge discovery?

1.4 Aim and Objectives

This research aims at in-depth investigations into innovative personal visual analytics approaches to support health living and personal life management such as life pattern discovery and personal memory recall through interactive data exploration and knowledge discovery. The thesis introduces a novel automated latent semantic enrichment, an inventive significance ranking model and highly interactive visualisation techniques. The specific objectives are as follow, to:

OBJ1. Review and identify the state-of-the-art research in data-mining, time-oriented, and spatiotemporal visual analytics with a particular focus on information extraction, semantic enrichment, event ranking and interactive visualisation of cumulative personal data for the purpose of identifying technical gaps and shortfalls.

OBJ2. Identify adequate evaluation methods and metrics within the visual analytics context.

OBJ3. Address scalability and data quality for the data mining algorithms through working on data preprocessing such as filtering, aggregation, and density-based clustering.

OBJ4. Design, implement, and evaluate an innovative semantic enrichment model to address the problem of lack of semantic information within personal data and hence enhance the quality of the extracted knowledge.

OBJ5. Design, implement, and evaluate a multi-significant event ranking model to reveal significant events from a large number of events in everyday life by involving the user preferences during the process.

OBJ6. Identify the best practice and key requirements for designing an effective information visualisation including visual encodings and interface.

OBJ7. Design and implement an advanced visualisation technique and visual components coupled with human-computer interaction to present the extracted knowledge from the daily life logging data.

OBJ8. Conduct a number of iterative evaluations within the process of visual component design and implementation by involving users in the process and hence enhance usability together with productivity.

OBJ9. Design and develop integrated visual analytics tools to demonstrate and allow evaluation of the proposed personal visual analytics approach for accuracy, usability, user satisfaction, and effectiveness.

1.5 Scope and Limitation

The scope of this research covers constructive knowledge discovery and effectual visualisation of personal daily life data by using a robust analytical reasoning, inventive information visualisation, and human-computer interaction as a web-based solution. This thesis strives to demonstrate the effectiveness of the proposed visual analytics approach by implementing a number of integrated visual analytics

tools towards supporting certain aspects of life such as daily lifestyle, life pattern, and reminiscence. However, it does not investigate nor suggest any medical terms in any form towards improving lifestyle or memory recall as producing any form of medical result is beyond the scope of this research. Instead, this work may be utilised by other research programmes to facilitate the investigation of such terms. In addition, this work does not investigate any form of data repository and its impact on the proposed visual analytics approach. Of course, in the future, using a discrete database including a novel data retrieval can be considered by other work.

This research comprises a limited number of individuals with recorded life logging data. Acquiring public personal life logging data is not possible owing to the confidentiality of such data. As a result, this places a limit on evaluating the proposed integrated visual analytics tools in Chapter 7 – particularly MyEvents. This is partially addressed during the process of design and implantation by creating synthetic data propagated from real personal life logging data. However, to evaluate an integrated tool such as MyEvents, which implies personal user preferences, only the participants with trackers with at least one year's data could be recruited.

The process of automated semantic enrichment in this research uses a free Point of Interest (POI) service to retrieve textual information. Although here this uses the most flexible service, the limitation in using such services prevents the process from being used at a large scale unless acquiring a business API account. Also, the popularity of POI services vary within different countries, which may result in varied quality of information (In our case, these services are mostly used to get place information within Europe). Thus, the place annotation outcome may be affected by different geographical regions and the level of information held by the POI service provider.

1.6 Contributions

This research decisively contributes to the process of knowledge discovery and interactive information visualisation for personal data. The main contributions are portrayed as follows:

C1. An extensive survey of the foremost existing visual analytics literature about data mining and interactive visualisation in various fields, such as time-oriented and spatiotemporal, with a particular focus on personal data visual analytics.

C2. A novel multi-level probabilistic latent semantic analysis model to automatically determine and attach a prominent level of semantic information to the users unidentified trajectory data assisted by contextual information (e.g. place of interest), historical location and prior-knowledge.

C3. A novel multi-significance event ranking model to identify significant events in a personal history according to user preferences through ranking, allowing the user to efficiently identify key events over a selected period of time based on their personal preference settings.

C4. A robust visual analytics pipeline that incorporates a novel data mining and interactive visualisation to exhaustively support the process of knowledge discovery and exploration within the personal life logging data.

C5. An interactive information visualisation with inventive visual properties to present personal daily life, such as personal events, along with extracted knowledge and a set of heterogeneous information, involving users in the process of design and extensive evaluation.

C6. Demonstrate and evaluate the proposed visual analytics approaches including the data mining models and the interactive visualisation techniques in integrated platforms.

C7. The ActivityTimeline platform to visualise lifestyle data collected from tracking devices and mobile applications to improve an individual's lifestyle by offering a set of interactive visualisation techniques.

C8. The LifeTracker platform to support effective exploration of individual lifestyle and patterns from a sequence of self-logging data over years through integration of a range of analytics and visualisation techniques.

C9. The MyEvent platform to support reminiscence with an integrated environment of data analytics, visualisation and human-computer interaction, featuring new data mining techniques (semantic enrichment and significance ranking) to support user involvement and novel visual presentations.

1.7 Definitions

A number of general terms are defined within this work. These terms are utilised in different parts of the research, such as the literature review, data mining, knowledge discovery, and information visualisation. A brief definition of each term is outlined in this section.

- **Personal data**: as partially mentioned by Huang et al. [96], in all respects, this is related to the individual and can be used solely for personal exploration along with knowledge discovery by a human with different skills, preferences, and experience.
- **Event**: an activity that takes place at a physical location at a particular date and time, for example, shopping at a local shopping centre on 7th July 2016, studying at a local library on 12th April 2016, etc. The event can be

repeated by revisiting the same place, e.g. 10 revisits to the local library. Notably, here the concept of event and activity are interchanged by assuming that the related activity takes place at one place, for example, study in a library or eating in a restaurant. Such an assumption, to a certain extent, is reasonable due to lack of information about possible activities within one place.

- **Movement**: a physical action that connects two or more events comprising a number of track points as a path within a certain date and time, for instance, walking from home to work, running, transportation, etc. The movement can be also repeated with the same or different path way.

- **Point of Interest (POI)**: complementary information related to various places and attractions that are considered as interesting points by users, such as restaurants, parks, malls, and the like.

- **Category**: the classification of the places and attractions according to their nature. Places, POIs, and any attractions are classified into different categories, such as food, shops and services, university and college, etc. This classification considerably supports the data analysis and visualisation of the events.

- **Geographical Coordinate**: specifies every location on Earth by a set of numbers, namely, latitude and longitude. The former represents a vertical line while the latter corresponds to a horizontal line. The tracking devices that are equipped with a GPS sensor are able to collect such data.

- **Physical Activity**: four types of movement with GPS track point information comprising start and end times that are automatically recognised by the application, namely, walking, running, cycling, and transport.

1.8 Thesis Outline

This thesis is organised into eight chapters. The thesis milestones include the relation between each chapter together with the research questions, objectives and contributions as shown in Figure 1.3.

Chapter 2 is divided into three parts: data mining and knowledge discovery, data visualisation, and evaluation within visual analytics which addresses [**OBJ1, OBJ2**]. In the first part, related work in semantic enrichment and place annotation is investigated as well as event detection and ranking to determine the current models and algorithms. Within the second part, the theory of visual information is discussed and subsequently the foremost existing work is reviewed on visualisation of time-oriented data, personal data, and spatiotemporal data, respectively. In the evaluation section, the existing and credible evaluation methods are reviewed within the visual analytics context to identify the best practice for evaluating this approach. And lastly, the chapter concludes by summarising the key points derived from the literature.

Chapter 3 describes the methodology undertaken for conducting this research in detail. It includes the underlying structure of data acquisition, literature review, design and prototyping, experiments, and evaluation.

In Chapter 4, the automated place annotation is introduced with multi-level probabilistic latent semantic analysis. This chapter explicitly describes the process of design and implementation of the model in depth. Moreover, the empirical evaluation results of the model are unfolded in this chapter to exhibit accuracy and efficiency [**OBJ3, OBJ4**].

Chapter 5 depicts the process of designing and implementing the novel multi-significance ranking model and its algorithms in depth. The process of evaluation and results obtained for this model are also illustrated in this chapter to show accuracy and efficiency, respectively [**OBJ3, OBJ5**].

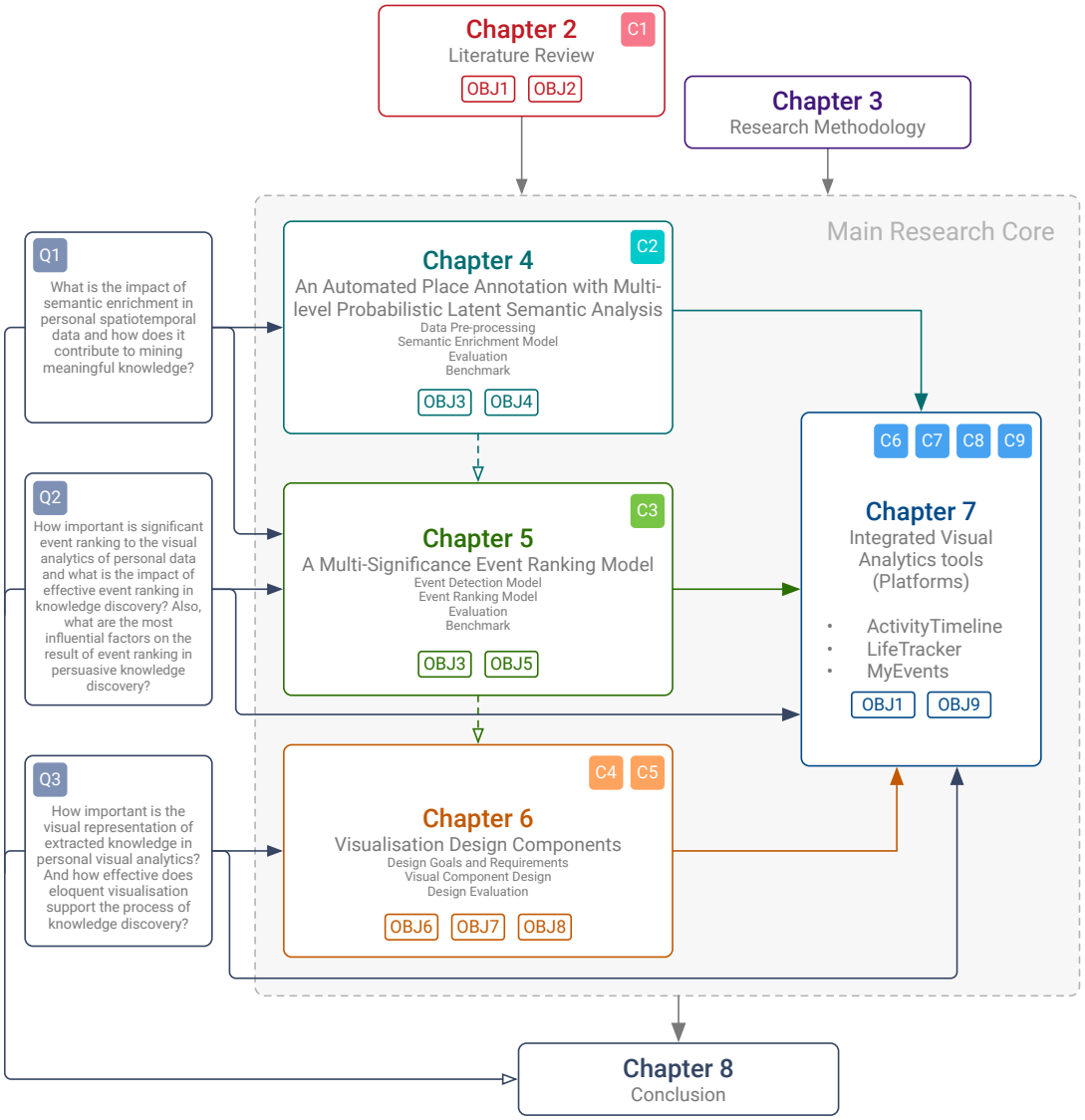


FIGURE 1.3: The structure of this thesis. The research questions are shown with Q, objectives with OBJ, and contribution with C.

The visualisation technique and its design principles are discussed in Chapter 6 to show how the extracted knowledge from the data mining model are encoded. A handful of visual properties are introduced together with visual components that can encode the knowledge and information to the highest standard with considering preattentive processing and the human perception. These components are evaluated in terms of design and usability and the results are cast accordingly [**OBJ6, OBJ7, OBJ8**].

Chapter 7 presents three practical platforms with different purposes that embody the data mining models along with interactive visualisation to show different perspectives of the visual analytics approach. These platform experiments are conducted sequentially during the progress of this research and the lessons learned are used in the succeeding trials. Platforms are described thoroughly within a number of sections, including the design intention, prototype, evaluation result, and discussion. The evaluation process involves task completion, usability, user interface satisfaction, and effectiveness and these are detailed independently within each integrated visual analytics tool (platform) [**OBJ1, OBJ9**].

And lastly, the thesis concludes with its major outcomes in Chapter 8 by summarising the thesis, highlighting the contributions, and identifying the upcoming opportunities from the current work, outlining them as future research directions.

CHAPTER 2

Literature Review

This chapter provides a brief introduction to visual analytics and related visual information theory followed by the literature review of relevant work within the context of personal data for this thesis.

The aim of this thesis is to introduce a novel approach for supporting knowledge discovery from personal life logging data by establishing two data mining models (place annotation and significant event ranking) and interactive visual exploration. Thus, this chapter systematically reviews existing methods in three parts, namely, data and knowledge mining, data visualisation techniques, and visual analytics evaluation. In data and knowledge mining, the related methods to place annotation and significant ranking are demonstrated. In data visualisation techniques, firstly visualisation information theory is described and then visualisation methods for time-oriented data, personal data, and spatiotemporal data (particularly event-based data), are reviewed. And lastly, the closely connected evaluation methods that can be employed to assess the proposed approach in this thesis are depicted.

2.1 Introduction

There is a huge demand for visual analytics in a great number of domains for innovative knowledge discovery as it is very capable of exploring and displaying the burgeoning amount of valuable data. Nowadays, a volume of varied complex data, which are time-varying, heterogeneous and multi-dimensional, has greatly increased to discover new insights and also make appropriate decisions in time-critical situations by means of visual analytics approaches [68, 73, 204]. Visual analytics, in the main, aims at turning information overload into an opportunity as well as providing transparent data processing [108, 112] and address the key challenges within this area by exploiting three main areas, namely, data-mining, information retrieval, and human interaction methods [107, 110, 111]. In fact, incorporating human perception, allowing interaction and analysis capabilities together with automatic data exploration, results in a higher quality of knowledge discovery in visual analytics procedures [151, 198].

In addition, visual analytics has a number of tuning tools, such as filtering, clustering, aggregation, noise or dimension reduction with well-established algorithms, and interactive design to reinforce insight discovery, exploring, monitoring, and comprehensible demonstrations to domain experts and the general public [151, 165, 207].

The first visual analytics mantra was introduced by Shneiderman [191, pg. 337]:

> "Overview first, then zoom and filter, and finally, details on demand."

However, the mantra was reshaped later to meet prevalent visual analytics criteria by Keim et al. [112, pg. 16]:

> "Analyse first, show the important, zoom/filter and analyse further, details on demand."

The latest mantra, to a great extent, focuses on an assortment of interactive visual interfaces and algorithmic data analyses compared to the information visualisation mantra. The original framework, correspondingly, was introduced by Kohlhammer et al. [118] to describe the process of visual analysis. The process begins with data retrieval and transformation, then an automatic data analysis approach such as data mining is employed to evaluate models and extract the information. Next, information visualisation implies that users collaborate with the visual interface towards exploring and analysing data to a greater extent.

The next section describes the principles of visualisation which have been taken into consideration in this work in order to provide a compelling data visualisation.

2.2 Data and Knowledge Mining

Geospatial data have become important as many visual analytics approaches require finding spatial patterns and relationships between the data points. Nevertheless, the key challenge in this field is the great amount of trivial data points and the lack of high level semantic information [173, 195, 236]. This can lead to an impractical encoding and also prevent the visual analytics approach from discovering heuristic knowledge. These challenges are already identified and dealt with accordingly by annotating and calculating the significant data points. Therefore, the pertinent literature is reviewed for both challenges in this section.

2.2.1 Place annotation

There are various works that attempt automated semantic enrichment for trajectory data via using different data mining approaches or supervised machine learning frameworks with a training dataset [9, 45, 83, 119, 120, 156, 172, 195, 236, 243]. These methods, typically, include determining the stops, movements, sequences, and episodes before labelling the points. One of the key challenges in this field is to identify unknown points and their characteristics, which have been extracted and

classified into different episodes. In this section, the relevant works are reviewed on semantic enrichment that use POI services to facilitate the process of enriching unknown places.

According to the literature review, semantic enrichment work can be classified into two groups. The first group deals with the preprocessing of trajectories in order to transform these complex data into a consequential series of stops, movements, and episodes and to extract knowledge by using an ontology or the like. The second group attempts to identify the unknown stops along with movements and then attach additional information by means of POI services. In the following, both groups are discussed in detail.

Group 1: Identifying places and movements

Renso et al. [172] present a semantic enrichment knowledge discovery approach to support human behaviour interpretation based on inductive reasoning to analyse the movement pattern and deductive reasoning for behaviour inference. Their method utilises human domain knowledge by creating a behaviour ontology when there is a need for classifying the movement pattern. The process of semantic enrichment in this work has the following steps: 1) data preprocessing, 2) ontology mapping, 3) data mining, and 4) reasoning. The preprocessing identifies the stop, move, and patterns while the ontology mapping including axioms which establish complex concepts such as [172, pg. 341] "a place that the user frequently stops during the night is home" or "a place that the user spends most of his time is work place". The data mining analyses the data by using a frequent mining algorithm to find the most frequent stops. And, eventually, the reasoning engine assigns the relevant labels to the places that have been defined by the ontology (e.g. assign "work" to all the places that the user stops during the day).

Alvares et al. [9] present a generic idea about preprocessing trajectory data which provides a set of stops and movements, and then appends to the rearranged and simplified data the geographical information provided by the users or prior system knowledge. This method argues that semantic enrichment of trajectories with

geographical information can remove the complexities from further analysis and the data mining process.

Spaccapietra et al. [195] introduce a conceptual modelling method which adds behavioural semantics as well as weather information to the animal trajectory data. This paper discusses that idea that the semantic annotation enables users to understand the trajectories and facilitates the process of data mining.

Group 2: Semantic enrichment

Janssens et al. [100] collected ground truth data for validating the semantic enrichment by providing a tool that annotates GPS information. Although the obtained dataset has been used to validate several approaches in the field of semantic enrichment, it has not been made public for use by many researchers. However, some research, such as [82], uses this data source to assess their model, which is dependent on spatial and temporal rules. They propose an approach to automatically infer the user's activities by initially detecting the stops, finding the most likely POI category, and inferring the most likely activity. The POI is identified by using a probability algorithm based on the Gravity Model that incorporates the duration and time information of the stop.

According to their evaluation, they achieved mixed accuracy results, e.g. shopping 0%, food 83%, education 3%, and leisure 49%. They identified two main issues within their work, namely, lack of a robust POI service, and mapping issues between the activities and the categories.

Another approach by Reumers et al. [173] attempts to infer the activity type by establishing a decision tree model that considers the activity start and duration. They created a prediction model and a probability distribution based on the decision tree. Their model shows that the time information has a vital impact on the process of semantic enrichment.

Moreover, Parent et al. [156] provide an extensive survey on modelling and analysing semantic trajectories. This survey studies approaches in three main areas:

1. trajectory construction from raw movement data to identify movements and stops.
2. semantic enrichment of trajectories to effectively interpret movements.
3. analysing semantic trajectories to gain knowledge as well as reasoning about the behavioural patterns, characteristics of movements, and life pattern.

Andrienko, Andrienko and Fuchs [17] proposed privacy-preserving semantic analysis of movement, an abstract form of semantic space by employing the concept of cartographic chorems. They convey the process of semantic enrichment by using a POI database and then visualise the output as flow between the predefined typology of human activities.

Spinsanti et al. [196] suggest an approach to deduce human activities based on the sequence of places visited by users. To this end, they introduce an algorithm to identify a most probable place based on the list of places from POI APIs and two semantic rules, namely, the constraint rules and the probability rules. The constraint rules target eliminating implausible POIs and stand on commonsense whilst the probability rules intend to accommodate the best POIs based on domain knowledge (Figure 2.1). The output of this algorithm is a list of measured probabilities between each pair of stop and determined POI. This process has the following steps:

1. Selecting POIs by taking two conditions into account
    - There is adequate time to go and come back from the stop to the POI;
    - There is adequate time to spend at the POI. [1]

[1] For instance if the POI is a restaurant, the average time that a normal user would spend would be between at least 30 to 120 minutes (2 hours).

2. Computing the probability of POIs based on their distance to the associated stop and also their average visit time by using the below equation [196, pg. 46]:

$$P_{i,x} = \frac{TempP_{i,x} + \alpha \times SpatP_{i,x}}{\alpha + 1} \tag{2.1}$$

Where $P_{i,x}$ is the probability of each pair of stop (S_x) and POI (P_i), $TempP_{i,x}$ represents the temporal probability linked to the average visit time, $SpatP_{i,x}$ denotes the spatial probability linked to distance, and α is a coefficient to assign weight to the distance criterion.

3. Updating the probability based on the domain knowledge by using a heuristics when step 2 identifies the POI within the specified threshold [196, pg. 46]:

$$\begin{cases} P(Cat_x) = \sum_{k=1}^{i} P_{k,y} & \text{where } Cat_y = Cat_x \\ P_{i,x} = \frac{P_{i,x}}{\sum_k P(Cat_k)} \times P(Cat_x) & \text{more than one POI per category} \end{cases} \tag{2.2}$$

4. Amending the previous history of probability to increase the the level of certainty for uncertain stops.

Figure 2.1 shows an example of how the approach calculates and updates the probabilities for each category (B, D, G) by considering the probability of the past stops (S_1 and S_2). The third stop (S_3) has four POIs where #2 and #3 are associated to category G. The current categories can be updated by contemplating the past stops and considering that category G has an aggregated probability $(16.51 + 22.01 = 38.52)$ [196, pg. 47]:

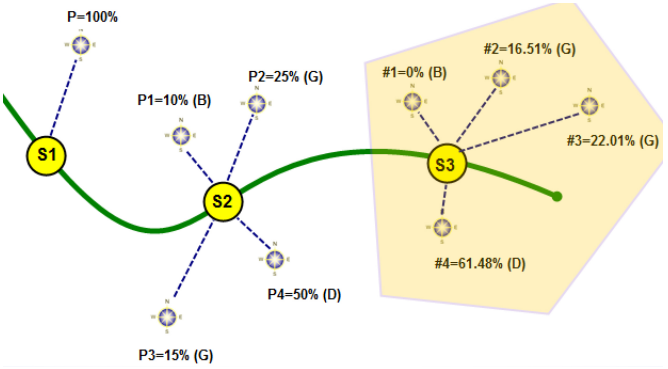


FIGURE 2.1: How the Spinsanti et al. [196] approach works

$$
\begin{aligned}
P(G3) &= \frac{0+40+38.52}{3} = 26.18 \\
P(D3) &= \frac{100+50+61.48}{3} = 70.49 \\
P(B3) &= \frac{0+10+0}{3} = 3.33
\end{aligned}
\tag{2.3}
$$

probabilities for current POIs will be updated as:

$$
P_{T1} = 3.33\% \qquad P_{T2} = \frac{16.51}{38.52} \times 26.18\%
$$

$$
P_{T3} = \frac{22.01}{38.52} \times 26.18 = 19.96\% \qquad P_{T4} = 79.49\%
$$

This approach is evaluated by using the POIs from the city of Milan in Italy. They used conceptual hierarchical POIs by classifying 39,256 POIs into four main categories and assigning an average duration. They achieved high accuracy for trajectory classification without involving the semantic rules in the evaluation process.

Yan et al. [236] present a framework that allows a flexible trajectory annotation at different levels and a novel annotation algorithm by exploiting the third party POI data sources to identify suitable points of interest based on the time period. They designed an algorithm based on a Hidden Markov Model (HMM) to filter POI category according to their previous state and work with numerous possible POIs within a densely populated area. However, they use types of POIs (POI category) instead of POIs as, in their view, inferring the exact POI from the large

number of POI candidates, which are queried from inaccurate location records, is a complex and expensive process.

They use the Milan dataset to evaluate their novel algorithm on the semantic point annotation. Moreover, they used a web interface to allow users to query their own GPS traces through a browser. Their results show that semantic annotation of stops performs very well on heterogeneous trajectories whist the storing time (write back the result) is low.

The most relevant work to the current approach is the semantic enrichment of movement by Krueger et al. [120]. This work attempts to add semantics to the trajectories by using POI service Foursquare – a venue-based social networking service with robust information about POIs. They undertake a preprocessing on the trajectory data before enriching the data semantically to determine stops, movements, and destination clusters. The authors introduce a POI decision model to interpret the imprecise and incomplete points within the data in a categorical way. The Foursquare service is employed by their decision model to discover nearby POIs and corresponding categories for a given coordinate. This model is built to handle the hierarchical categories provided by Foursquare and enrich the trajectories via:

- Creating a categorical result tree to organise the three level hierarchical categories e.g. top level (level 1), subcategory (level 2), and sub-subcategory (level 3).
- Considering distance of each venue to a given query point and calculating the average Gaussian value for all POIs within each category by means of Gaussian Kernel.
- Calculating certainties score based on the average Gaussian model
- Giving consideration to the number of check-ins to each categories as well as the number of Foursquare users who are logged in.

However, this work attempts to enrich the trajectory only by calculating the overall category score of nearby places. To calculate the category score, they formulate equation 2.4 that incorporates $cDist_{cat_l}$ the certainty score of the average Gaussian value based on the distance, $cCheckins_{cat_l}$ the number of check-ins for each category in Foursquare, and $cUsers_{cat_l}$ the number of logged in Foursquare users. Their equation can be used as a linear regression model by including α, β, γ as factors [120, pg. 908].

$$cM_{cat_l} = \frac{\alpha * cDist_{cat_l} + \beta * cCheckins_{cat_l} + \gamma * cUsers_{cat_l}}{\alpha + \beta + \gamma} \tag{2.4}$$

This model is evaluated by using two datasets with known ground truth and it achieved a reasonable accuracy of over 80%. However, this work is not designed to determine the real places within the trajectories or provide multi-level information.

Based on the literature review, most of the approaches infer the human behaviour and perform semantic enrichment by adding the categorical information of POIs. This means that although the category of the POI is is predominantly identified, the real venue yet is not fully determined. To address this problem, unlike the Google Map Timeline with tremendous geographical information such as places' boundaries, sizes and nature of businesses that can be shared across its services, this work attempts to use only raw spatio-temporal data and Foursquare POI service to discover the foremost venues, annotate the visited places according to the highest score calculated by our novel approach and provide shortlisted venues in order to compensate for inaccurate annotations.

2.2.2 Event detection and ranking

Extracting the significance of the data points is challenging. There are a series of factors which need to be considered in the direction of identifying accountable significance. There are a number of works that strive to provide a sensible visualisation by identifying the most important data points.

Andrienko, Andrienko, Hurter, Rinzivillo and Wrobel [13] introduce a place-oriented analysis of movement data to extract significant POIs from population trajectory data by examining the events that occurred repeatedly and their temporal distribution. They propose a scalable visual analytics approach that implies four main parts, namely, event extraction, obtaining the relevant places by clustering the events, event aggregation, and analysis. In this approach, the repeated event occurrences are determined by using density-based clustering [69] such as DBSCAN or OPTICS according to the event's position in space and in time. They provide a flexible clustering that can accept a customised distance function between the events. Therefore, the distance function is required to get a vector distance threshold from the user manually. Next the user is required to state the method to aggregate the distances, either by selecting the maximum value or by using the Euclidean distance.

Stream of our lives (LastHistory) [25] visualises everyday music listening history together with contextual personal information such as photos and calendar events. Daily-streamed songs are shown by using colour-coded circles to represent the genres and the frequency-based song-ranking algorithm for the mood analysis. The songs are categorised based upon genre and assigned a distinct colour. Additionally, to show the personal relevance of a song/track, the sizes of the circles are influenced by a song's comparative importance as well as the overall importance (Figure 2.2). The size of the circle is calculated by equation 2.5 [25, pg. 1123]:

$$w_h = \frac{\left|t(h)^P\right|}{|t(h)|} \times \frac{|t(h)|}{|t|_{max}}(1 - \frac{M}{2} + \frac{d(h)}{d} \cdot M) \tag{2.5}$$

where (h) represents the history, $t(h)$ is the overall number of entries (history) for track t, $t(h)^P$ is the numeric amount of history entries within the range of P time surrounding h while $|t|_{max}$ is the maximum number of entries (histories) for all tracks. The (d) indicates the time span between the first and last entry within the listening history while $d(h)$ is the time span between the first history entry h. Here, M is a constant denoting the weight influenced by m factor which boosts the weight of younger songs with the lower play counts compare to the older songs

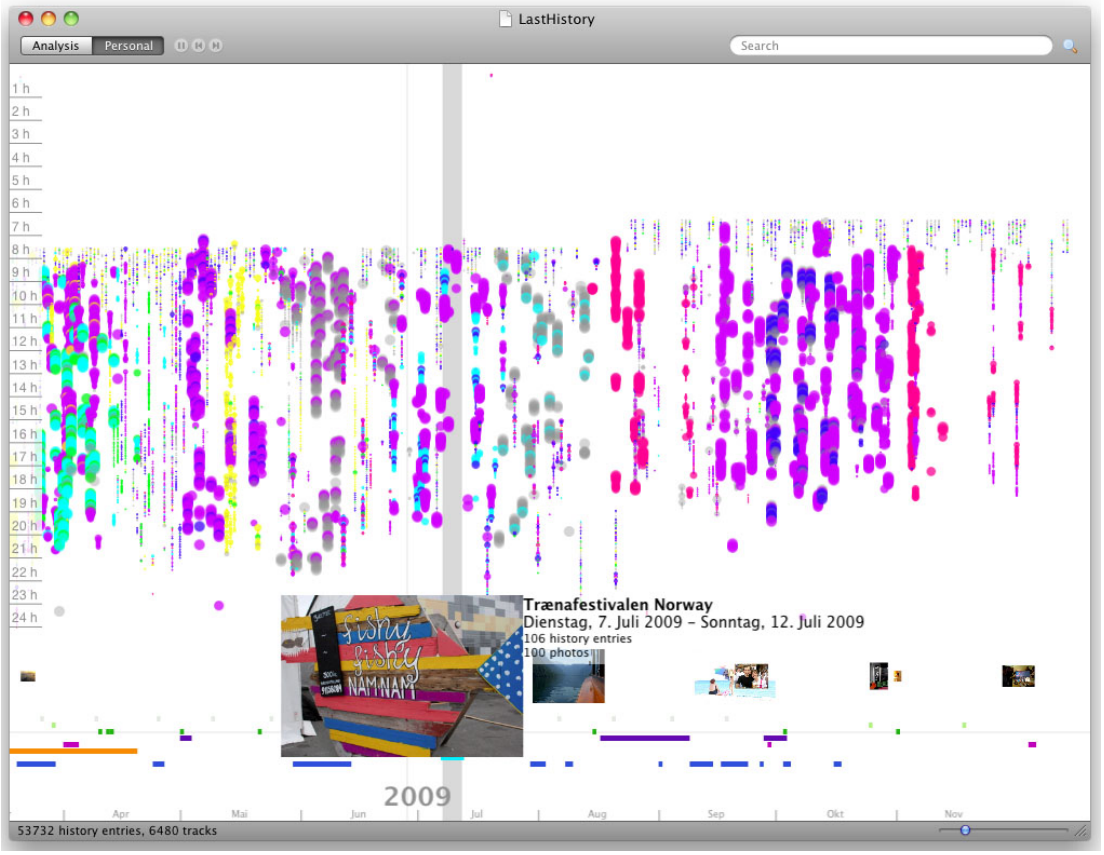


FIGURE 2.2: Last history [25] interface with the colour coded circle depending on the genre and various sizes related to the personal relevance of the song

with considerably higher play counts. The rang of m is $(1 \pm \frac{M}{2})$ and m for any song, precisely in the middle of the listening history, equals to 1. According to the paper, the authors considered $(M = 0.5)$ based upon a number of listening histories sample. The visualisation of this work is reviewed thoroughly in the personal visualisation (Section 2.3.3) within this chapter.

Dias et al. [60] similarly attempt to visualise music listening history by laying out the songs as stacked dots along a timeline. The dots are encoded based upon the frequency of playings of the song and its relevancy in the whole available history of the song. In this approach, the songs are classified and colour-coded according to their release dates. Figure 2.3 shows the encoded dots in detail. The brief review of this approach is in Section 2.3.3 in this chapter.

In another work, Wood et al. [233] propose a flow map to visualise the dynamics

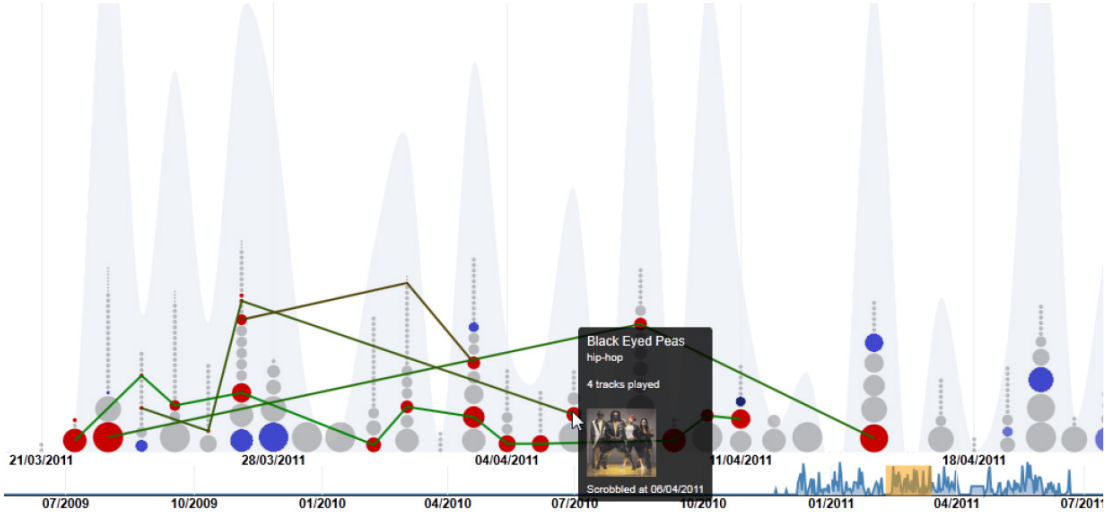


FIGURE 2.3: The representation of the streamed songs over the course of a day on the timeline by using different size of stacked dots [60]

of London's bicycle hire scheme and determine alteration in travel behaviour over space and time. This work endeavours to design a careful approach to prevent the visualisation from salience bias, occlusion, and information overload while encoding a great number of flows (Figure 2.4). To emphasise the significant and more relevant flows, they use different transparency, width, and colour in accordance with a calculated weight based upon the relative frequency of the trip:

$$w_{od} = \left(\frac{f_{od}}{f_{max}} \right)^{1.5} \tag{2.6}$$

where f_{od} is the trip frequency between the start (origin) o and destination d, and f_{max} is the maximum frequency between any start and destination within the whole data.

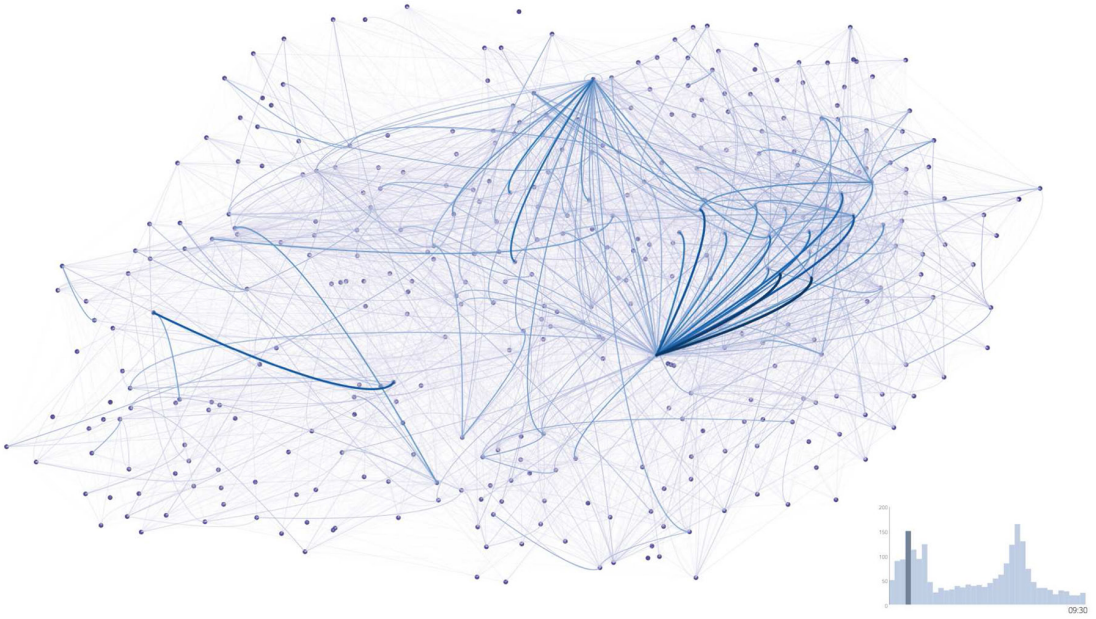


FIGURE 2.4: Flow map [233] to visualise the dynamics of London's bicycle hire scheme by indicating the significance with weighted lines

2.3 Data Visualisation

2.3.1 Visual information theory

Visual analytics is highly dependant on the information visualisation as well as the interaction to deliver a meaningful feature of analysed/mined data and allow for acquiring knowledge. There are many approaches available nowadays that have been proposed and implemented to achieve this goal [6, 11, 30, 68, 72, 85, 122, 209, 215, 218]. The majority of these methods have been developed by means of visualisation theory including the human cognition model, visualisation modes, and visualisation type. Henceforth, this section briefly looks at the key parts in the rest of this section.

2.3.1.1 Human cognition model

Visualisation plays a vital role in the human cognitive system. Humans obtain information via the visiual system more than any other sense. The most essential part of human cognitive activity involves a pattern finding procedure. Human cognition in visualisation implies a mixture of abilities to discover and learn, in which they are occasionally binary, certainly not entirely sequential, and come with eluded model-based prophecy [11, 15, 86, 124, 126, 126, 228, 229]. This means that humans utilise the most likely and straightforward heuristics to conduct tasks. Human visual structure has adaptive decision making together with a flexible pattern system while the computer benefits from computational power and extensive information resources. Hence, interactive visual exploration is considered as a connection between the two, and enhancing this connection can result in increasing the performance of the data visualisation system. Thus, human and computer both strengthen the visualisation task by working together to explore and gain new knowledge. However, human and computer have different abilities to attain this goal. The most distinct human ability consists of:

- Adaption, the ability to incorporate the newly learned information in existing knowledge
- Accommodation, the ability to lay the new knowledge in the nearest corresponding information or generate a new one when the discovered finding is novel and does not fit the current knowledge depiction.
- Brief reasoning analogy and problem solving, the capability of eliminating available insignificant attributes by means of cognitive effort in order to narrow down accessible options.

Whereas a computer can accompany the human with the two following processes:

- Remarkable processing memory, the ability to keep all relevant information with respect to the mental model

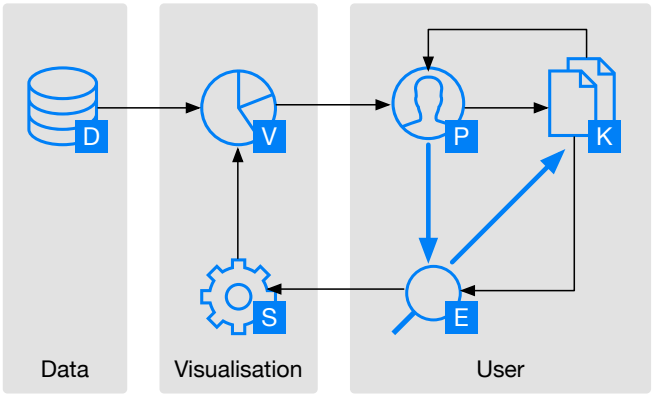


FIGURE 2.5: Improved visualisation model with higher cognition by Green et al. [86]

- Processing with no cognitive preferences (biases), the ability to analyse and encode data without loosing or filtering the relevant information based on a perceived hypothesis.

Green et al. [86] proposed an improvement on the human cognition model based on the Wijk [228] original visualisation process model to intensify human and computer's supportive cognitive strengths and provide effective visualisation design guidelines. This research employs the enhanced model to improve its knowledge discovery. The model is illustrated in Figure 2.5, where the user as perception cognition (P) observes the encoded image (V) and uses the visualisation specification (S) by means of exploration (E) to gain knowledge (K).

In this model exploration (E), knowledge (K) and perception (P) are all linked together as a cognitive process in order to exchange information. This model shows that perception performs a vital role in exploration.

Moreover, it points out that interactive exploration can provide reasoning knowledge during the exploration by a human and lead to more focus for further exploration and henceforth further knowledge discovery. Therefore, Green et al. [86] articulate the process of knowledge discovery as follows:

$$K(t) = K_0 + \int_0^t \mathrm{E(P,\ K,\ t)dt} \tag{2.7}$$

where the knowledge is expansion of existing knowledge and held by the user via perceptual progress of the exploration. Additionally, as stated by Wijk [228], the user may adjust the visualisation specification (S) for further data exploration by means of interactive exploration, here known as $E(K)$:

$$S(t) = S_0 + \int_0^t (\frac{dS}{dt})dt \tag{2.8}$$

This means that the current specification evolves over time followed by the initial specification (S_0).

This work uses this fundamental term in human cognition towards designing the effective visual analytics approach which benefits from the computer's strengths in analysing the personal life logging data and interactive visual encoding in order to meet the knowledge discovery goals in this context.

2.3.1.2 Visualisation modes

To propose, design, and implement a novel visual analytics technique, the fundamental terms are investigated, not because the existing works in this context have not studied or employed such terms, but to establish a better understanding of the included terms and also to design empirical modules for personal daily life data with the current challenge in mind.

According to Wills [229], data visualisation incorporates three individual modes, namely, interactive visualisation, presentation visualisation, and interactive storytelling. Interactive visualisation corresponds to a process of knowledge discovery by encoding the data based upon user input and delivering a prototype quality visualisation. This type of visualisation is intended for a single user who controls almost everything together with datasets. In contrast, presentation visualisation is used for communication by means of highly polished visual encoding and hence intended for a mass audience or large group but does not support user control or user input. The last and the most flexible mode is interactive storytelling as it

provides a presentation via an interactive web-based approach and allows users to filter and examine details of the datasets. In general, the ideal visualisation should infer the following:

- Help the individual to learn features of the data (knowledge discovery);
- Envisage known features of the data to users; and
- Provide interactive exploration for further knowledge discovery and delving into the details.

Additionally, as stated by Wills [229], data visualisation techniques should determine their target users and how users mainly employ them to gain knowledge. There are two main modes that can be considered for use in visualisation techniques, immersive mode and reflective mode [30, 229]. Immersive mode is mostly selected for visualising data in an expert mode as it involves the user in the data processing, decision making, action taking, and observing the outcome as part of its workflow. In contrast, reflective mode allows the user to delve into the visualisation unhurriedly, contemplate a conceivable hypothesis, and subsequently validate it by further interactive exploration of the visualisation.

2.3.1.3 Visualisation types

The design method based upon the information seeking mantra from Shneiderman [191] is considered a top-down approach. The system encodes all the data at a low level of detail – overview first – and enables users to establish subsets, and tune their observations to gain better understanding. The visualisation provides narrowing-down, filtration, and interaction – zoom and filter – in order to allow users to explore further to discover their answers. Finally, details-on-demand allows users to get more information and understand data better. This process starts with the entire dataset as an overview and continues by exploring, drilling-down, and taking further actions [3, 110, 115]. This process includes a number of

trials and perhaps errors as users try to familiarise themselves with the system. Providing guidelines and also previewing the procedure can prevent users from numerous false starts. Moreover, using tooltips can help users to swiftly and deliberately uncover the desired information within the process. The seeking mantra stated that the details should be solely revealed at the lowest part of the analysis. A more fitting statement for the current data – considering size and complexity – would be details-always-on-demand [19, 229].

2.3.2 Time-oriented data visualisation

Time is a unique dimension and comprehending the relation between time-oriented data allows the user to gain insight from the past towards planning or predicting the future [3, 5, 85, 202]. Henceforth, pertinent visual analytics methods can support such data where the main goal is knowledge discovery and gaining meaningful understanding.

The life logging data – particularly daily life – are highly associated with time and space. The visualisation of this type of data in revealing compelling facts and insights needs to extensively consider the time points, start and end of each course as a time intervals, and extreme changes [3, 4, 21, 188]. There are various infoVis techniques available in the visualisation community which try to handle such data by either proposing a novel method or combining existing ones.

However, by reviewing several approaches, it can be established that analysing and subsequently visualising the time-oriented data have several aspects which make the process of analysis and visualisation extremely hard [3, 21, 115, 189]. Many techniques have been developed and published within the visualisation community to address such data but the majority were only able to tackle a specific problem. This is due to considering time as a typical quantitative variable that can be envisaged via generic methods rather a distinct dimension. Theoretically, generic methods are not suitable for analysing temporal data with numerous dimensions

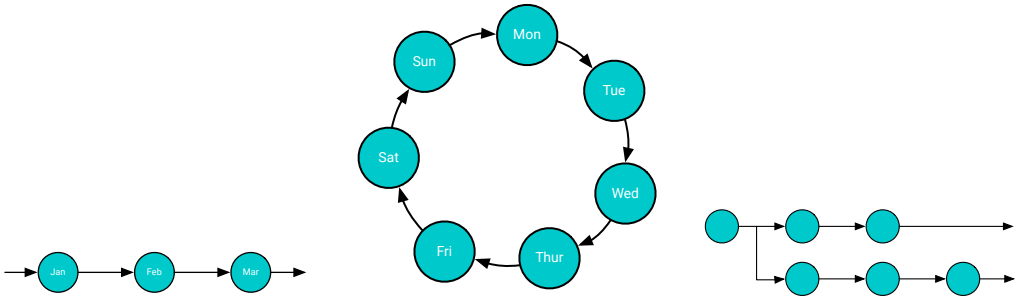
as they can not establish a perceivable connection between time and multiple variables.

Aigner et al. [5] categorise the techniques for time-oriented data visualisation into time, data, and representation. Time is divided into two subcategories with respect to Frank [78], namely, temporal primitives and time structure.

Temporal primitives create the time axis, which can be time points (considered as an instant) or time intervals (a time point with a duration or two time points). Subsequently, the structure of time can be divided into linear, cyclic, or branching (Figure 2.6). *Linear time* is a collection of temporal primitives in a linear order which is similar to the natural human perception of time in life from the past to the future whilst *cyclic time* represents a fixed set of temporal primitives such as days, weeks, months, and seasons. *branching time* is related to when time can be split into alternative scenarios at the same time. Moreover, time-oriented data is categorised by Aigner et al. [5] into three parts:

- **Contexts**: The data can be abstract or spatial. The former corresponds to collected data that are not connected to some spatial context while the latter denotes data that implies inherent spatial content. However, the process of visual analytics for these data is not the same.
- **Number of variables**: The number of time-dependant variables is another criterion that is considered within such data. It is important to distinguish between temporal data that are associated with a single value (univariate data) or a set of values (multivariate data).
- **Level of abstraction**: Visualising the data is a valuable asset of visual analytics. However, when it comes to visualising large-scale data, visual clutter is a common problem. In this case, it is essential to abridge the data in the light of users' needs and interests. This allows the visualisation to envisage large-scale data effectively [181].

The last part is the representation of this data. The representation is divided into two major parts: time dependency and dimensionality. Time dependency



(a) Linear time (b) Cyclic time (c) Branching time

FIGURE 2.6: The structure of time defined by Frank [78]

implies two terms, static and dynamic, in which the former is related to the representation in still images whereas the latter is a representation that shows – by using animation – the changes over the time. Both terms have their weak and strong points. Visualising large-scale data as a single image (static) requires an enormous display or set of displays whilst using an animation (dynamic) may take considerable time for such big data and limit the process [194]. Moreover, data can be envisaged in a traditional way as 2D or 3D but the main challenge is that there are still not strong grounds for using particularly 2D or 3D within the visualisation community. Some researchers claim that employing 3D can open up the possibility of encoding further information by using the third dimension, whereas other researchers argue that 3D can lose information on back faces and leads to visual complexity.

Mapping time-oriented data on spirals

As mentioned in Section 2.3.2, temporal data can appear linearly, cyclically, and branched. Days, weeks, months, years all reoccur periodically and although time is inherited and moves forward linearly, days, weeks, months, and years can be considered as cyclic temporal data. Carlis and Konstan [42] introduce a new visualisation technique to represent the serial and periodic properties of time-oriented data – known as serial periodic data – by mapping the data onto an Archimedean spiral layout to draw periodic attributes on the radius and position

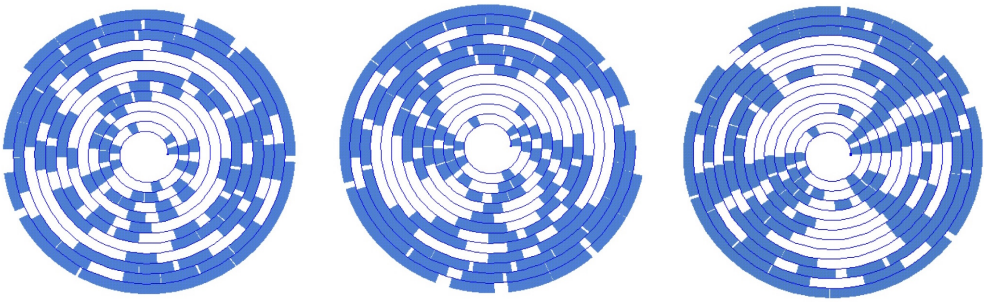


FIGURE 2.7: General example of spiral layout [42]

serial elements alongside the spiral axis in 2D as well as 3D (Figure 2.7). The authors alleged that the spiral visualisation facilitates the process of comparison and understanding the temporal data in a seasonal and periodical way.

Weber et al. [221] present a comparable interactive spiral approach for visualising time-related data which supports large-scale and periodic data structures, and comparative analysis. This approach reviews a different type of spirals and considers the Archimedean spiral as an appropriate algorithm to map time-related data. The form of the spiral in the Weber et al. [221] approach remains untouched but additional visual elements – though not adequately effective – such as colour, texture, line width, and icons have been used to visualise the data. This technique facilitates detecting the useful cycles and identifying periodic patterns within the temporal data (Figure 2.8).

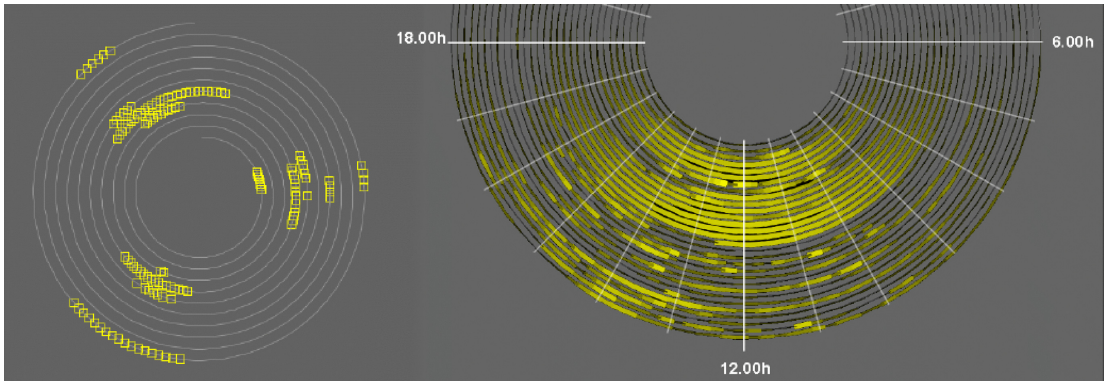


FIGURE 2.8: Example of 24-hour spiral layout by Weber et al. [221]

Although this method claims that the users found the visualisation beneficial in extracting periodic and seasonal information, the major obstacle in this method is

the visualisation. Encoding the temporal data starts from a small-scale ring and continues to the larger rings. This results in an inconsistent view of the cycles in the visualisation and can lead to misinterpretation of data.

Cluster-based and Calendar-based

Van Wijk and Van Selow [211] propose a technique to concurrently determine potential patterns and trends on multiple time scales, such as days, weeks, months, etc., by clustering identical patterns, mapping trends and patterns on a calendar, and interaction. In this approach, the temporal data are considered as two dimensions – days and hours – and plotted on two individual axes. Moreover, another dimension can be displayed as the height of the plotted data over time – in this case, days (Figure 2.9). However, identifying the differences over the weeks is cumbersome, and additionally the weekends patterns are almost blocked.

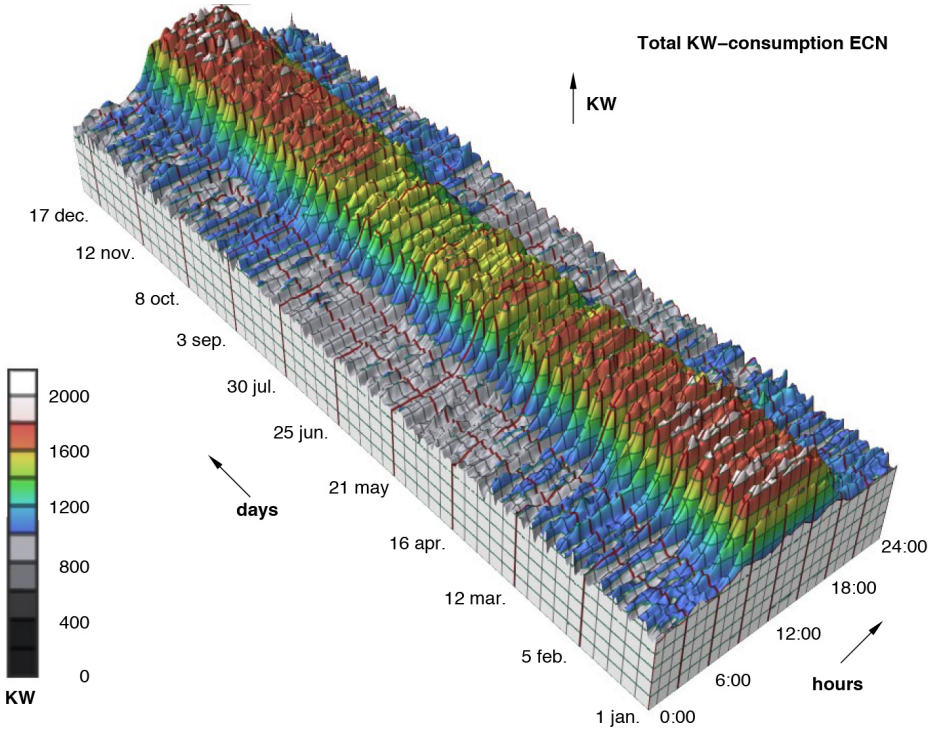


FIGURE 2.9: The visualisation uses hours and days to display the power consumption of the Energy Research Centre of the Netherlands (ECN) by Van Wijk and Van Selow [211]

To address this issue, they attempt to use a bottom-up hierarchical clustering algorithm by employing a root-mean-square distance (equation 2.9 where y_i and z_i are two-day patterns and $i = 1, ..., N$) in order to merge similar patterns into a number of clusters and map them on the developed combined calendar (Figure 2.10).

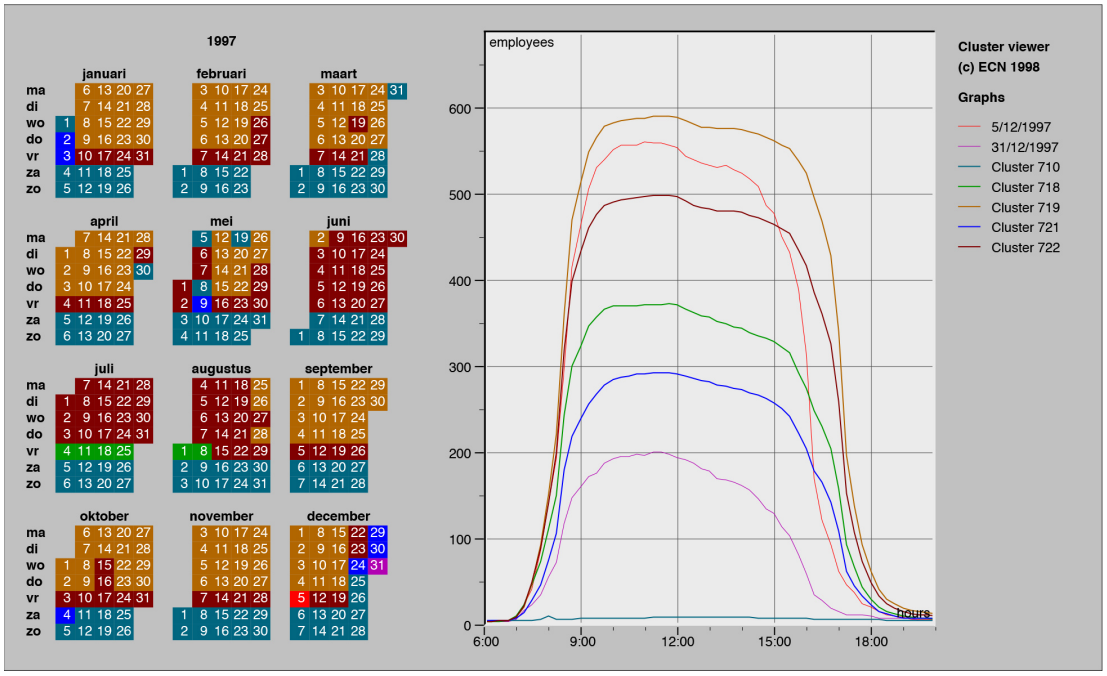


FIGURE 2.10: Calendar shows the clusters by distinct colours on the left and patterns on the right with the same colour scheme [211]

$$d_{nm} = \sqrt{\sum_{i=1}^{N} (y_i - z_i)^2 / N} \tag{2.9}$$

This technique shows that incorporating the cluster analysis in calendar-based visualisation not only uncovers meaningful insight – in this case about the power consumption – but could help predictions in different domains.

TheMail

TheMail [214] is a typographical visualisation tool that shows how a relationship changes over the certain period based on the conversational histories, particularly

the words specifying the relation between sender and receiver. This technique presents multi-layer information by lining up the keywords – with different sizes based on their uniqueness as well as frequency – on a timeline, using circles to show the length (size of circles) and direction of the email, colours in order to distinguish the keywords – in terms of monthly or yearly frequency – and the circles for indicating the email direction, and interaction for further details (Figure 2.11). TheMail scores the words based on their pertinent frequency by using $tf - idf$ algorithm [28] that provides two monthly and yearly scores for each word – within the exchanged email – in every month and year.

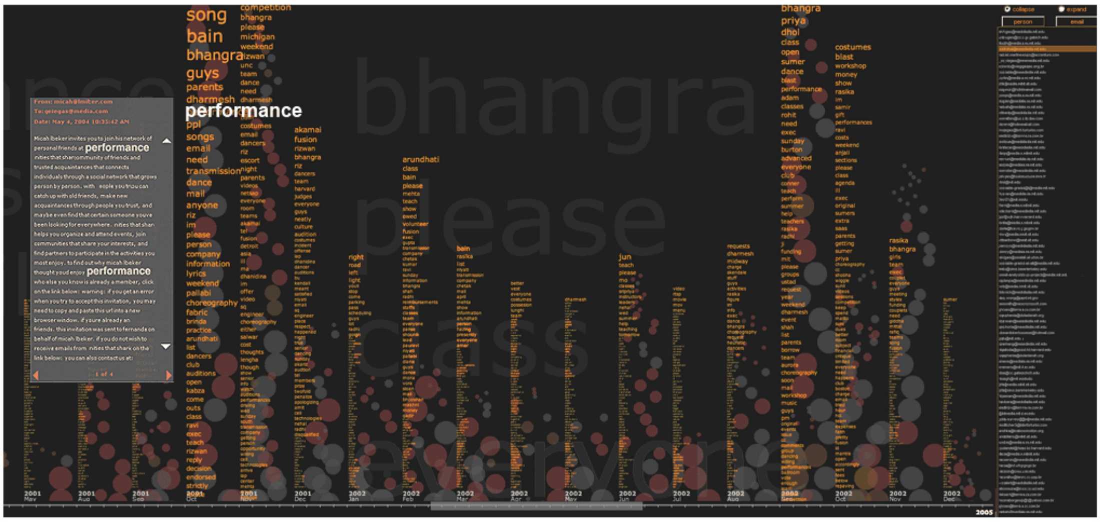


FIGURE 2.11: TheMail visualisation [214] presents how the particular relationship moves forwards, here, by showing the most frequent words in emails that the user has exchanged with a friend over a period of 18 months. Interaction shows the original email in which the selected keyword has been mentioned. The circle size indicates the length of the email while the colour shows the direction (incoming or outgoing).

TheMail finds two different types of users – the ones who are looking at the overall pattern, and the others who are interested in finding more details on specific keywords – for its visual analytics method. This method provides the users with effortless information to grasp interesting information about how their relationships are developing over the time. However, the authors identify two major limitations: the first is the content analysis algorithm quantifies all the email in a similar fashion without considering the particular conditions. The second is the parsing method that contemplates only single words without any knowledge regarding expressions and such like.

PatternFinder

PatternFinder is introduced by Fails et al. [71] to provide a visual query and then visualise a set of results in order to facilitate the process of temporal pattern discovery within the bounds of multivariate datasets. This approach allows user to define the time range and the event elements towards of making an effective query, exploring, and yielding a meaningful temporal pattern discovery (Figure 2.12). It offers a simple visualisation together with interaction to show the query results following by zoom/filter and details-on-demand.

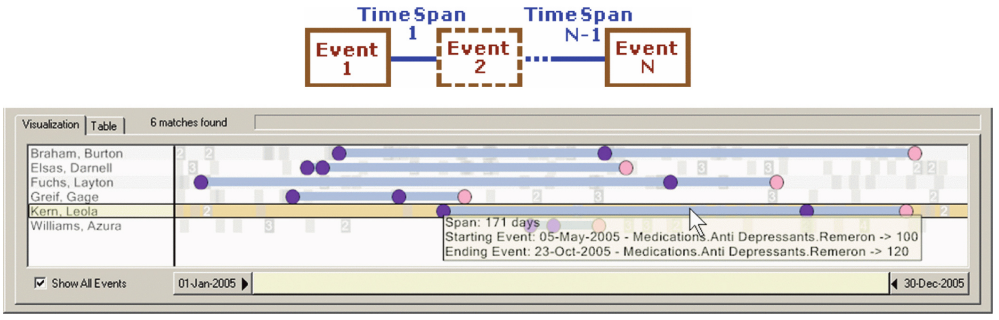


FIGURE 2.12: PatternFinder shows the result of the visual query made by the user. Each line shows a pattern of the patient matched with the query. All the other events which took place in that day are shown by grey slabs followed by the number of events. Interaction provides more details about the events. Note that there is no intensity or any kind of visual encoding employed to emphasise severity of the patient's condition [71].

The authors claim that PatternFinder's visual query interface is a major component of their work that allows for defining an influential pattern query by the user which is not supported by other existing systems at the time of implementation. Additionally, listing the matched results as summary is proved to be an effective way to test a more comprehensive set of hypotheses and discover more applicable patterns. However, it has been pointed out by the authors that making a visual query needs complete training as the pattern query panel comes with technical terms – e.g. time span or formulating events – and cannot be served as an easy-to-use interface. Moreover, visualising the multiple result of the query is rather simple and visual encodings are not carefully selected.

Timeline

Timeline is an intuitive and robust method to display continual time-oriented sequences by forming the time span of actions, linearly, along a vertical or horizontal line [122]. This method has been used by much research in the visual analytics community to delve into time-related data and discover meaningful information [7, 10, 23, 41, 91, 122, 177, 204, 241, 242]. According to [91], the main purposes of a timeline are: 1) informing the user; 2) showing the context; 3) providing related information. Allen [7] states that a timeline can present events and their temporal ordering considerably better than many other techniques, and can be understood easily by many users without much training. A timeline can also be used to display historic events, and temporal and semantic data [64]. One of the simple yet meaningful timelines is the Euler timeline [70] – see Figure 2.13.

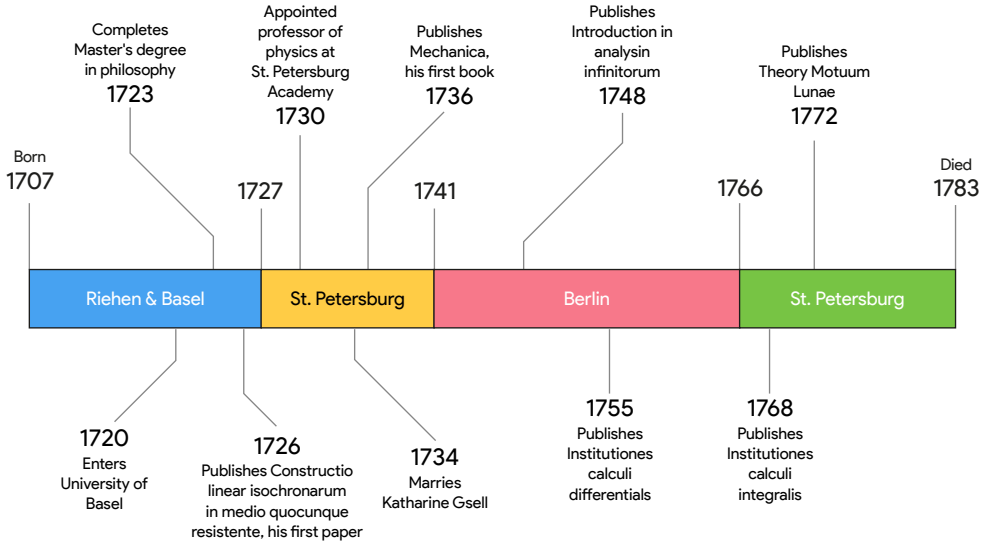


FIGURE 2.13: Euler's life timeline (re-drawn) inspired by this technique

LifeFlow, OutFlow and DecisionFlow

LifeFlow [232] is designed to deliver a scalable visual overview of event sequences interactively. It supports users' exploration for the medical purposes such as accident response time to determine the best practice (Figure 2.14). This approach

can encapsulate all potential sequences and portray the time-related spacing of the events within sequences. LifeFlow is evaluated by using two case studies for transportation and healthcare.

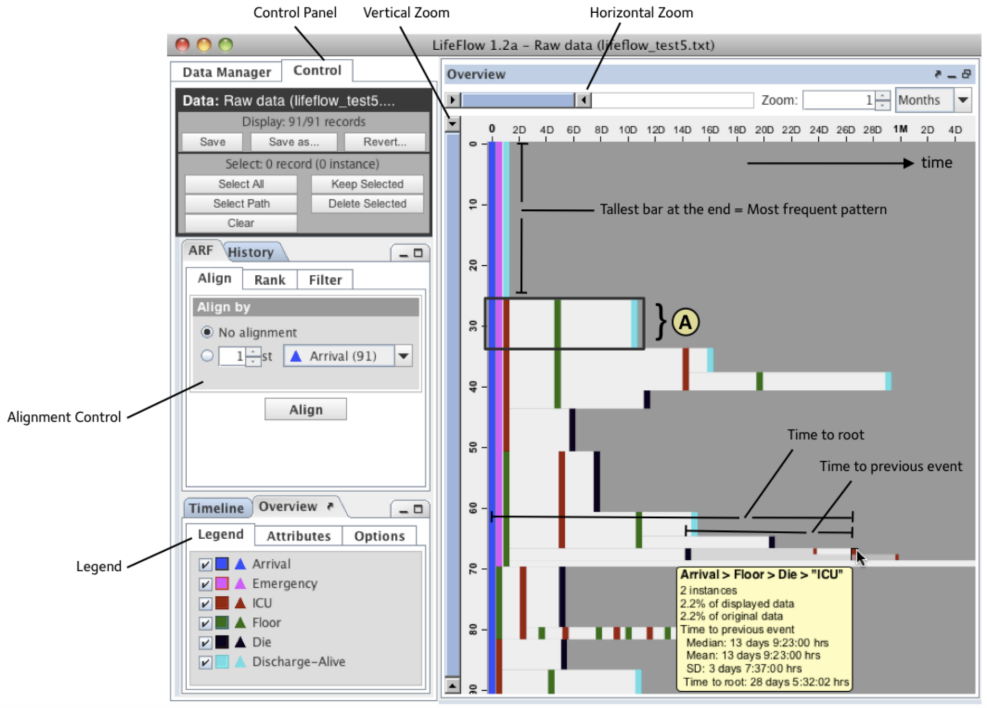


FIGURE 2.14: LifeFlow interface shows the sample patient medical data [232]

Outflow [231] and DecisionFlow [85] use Sankey-based visualisation [175] to support visualisation and analysis of the causal relationships of events within the complex temporal event sequence data. These methods use clustering and inverse document frequency to extract significant events in their work. DecisionFlow employs an interactive juxtapose visualisation and statistical analysis to propose a scalable and dynamic event visualisation (Figure 2.15). This technique is studied via a task completion process which recruited 12 participants. The evaluation results show that the approach enables the participants to complete a range of sequence analysis tasks swiftly and accurately.

Similarly, OutFlow – a successor of LifeFlow – is designed to encapsulate the patient Electronic Medical Record (EMR) temporal event data by an interactive visualisation that merges the patient records into a Sankey-based graph visualisation (Figure 2.16). This method offers an interaction to enable users to control

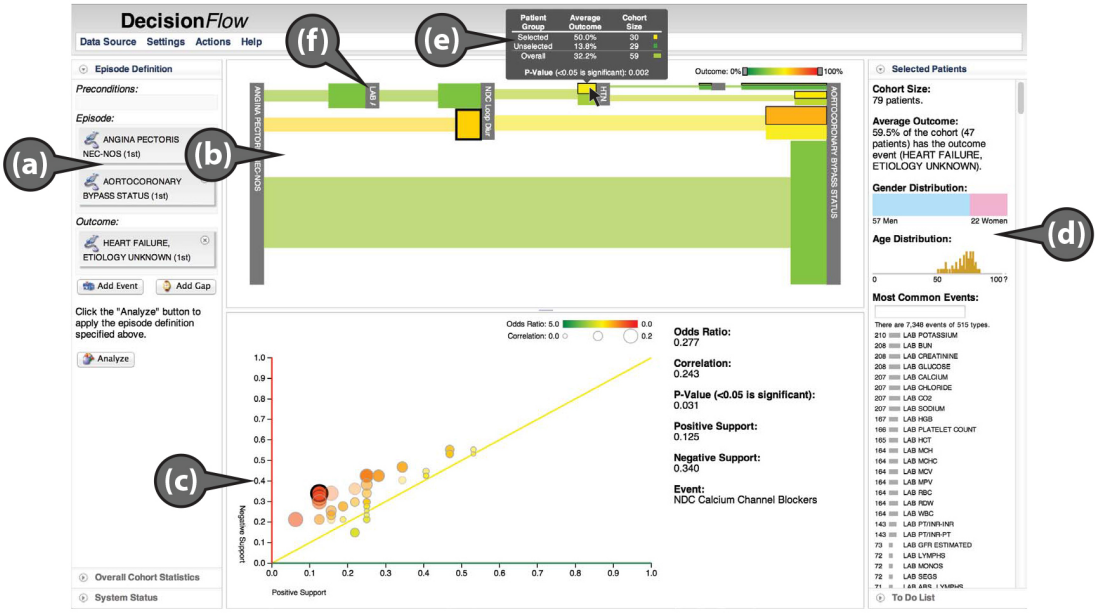


FIGURE 2.15: DecisionFlow interface to analyse the medical data [85]

the visual representation. This approach is evaluated by including two sample analyses to demonstrate the level of insight that can be gained from this technique.

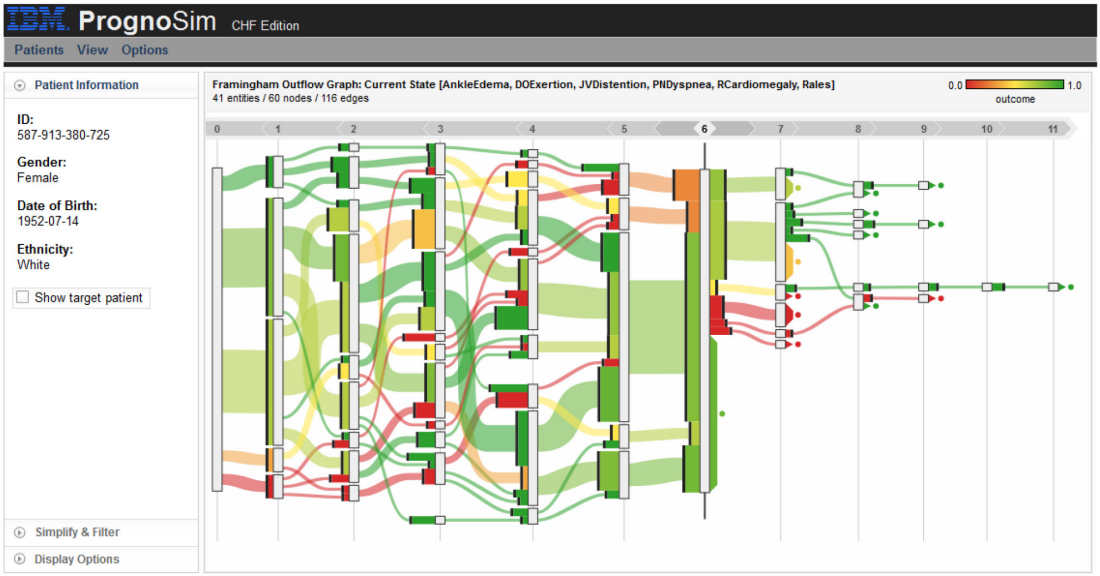


FIGURE 2.16: OutFlow [231] shows the aggregation of a cohort of patients including temporal event data

2.3.3 Personal data visualisation

As mentioned in chapter 1, the amount of practical personal data related to an individual's fitness, movements, health, and lifestyle – which are captured automatically – is growing day to day as a result of widely available sensors in mobile and wearable technology. In order to gain knowledge and empower an individual's life, this data, given its wide availability, needs to be reachable, comprehensible, and explicable [96]. This allows exploring the personal data to gain insight and better understanding of daily life, which can result in reminiscing about episodes of life, improving lifestyle, and making effective decisions. Moreover, the number of systems and applications that provide personal visual exploration and reasoning have also become greater [23, 25, 53, 80, 132, 202]. However, the weight of visual exploration is noticeably higher than the reasoning at the time of this review.

Huang et al. [96] provided a comprehensive review of personal life data visualisation and visual analytics, which this topic is distributed amongst various research communities and still many lessons learned, gaps and challenges might fail to be included despite sharing the results and outcomes. She introduced a new scope in this field, namely, personal visualisation and personal visual analytics.

According to Huang et al. [96], personal visual analytics is the use of analytical reasoning together with visual representation in making sense of personal life data. Similarly, personal visualisation comprises the design and use of interactive visual representation within personal life-related data. As stated by the definition, personal visualisation focuses only on how data should be visualised by means of various visual encoding but personal visual analytics includes visual representation together with computer-aided analysis.

A small amount of infoVis research concentrates on visualising personal and particularly life logging data. Huang et al. [96] identifies eight main areas in personal life that have been addressed by infoVis research, namely, healthcare, life logging, finance, social networks, political views, residential environment

and power consumption, movements, and recycling. In this section, the typical meaning of personal data is depicted and then the previous work around such data summarised, particularly that relevant to this research.

Personal data

Personal data, as partially mentioned by Huang et al. [96], in all respects, is related to the individual and can be used solely for personal exploration along with knowledge discovery by humans with different skills, preferences, and experience. This definition, to a certain degree, covers personal context [2], that has been argued by Ellard [66] to be in which the context can be simultaneously internal (e.g. specified goals and objectives) and external (e.g specified artefacts and settings) to people.

Personal data can be explored with different backgrounds, preferences, goals, experiences, and assumptions. Moreover, many people are not expert in data visual analytics and have dissimilar behaviours, priorities, and times to conduct analytical tasks. This signifies that research work should consider an individual's culture, understanding, and capability when designing a visualisation/visual analytics approach to identify how humans apprehend visualisation and depict their own data.

The data captured by wearables (e.g. Fitbit, Withings) or smartphone applications (e.g. Moves, Google Fit) can be strongly regarded as personal data. The data acquired comes with a standard timestamp format and are accessible. This data, depending on the device or application, can obtain activity-related data, geospatial data, and health data. Each wearable or application is equipped with its own tool in order to provide exploration and better understanding of personal data. These tools, evidently, share common ground in visual exploration work.

[2] Personal context means that, for instance, a user concerned about exploring traffic patterns for which the data may be public and do not have much personal pertinence or connection with the user would not have much personal context.

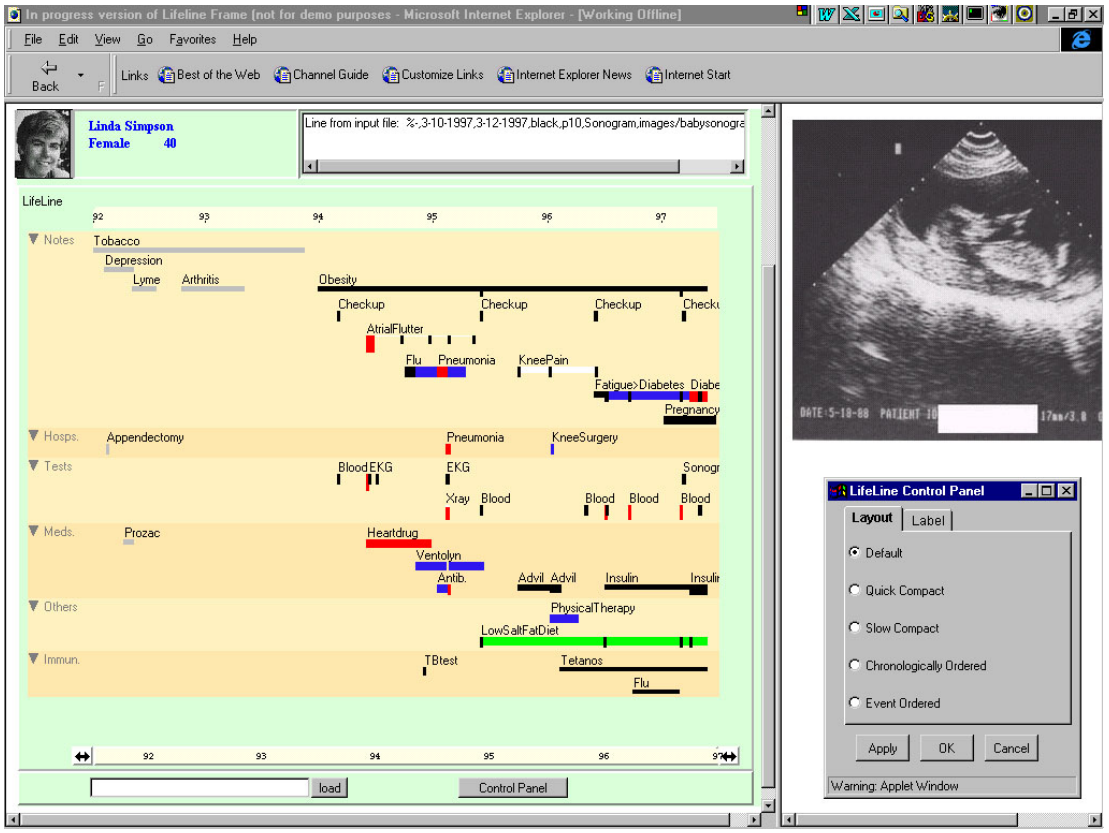


FIGURE 2.17: LifeLine multiple-view interface demonstrates the patient's medical records on the timeline in several facets with the facilities to zoom and filter. Source: [163]

LifeLines

LifeLines [162, 163] presents a one-screen approach in visualising time-related data of personal histories which can be applied to medical records data or other classes of biographical data (Figure 2.17). This approach displays different aspects of data by means of separated facets, lines, icons, colours, and thicknesses along the timelines to encode information and demonstrate individual events, relations, and importance. For instance, the thickness and the colour of the illustrated line on the timeline may correspond to the importance and the severity of an event. Moreover, this tool allows the user to narrow-down and focus on a particular part of the encoded data by offering scaling and filtering tools. LifeLines claims that it speeds up the process of identifying trends and anomalies, lessens the risk of missing information, provides access to details efficiently, and is customisable.

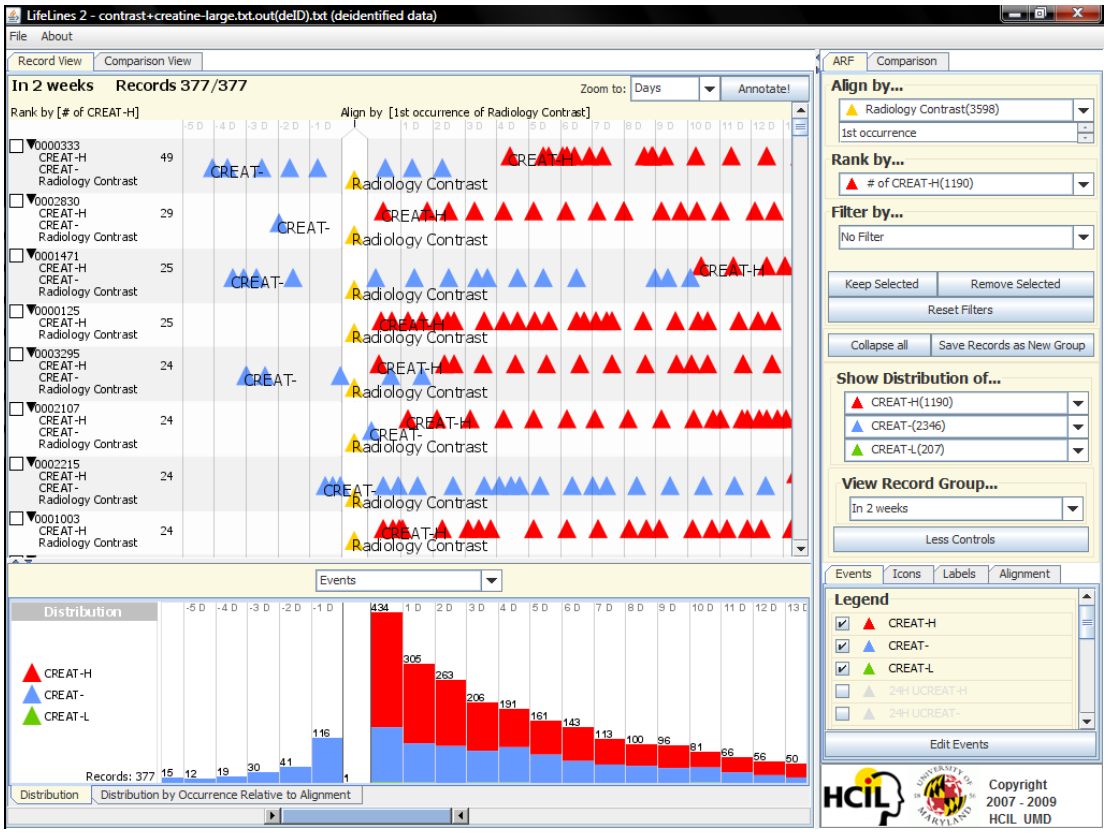


FIGURE 2.18: In LifeLine2 all records are aligned by the first occurrence of the radiology contrast. This feature facilitates comparison as well as determining the diagnosis within the defined days. Ranking, filtering and showing the distribution help to gain better understanding of multiple data. Source: [162]

LifeLine2 is designed to visualise parts of multiple patients' medical records based upon the LifeLine technique with enhanced features such as rank and filter of the results queried by the user. This feature provides effective patient record comparison and support for detecting hidden patterns within the entire dataset.

The Streams of Our Lives

This approach provides an interactive visualisation tool to display an individual's music streaming history together with available contextual information and aims at allowing an individual to examine and reminisce about their streaming [25]. This visualisation tool is designed for non-expert users and focuses on three tasks,

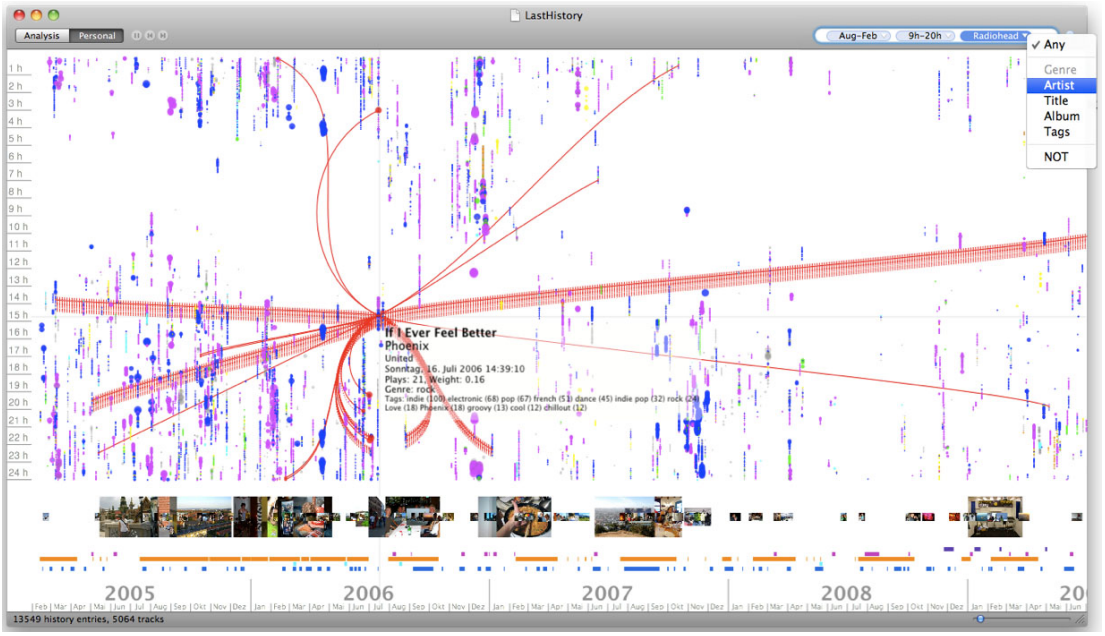


FIGURE 2.19: Stream of Our Lives [25] interface shows its interaction ability

namely, analysis of streaming activities, finding patterns, and examining theories based on the streaming data.

This approach uses a traditional timeline to encode the streaming data into a visual representation. The horizontal axis is used to line up days and the vertical axis is used to show hours and minutes of the day, respectively. The streaming songs are represented only as circles without displaying duration. This method utilises eight fixed genres with distinguishable colour to reduce unclear and incorrect classification. Additionally, to show the personal relevance of a song/track, the size of the circles is influenced by a song's comparative importance as well as its overall importance (Figure 2.19).

This approach offers three types of interaction, namely, zooming and panning, filtering, and mouse hovering in order to impart supplemental information within the current encoding. The following, in brief, unfold the aforementioned interactions' merits and flaws:

- Although zooming and panning helps to increase the focus on the particular part of visualisation and also lessens overlapping, the visualisation still

suffers from overlapping as the zooming has no impact on the vertical axes – hours and minutes (time) – which might have a dense distribution of circle (songs).

- Filtering offers inputting a combined form of arbitrary terms in order to refine the visual outcome which may decrease the clutter and provide a better view.

- Mouse hovering highlights and illustrates potential song sequences by means of steady or dashed lines, and moreover provides additional information regarding the streamed track/song history in a form of a tooltip. However, linking the large number of related instances together by displaying a tooltip can lead to extra clutter and overlap which make the visualisation ineffective.

The advantage of this approach is that the vertical axes are used to display time, which provides the ability to collate daily streaming actions by setting them side by side. However, the following downsides are identified:

1. It is slow in fetching data, processing data, and encoding them to the visual form.

2. The visualisation, in general, is cluttered.

3. It is too dense as it shows all available years in one screen – considering the screen and pixel limitation – it does not fit the non-expert users' standard screen and low computing power.

4. The weighting algorithm is not effective and does not include users to decide relevance, frequency, and importance.

5. The designed interactions do not intimately support the visualisation due to the problem pointed out earlier.

AppInsight

AppInsight [23] demonstrates a visualisation tool that empower users to recall their past memories regarding general computer usage and gain relevant information. This approach uses three contextual cues (e.g. Windows title, url, and name of the program) to identify the user's actions on their PC or laptop together with progress over time (Figure 2.20). The main part of the visualisation are application (shows the most frequent applications used), hourly usage (displays any activity on an hourly basis), and usage evolution (indicates the overall user activity and its progress over the period).

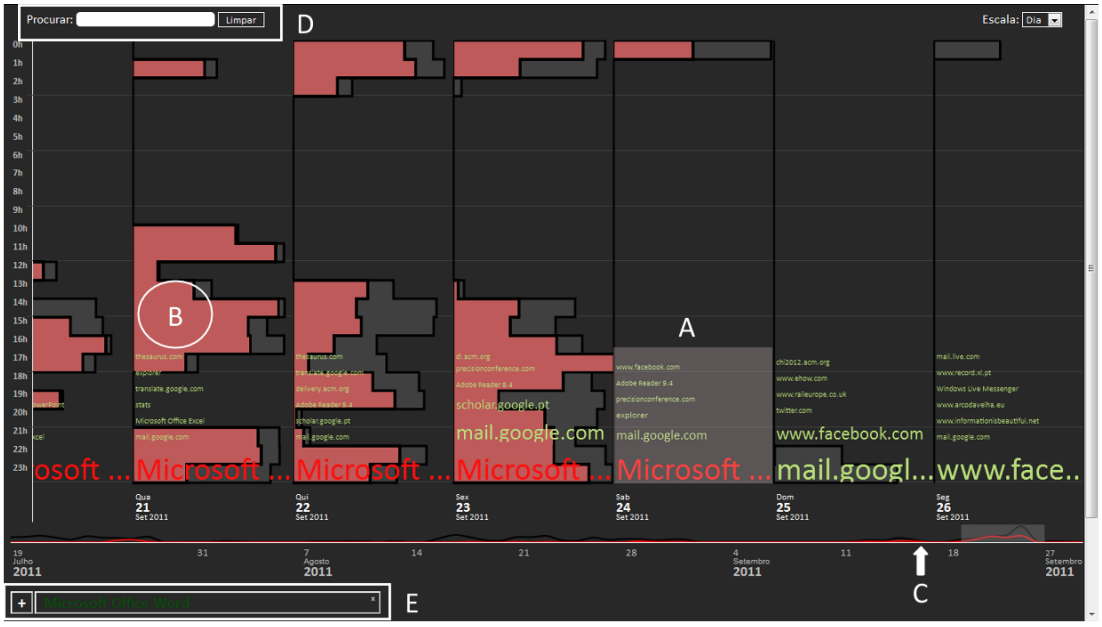


FIGURE 2.20: AppInsight [23] interface that shows computer usage over the time

The most used program in a day, month, or year is calculated by the following formula, where r is the usage ratio for program i and $n = 1, ..., N$.

$$r(apps[i]) = \frac{usage_p(apps[i])}{\max(usage_p(app[1]), ..., usage_p(app[n]))} \tag{2.10}$$

The overall usage of program i in a certain period p is calculated as follows:

$$overallusage(p) = \sum_{i=1}^{n} usage_p(apps[i]) \tag{2.11}$$

This approach can help the process of recalling the past, examining digital histories, and portraying productivity and work behaviour of the user. Nonetheless, this tool does not allow for further exploration of distribution of usage within every hour.

Interactive Music Exploration

Dias et al. [60] introduces a visual analytics tool that allows browsing and exploring streaming music history in order to discover any patterns, habits or interesting moments by offering a multi-facet timeline, filtering tool, and interaction (Figure 2.21). In this technique, the streaming history maps on a timeline-based visualisation to reflect the structure of the data, which is inherently based upon time. The main body of this visualisation tool consists of different sizes of stacked circles in individual columns lying down on the timeline. The columns represent the time intervals, while the stacked circles encode the data element, e.g. artist, track, or album. In addition, there is a histogram in the background which indicates the overall frequency of streaming music over the same period. Interaction includes brushing and highlighting techniques by which additional textual information is shown regarding the selected point together with the linkage between the similar – in this case, songs – circles to help the user track the listening trend. The filtering allows for limiting the result and focusing on interesting parts based upon user interest.

This technique shows a good result from its evaluation and these show that the developed application is easy to use as the users completed the given tasks with minimum error rate and considered the experiment beneficial. The authors stated that using this approach can facilitate the process of remembering in this context. However this approach – in my opinion – has a series of flaws

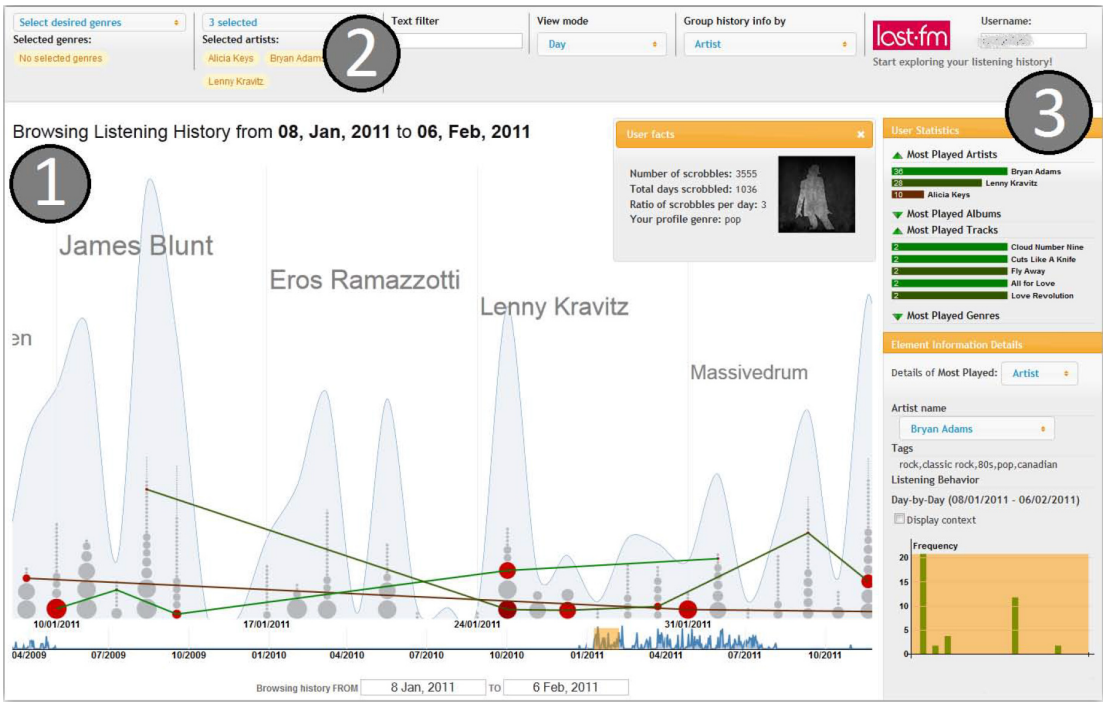


FIGURE 2.21: Dias et al. [60] technique uses multi-faceted visualisation to provide statistical or factual information along with the facility to filter and see the result interactively.

in analysis and visualisation. Thus, the analysis algorithm is not capable of finding any special streaming condition that might be interesting or informative within the history of streaming. The filtering does not have any influence in the process of analysing the history. The filtering is rather scattered, without robust structure. The visualisation suffers from measurement guides on the side which prevent understanding of the height of the histograms as well as stacked circles. Interaction makes the visualisation cluttered and as a result blocks some parts of the visualisation viewport by providing informative tooltips.

Visual Mementos

Visual Mementos [204] analyses and visualises personal movements at different temporal and spatial scales via an integrated timeline including a map, photos, etc., followed by semantic clustering of GPS logs to support reminiscing for self-reflection and memento sharing (Figure 2.22). This approach uses different sizes of

circular map segments along a time axis in chronological order to represent visits or repeated visits and associated duration within a geographical area. The size of each circular map is directly related to the duration of the visits and can be resized by adjusting the time interval. Visual Mementos can present familiar places and allows for creating and sharing a visual memento. The approach follows two main goals: reminiscing and sharing of personal experience. This work introduces five different case studies to examine their approach. The case studies are collected by the authors based on the participant personal data experiments. The use cases include mementos of short trip, activity, everyday life, multiple trips, and historic memento. The results of the evaluation are positive but only descriptive. However, as is evidenced in the paper, one of the drawbacks of this approach is that it is particularly designed for travel memories. This means that the approach is not capable of assisting its users to explore the everyday movements or a combination of travel and life logs, and hence, delivers inappropriate significant moments.

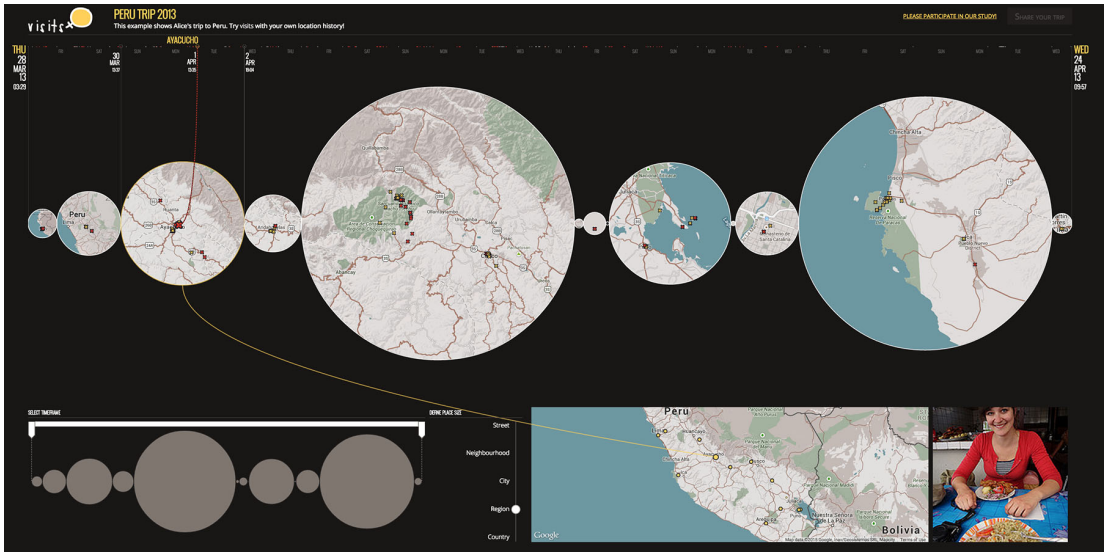


FIGURE 2.22: Visual Mementos [204] main interface with a number of circular maps over the timeline showing the trip history

2.3.4 Spatiotemporal visual analytics

Most of the work in the spatiotemporal visual analytics field is related to visual analysis of massive trajectory data [11, 13–15, 26, 45, 73, 119, 149, 190, 197, 215].

Andrienko, Andrienko, Bak, Keim and Wrobel [12] provide a detailed description of work on spatiotemporal visual analytics. This work identifies a need for novel visualisation methods with tightly integrated algorithmic data analysis to extract meaningful knowledge from such data as the current visual analytics are not capable of handling the expansive challenges within the movement data. The authors show that their approach – visual analytics of movement data – can provide significant understanding about movement behaviour and events that have occurred. Krueger et al. [119] introduce context data into trajectory data analysis and their main work is to find potential places from a large group of trajectory data with respect to the available POI information and analyse the movement behaviour. They use a density-based clustering technique to extract frequent destinations. Foursquare APIs are employed in their work to enrich the destination semantically (Figure 2.23).

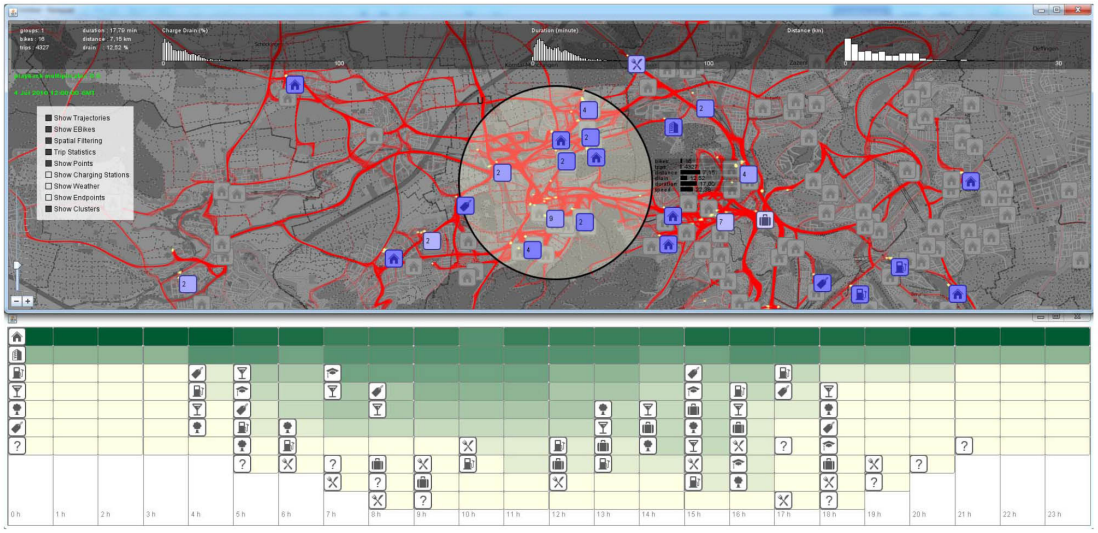


FIGURE 2.23: A temporal view to envisage the temporal daily pattern; Visualisation of frequent destinations including clusters, routes, and POIs on the top [119].

Yu et al. [238] introduce iVizTRANS to distinguish home and work places from public transportation data by incorporating a combination of visual analytics and machine learning methods. Von Landesberger [216] combines spatial and temporal simplifications for graph-based visual analysis to analyse the mobility data and support decision makers. Beecham et al. [26] study commuting behaviour by

visually analysing the London Cycle Hire Scheme data by developing a classification technique including a kernel density-estimation, a weighted mean-centres calculation, and spatial k-means clustering to group the behaviour of such data (Figure 2.24).

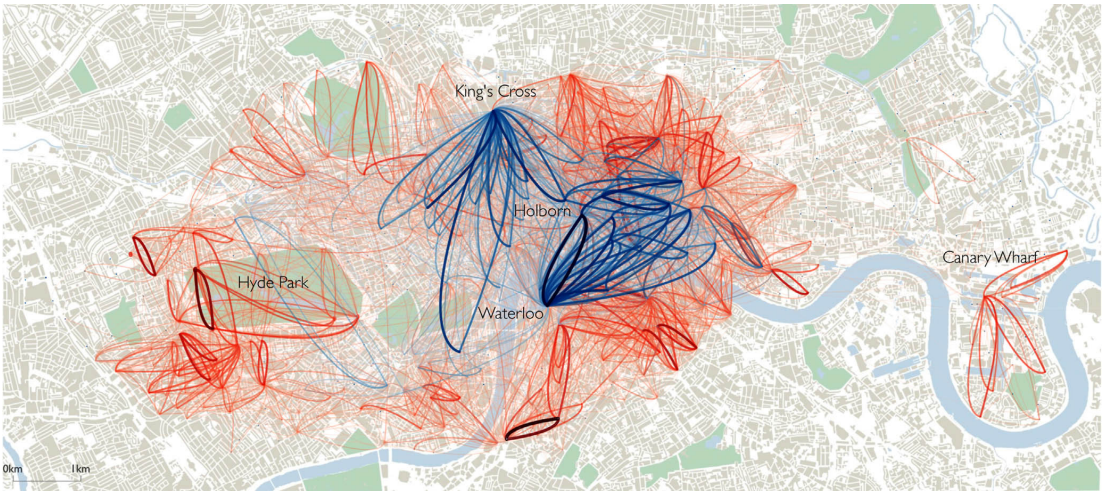


FIGURE 2.24: Cyclist's commuting journey in London. Flow lines appear with calculated journey frequency weight [26].

Chen et al. [45] discover movement patterns by analysing geo-tagged social media data of a large population via an interactive visual analytics approach that supports sparse trajectory data.

2.4 Evaluation in Visual Analytics

Evaluation in the field of visual analytics is rather intricate as it involves accessing the visualisation and, furthermore, all the related gears and processes (e.g. collaborative analysis, preliminary data analysis and data reasoning, imparting visualisation, etc.) that support the method [178, 210]. Numerous challenges have been identified by researchers whilst preparing, conveying, and implementing an evaluation of visual analytics approaches. Determining the appropriate evaluation method can be difficult as it requires selecting suitable tasks, well-suited questionnaires, appropriate users, and the right time to conduct it. There is varied literature regarding evaluation, particularly focusing on how to carried it out but

without any advice on when. In this work based on the research methodology, I follow the compound design stage strategy to asses the proposed approach.

2.4.1 Evaluation scope

The process of evaluation is not limited to the particular analysis or representation. The evaluation outcomes are vary and can be very general or overly specific to a visualisation approach [31, 124, 178, 210]. According to Lam et al. [124], the evaluation can be conducted at different phases within the visualisation process:

1. **Precondition**: to identify the gap and potential users with the environment of work;
2. **Design**: to identify the appropriate design including the visual encoding and interaction derived from human perception and cognition;
3. **Prototype**: to build and examine a visualisation prototype to observe how the design fulfils its goals and requirements by comparing with the state-of-the-art methods;
4. **Deployment**: to determine the effectiveness of the visualisation and its process in practice; and
5. **Redesign**: to enhance the design, functionality, and usability of the visualisation based on the problems identified within the process of evaluation.

2.4.2 Evaluation methods

Lam et al. [124] introduce seven scenarios for evaluating visual analytics approaches. The most relevant evaluation scenarios are outlined here:

- **Evaluating visual data analysis and reasoning**: is to examine visual analytics approaches' outputs, which can be quantifiable metrics (e.g. a

number of clues or insight by using the analysis) or subjective feedback (e.g. the quality of the analysis). The questions in this type of evaluation are related to the data exploration, knowledge discovery, decision making, and hypothesis generation.

- **Evaluating user experience**: is used to determine the reaction of the user to the visualisation approach over a short or long period and to inform the design. The visualisation approach here can similarly be referred to as an initial sketch, an initial working prototype, or a finalised approach. This type of evaluation includes a series of questions regarding useful features, missing features, perceptibility, and adoption.

- **Evaluating visualisation algorithms**: is used to examine the quality and performance of the algorithms used in the visualisation process. The main goal of this evaluation is to investigate the algorithm according to data size and complexity. This evaluation includes a set of questions with respect to pattern quality, meaningful representation of the underlying data, cluttered view, performance, and scale.

Recently, Few [75] has proposed an interesting effectiveness profile for data visualisation approaches which gives seven criteria in two categories: criteria that are related to the process of understanding (informative) and criteria that are affiliated with the process of offering practical emotional response (emotive). The criteria are:

- Informative
  - Completeness: the approach should include all the information that is required to obtain the level of understanding of the visualisation by providing pertinent context;
  - Perceptibility: the approach should be perceived with minimal effort and acceptable clarity;
  - Truthfulness: the approach must provide accurate and valid information to the user;

  - Usefulness: the approach should communicate information with some value and importance rather information that does not add any value;
  - Intuitiveness: the approach should deliver the information in an easy-to-understand way that the user is familiar with. Providing a novel and unfamiliar visualisation should not be more difficult to understand than the familiar methods. It also requires incorporating simple guidance with minimal learning.

- Emotive
  - Aesthetics: the approach should attract attention and by providing acceptable visual properties and user interface to its users.
  - Engagement: the approach, as whole, should have adequate quality to draw the audience into the data and exploration.

2.5 Chapter Summary

In this chapter, the most relevant work to this research in data mining (semantic enrichment and significant ranking), interactive data visualisation, spatiotemporal visualisation, and evaluation of visual analytics methods has been reviewed.

By reviewing the underlying work in semantic enrichment and significant ranking, it is found that this area is still in need of refinement as the current works can only enrich data with certain top-level information – a good start, but still not useful for some of the processes. Similarly, significant ranking models are rather simple and do not involve external factors or user preferences within their calculation. This is another gap that needs addressing.

Three major areas of visual analytics have been reviewed, namely, time-oriented data, personal data, and spatiotemporal data. Despite the notable contribution, the majority of these methods lack scalability, robust analytical reasoning, and clarity. Addressing these challenges can greatly strengthen the visualisation of large-scale personal data.

Gainful evaluation techniques that can be employed to assess the proposed approach for the visualisation, user experience, and effectiveness have been studied.

Lessons learned from these works have been used to fill the gap by proposing an effective visual analytic approach.

CHAPTER 3

Research Methodology

This chapter provides an overview of the methodology employed, including all the techniques within the bounds of this research. The methodology of this research comprises five areas, namely, data acquisition, systematic literature review, design and prototyping, developing integrated visual analytics tools, and evaluation. The structure of the methodology and its utilisation is visualised in Figure 3.1. This chapter aims at depicting each area in detail.

3.1 Data Acquisition

This research is derived from personal daily life logging data. Hence, it is necessary to establish the potential sources of such data that can be obtained in conjunction with ethics and privacy concerns. In addition, the personal daily life logging data itself should include a set of required attributes such as timestamp, GPS coordinates, and the like for inclusion in this research. This thesis involves individuals in the process of the data acquisition in order to obtain real data and provide a beneficial approach that can be evaluated by relying on the participants and their established ground truth.

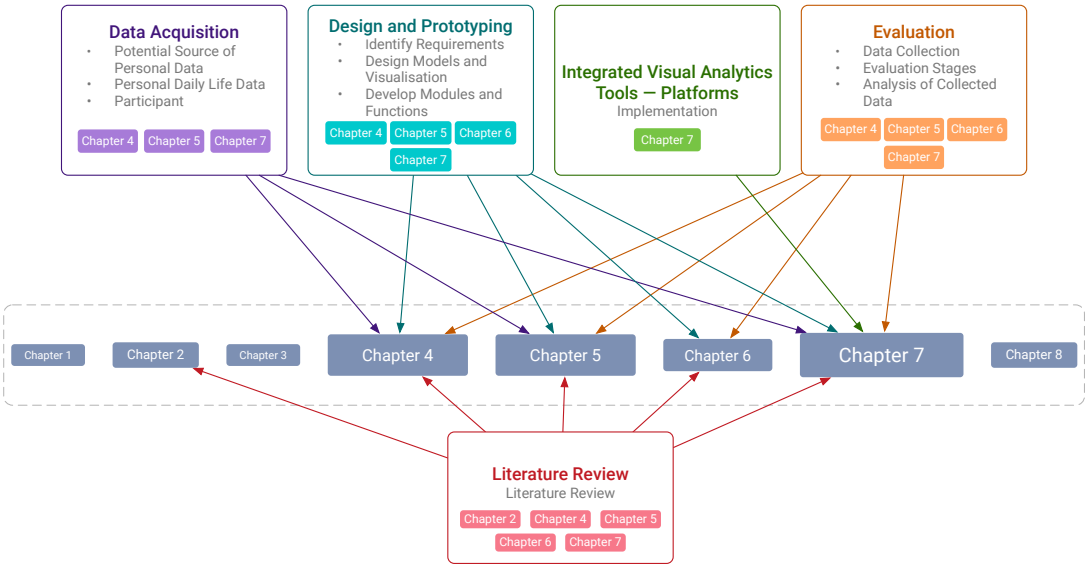


FIGURE 3.1: Research methodology application

3.1.1 Potential sources of personal data

Two popular wearable applications (*Fitbit Charge* and *Withings* O_2x) and two smartphone applications (*Moves* and *SmarTracker* [1]) are examined to ascertain which practice 1) fits better within the bound of the research requirements; 2) provides sufficient data with adequate quality; and 3) can be exploited with less complication (Figure 3.2).

The potential sources are compared in accordance with the definition of personal life logging data and its requirements as follows:

- contains movements (physical activities) and stops information;
- includes timestamp information including start and end times;
- holds essential metrics of the physical activities such as step counts and distance;
- carries a valid geographical coordinate including latitude and longitude; and

[1] SmarTracker is a smartphone application designed specifically for tracking daily life by our in-team research lab at the Centre for Visualisation and Data Analytics

FIGURE 3.2: Potential sources of data and their corresponding devices

- consists of duration of actions or any additional information (e.g. calories) and the name of the geographical coordinates (optional).

The two wearables, *Fitbit Charge* and *Withings* O_2x, provide corresponding temporal information such as step count, calories, distance, and elevation. The advantages of using these wearables are:

1. working independently with no need for internet connection or the like;
2. the built-in batteries last for 3-5 days;
3. data can be synced automatically via Bluetooth technology;
4. data is accessible via APIs; and
5. more accurate data as a result of using specific sensors and algorithms.

However, these wearables cannot recognise the type of the physical activities nor geographical location in any circumstances; see the sample data in Snippet 3.1 and 3.2. This means that the collected data lack movements information that is an essential part of daily life and can be used for gaining valuable knowledge.

```
{
    "date": "2015-11-02",
    "local_id": "7cfe069f-09ab-4889-91ef-2c8c5544b13a",
    "summary": {
      "source": "fitbit",
      "steps": 1320,
      "calories": 230,
      "calories_bmr": 1913,
      "calories_out": 2143,
      "elevation": 48.77,
      "soft_activity_minutes": 310,
      "moderate_activity_minutes": 620,
      "intense_activity_minutes": 0,
      "sedentary_minutes": 1166,
      "activities": [null]
    }
  }
```


Snippet 3.1: An example of JSON data from Fitbit


```
{
    "date": "2014-09-12",
    "local_id": "8c9adc1d-9453-4180-a750-47ddac842363",
    "summary": {
      "source": "withings",
      "distance_meter": 5783.6,
      "steps": 7065,
      "calories": 242.61,
      "elevation": 32.2,
      "soft_activity_minutes": 40,
      "moderate_activity_minutes": 30,
      "intense_activity_minutes": 2
    }
  }
```


Snippet 3.2: Example data from Withings

In contrast, the *Moves* and *SmarTracker* applications automatically record physical activity metrics (e.g walking, running, step count, calories, distance), and most importantly movements by using the smartphones' built-in sensors. The only difference between these two applications is the use of different underlying algorithms to determine the type of activity and GPS coordinates. The strong points of using these applications are:

1. Smartphones are carried almost everywhere by the user;
2. Both applications work in the background with no need for initialisation;
3. They recognise the user's activity types automatically (and allow modifying them);

4. They record the user's movements and stops via built-in GPS together with activities;

5. They provide a high level of accuracy by using GPS and WiFi; and

6. They provide APIs to access the raw data.

By comparing the data provided and advantages of each type of tracker – wearables and smartphone applications, the aforementioned smartphone applications were selected to collect life logging data towards gaining meaningful knowledge and providing an effective visualisation of personal daily life.

3.1.2 Personal daily life data

As mentioned in chapter 1, personal daily life data is the series of temporal logs captured by mobile and/or wearable technologies in an ordinal form. These data incorporate different aspects of individual life including physical activities, movements, diet, health, and the like.

In this research, according to the comparison of the potential data sources, two mobile applications – *Moves* and *SmarTracker* – are employed to automatically acquire personal life logging data including coordinates and movement data on a daily basis. The data from these application include:

- GPS location data such as manual place annotation, geographical coordinates, duration, and time-stamp information;
- Temporal data of movement such as walking, running, cycling, and transport with GPS tracking points, step count, calories, and distance.

The raw data can be retrieved in a standard format such as JSON or XML via the provided APIs. In this research, the APIs are used to retrieve the data in JSON format. The data is similar to the Snippets 3.3 and 3.4.

```
{
    "place": [{
        "name": "unknown",
        "Id": "GB232049236409273402",
        "centerPoint_lat": 51.9477599660822,
        "centerPoint_lon": -0.2811089100338727,
        "start_time_milliseconds": 1493593267292,
        "end_time_milliseconds": 1493676989595,
        "duration_seconds": 83722,
        "history_centerPoints": []
    }],
    "movement": [{
        "group": "Walking",
        "trackPoints": [{
                "lat": 51.9478026,
                "lng": -0.2811036,
                "accuracy": 18.39299964904785,
                "time_milliseconds": 1493670425674,
                "predict": false
            },
            {
                "lat": 51.9478026,
                "lng": -0.2811036,
                "accuracy": 18.39299964904785,
                "time_milliseconds": 1493670425674,
                "predict": false
            },{"..."}
        ]
    }],
    "activity": [{
        "steps": 1540,
        "calories": 64.975,
        "duration_seconds": 1320
    }]
}
```


SNIPPET 3.3: An example of JSON data from MyHealthAvatar tracking application – SmarTracker

3.1.3 Participants

In general, the participants for the evaluation process were selected within the University from a range of in-team colleagues to academics and students in different departments.

All the participants who contributed to this research were provided with a consent form and also were free to withdraw at any time without giving any reason by contacting the author. Participants, also, had the right to withdraw retrospectively any consent given, and to request that any data gathered on them be destroyed.

Two types of participants took part in this study: participants with and without personal daily data. Participants without daily life data took part only in the

```
[{
        "source": "moves",
        "type": "place",
        "date": "2015-10-23",
        "start_time": "2015-10-23 20:38:51+0100",
        "end_time": "2015-10-24 07:47:55+0100",
        "place": {
                "local_id": "00000000",
                "name": "unknown",
                "location": {
                        "lat": 51.8937338412,
                        "lon": -0.4229268157
                },
                "type": null,
                "foursquare_category_ids": [null]
        }
}, {
        "source": "moves",
        "type": "movement",
        "date": "2016-10-24",
        "start_time": "2016-10-24 18:14:20+0000",
        "end_time": "2016-10-24 18:46:34+0000",
        "activities": [{
                "activity": "cycling",
                "activity_group": "cycling",
                "duration_seconds": 1857,
                "distance_meter": 2431.0,
                "steps": 0,
                "calories": 290.0,
                "track_points": [{
                        "lat": 51.8779746192,
                        "lon": -0.4122447968
                }, {
                        "lat": 51.877854973,
                        "lon": -0.4118430306
                },{"..."} ]
        }]
}]
```


Snippet 3.4: Example of personal life data from the Moves application

process of evaluating the visual component designs and the platforms that are made for general purposes by using a synthetic dataset. Figure 3.3 shows more details of the participants.

3.2 Literature Review and Investigation

A complete review of visual analytics terms has been made to comprehend how the current visual analytics approaches and their components including data mining, visualisation, and interaction support humans to grasp better understanding of data. The literature review is organised into three main parts: 1) data mining

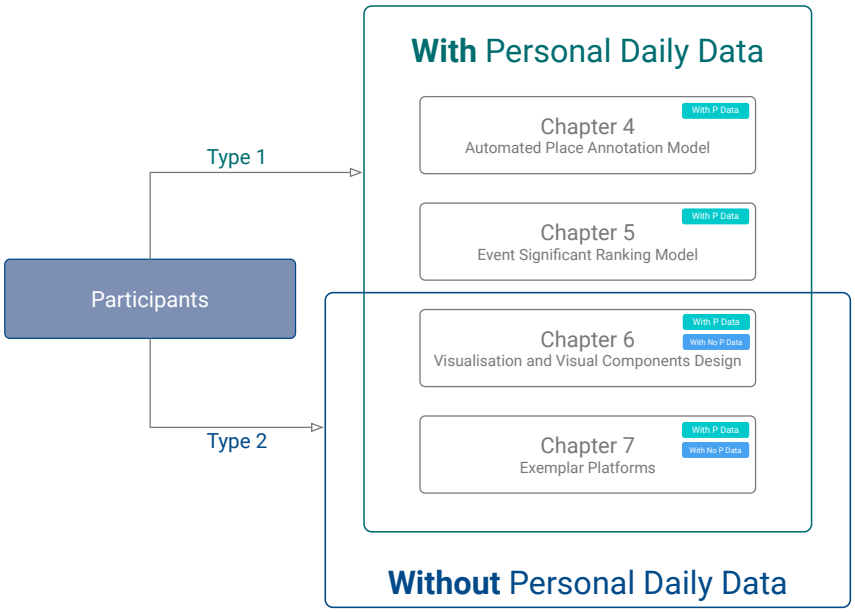


FIGURE 3.3: Two types of participants for the evaluation in this research

with particular focus on semantic enrichment and significant ranking; 2) data visualisation with focus on time-oriented, personal, and spatiotemporal data; and 3) visual analytics evaluation methods to identify the gaps and exploit the relevant terms and findings in the field of visual analytics to strengthen the proposed approach in this work.

3.3 Design and Prototyping

This research follows the agile procedure for designing the models and visual components. This process comprises three stages:

1. identifying the requirements;
2. designing the data mining models and visualisation approach including the visual components with respect to the requirements followed by prototyping; and
3. evaluating the design.

Subsequently, two prototyping approaches – concept prototype and vertical prototype – are employed [169] firstly, to encapsulate the overall vision with regards to the design, architecture, and functionality, and secondly, to assess the quality and usability of the approach implemented.

The concept prototype approach is typically initiated after setting out the architecture. It represents the essential design of the user interface, scope, and its related components. In principle, this approach can help this research to identify any limitation or problem during the process of design and evaluation.

The vertical prototype approach is able to demonstrate the working mock-up of the proposed technique with real data at an early stage in which some of the features may not be fully functional. This approach allows testing and improvement of the design and the main features of the implementation in order to adjust and accommodate all the requirements.

This work complies with the following steps to create a vertical prototype:

1. identify the required core functions of the approach which need to be included in the prototype
2. define how the prototype data source is used
3. create a temporary structure
4. determine and design key modules and core functions
5. develop the identified modules and functions based on the design.

Several scripting languages are exploited such as JavaScript, D3.js, and jQuery to develop the standalone prototypes. Correspondingly, a local development environment (Docker) is employed to host and test the prototypes.

3.4 Visual Analytics Tools (Platforms)

Several platforms are made available within the progress of this research to exemplify and study the integration and effectiveness of the novel data mining models and visualisation for different purposes within the domain of personal life. Each platform exhibits an independent structure to fulfil the requirements. These platforms in general provide the following:

- introduce a standalone platform as a sandbox that incorporates the novel data mining and data visualisation towards the process of data exploration and knowledge discovery;
- show the integration of the models and visual components in a single place;
- evaluate the integrated novel data mining and visualisation as a whole to assist analytical reasoning and support knowledge discovery in the personal life domain by measuring the accuracy, usability, and effectiveness; and
- obtain qualitative and quantitative feedback on novel approaches and their features.

The platforms are designed and implemented as an online working prototype (sandbox) for the process of test and evaluation. The implementation of the platforms follows the same procedure and is described in the following subsection.

3.4.1 Implementation

The platforms are developed as a web-based platform with similar back-end and front-end structure. Each platform has the logical structure to read the data from the database, process the data on the back-end as well as on the client side (on the fly), applying the data mining models to the data, and encoding the knowledge visually. The technologies that are employed to make this project possible are depicted in the following paragraphs.

All the platforms are implemented under MyHealtAvatar. The server side is built using Java with the Spring Framework web technology stack, and it is deployed on Tomcat 7 on Ubuntu 14.04. MySQL is used for data management while Apache Cassandra is utilised to store all the user's activity; the design of the two data stores takes into account data security and scalability. The data transfers from third parties to MyhealthAvatar services, and between MyHealthAvatar services and the mobile application, are protected using OAuth 2 Over HTTPS, which is also the standard protocol used by large corporations in data transfer, including Google, Facebook, and Twitter. The HTTPS protocol is enforced for the whole platform to protect the user from misbehaviour such as eavesdropping. Authentication and authorisation for user data access is built on the Spring Security framework, which is the most widely used Java web security solution and tested by millions of users around the world.

The data mining and data analysis are all implemented in JavaScript [2]. The raw data is queried via the APIs and fed to the data processing modules. The data mining is implemented with scalability in mind for the large amount of personal data. This ensures that the process can be run within the client browser.

The front end UIs are implemented in HTML5, CSS and JavaScript; the main libraries that are utilised to create the user interface and the interactive visualisation include Bootstrap [3], jQuery [4], D3.js [5], leaflet.js [6] and Google Maps JavaScript API [7].

3.5 Evaluation

The evaluation is conducted on the data mining models, the visual component designs, and the platforms, independently. The process of evaluation for each part

[2] https://developer.mozilla.org/en-US/docs/Web/JavaScript
[3] https://v4-alpha.getbootstrap.com/
[4] https://jquery.com/
[5] https://d3js.org/
[6] http://leafletjs.com/
[7] https://developers.google.com/maps/documentation/javascript/

is slightly different owing to the nature of the approach. The data mining models are examined in terms of accuracy and performance while the visual components are assessed to measure the constructiveness of the represented information. And lastly, the platforms which incorporate the data mining models and information visualisation – as complete visual analytics approaches – are evaluated to determine the effectiveness, usability, and accuracy.

3.5.1 Data collection

In general, a series of surveys and interviews were used within the visual component design and platforms to obtain the users' opinion regarding the interface, usability, and overall effectiveness of provided visual components.

The surveys embrace different types of questions including multiple choice, multiple answers, Likert-type, scale-type, and open-ended questions. Multiple choice, multiple answers, and open-ended questions were used to collect the participants' qualitative opinions whilst the 5-point Likert-type and ranking questions were used to obtain the quantitative measures of effectiveness and usability. The 5-point scales vary based on the question formulations, i.e. (1- very poor), (2- poor), (3- normal), (4- good), and (5- very good) or (1- strongly disagree), (2- disagree), (3- neutral), (4- agree), and (5- strongly agree).

Moreover, a number of tasks are designed in accordance with the design goals and requirements of the platforms to assess their fulfilment and measure the effectiveness of each function and visual component. Furthermore, two standard questionnaires were used – User Interface Satisfaction [89] and IBM computer usability satisfaction questionnaire [130] – across the evaluation. The effectiveness of the platforms were measured based on Few [75].

Two online survey platforms – BOS [8]) and Google Forms [9] – were utilised to collect and store the evaluation data.

[8] https://www.onlinesurveys.ac.uk/
[9] https://www.google.co.uk/forms/about/

3.5.2 Evaluation stages

3.5.2.1 Data mining models

The data mining models were assessed using a set of datasets with known ground truth to determine the accuracy of the results. The results from the models were compared against the acquired ground truth (from the participants with the personal life logging data) to examine the precision. Additionally, the models and their implemented algorithms underwent benchmark testing to work out the performance, efficiency, and scalability. The performance test plugged different sizes of datasets to the implemented algorithms and measured the performance whilst the efficiency test examined the achievable operations per second for each algorithm. The former test was developed as a built-in part of the implemented models while the latter was adopted – (jslitmus.js)[10], a JavaScript library – to assess the models.

3.5.2.2 Visualisation design

The visualisation design and its components were evaluated in line with user centred design and agile processes[117, 186]. The iterative evaluations targeted user centred design goals in which users' opinions and preferences are embodied in the process of design and implementation of the approach to determine the best practice in representing the extracted knowledge and information. The evaluation, in the main, determines the accuracy, perceptibility, design, and intuitiveness, respectively. To this end, a series of evaluations were conducted via a web-based prototype which enabled the user to work and feel the visualisation in practice. An unsupervised environment was used for each round of the evaluation to allow participants to perform the tests with no influence which might result in providing non-credible feedback. The process of evaluation includes the following stages:

- a brief introduction of the visual component and its meaning;

[10] https://github.com/broofa/jslitmus

- data collection via the online questionnaires and interview; and
- result analysis, discussion, and improvements.

The results are used to improve the design of the components in a different stage of this research.

3.5.2.3 Visual analytics tools (platforms)

The platforms are examined based on the visual analytics' effectiveness, accuracy, and user experience. Assessing the visual analytics' effectiveness allows investigating how the visual analytics tools can substantiate the reasoning and visualising of personal life data by providing quantifiable metrics (e.g. quantity of the insights) or subjective feedback (e.g. quality of the approach on each platform). Subsequently, assessing the accuracy is based upon the task completion result in which the output is, almost always, numerical and can be studied via descriptive statistics. The user experience involves understanding individual reactions to the visualisation tool by perceiving accuracy, efficacy, and efficiency.

Two types of evaluations – iterative design and conclusive design – are carried out within each platform in this work based on [32, 75, 124, 130, 227, 228]. In the iterative design evaluation, participants were asked to use the online prototype to form an opinion about the approach and complete the online questionnaire. The questionnaire was designed as part of the user centred design approach to enhance the design of the visual components, interface, and functionality. The conclusive evaluations were carried out as a final evaluation using a task-based questionnaire, a usability survey, user interface satisfaction, and a set of individual interviews. The participants were asked to complete a number of typical tasks that reflected the design goals and requirements of the related visual analytics platform.

The process of evaluating the platforms consisted of four main parts:

- A brief introduction to the platform and short tutorial regarding the interface and its functionalities. Subsequently, the participants were given free trial

time to familiarise themselves with the approach provided and then asked to complete the given tasks.

- Data collection via the online questionnaires including the tasks, usability, user interface satisfaction, system reliability, and effectiveness questionnaires.
- A number of participants with different backgrounds were interviewed to understand their opinions about the approach and the level of knowledge they had gained during the process.
- Results analysis and discussion together with elucidating the collected data.
- Describing limitations, lessons learned, and future work.

3.5.3 Analysis of collected data

The collected data from the questionnaires contained quantitative and qualitative information in the form of multiple choice, multiple answers, ranking scale, Likert-type, and open-ended questions. The quantitative data was analysed to measure the accuracy, satisfaction, and effectiveness of different components or platforms within the bounds of this work.

To analyse the Likert-type data, the responses ranging from 1 to 5 were mapped to a 3-point scale where Neutral = 1 (middle), Strongly disagree or Disagree = 0 (negative), and Strongly agree or Agree = 2 (positive). In addition, to test a hypothesis, the Likert scales were transformed to the nominal level by merging the responses into accept/agree and reject/disagree.

Correspondingly, the task completion results were analysed by comparing the results against the known answers. Tasks with two parts were considered as two individual sub-tasks with equal, individual 50% score. Moreover, open-ended qualitative questions were analysed carefully to determine the soundness of the platforms. Similarly, the interviews as an additional means were studied to assist with interpreting the collected data. The contents were used to modify or enhance parts of the approach.

3.6 Chapter Summary

In this chapter, the overall methodology utilised by this work has been described. The nature of personal life logging data, potential sources of acquiring such data, and the participants were depicted in detail. Subsequently, the organisation of the literature review and investigation into current works as the important part of this research was explained. Correspondingly, the process of design and prototyping, implementing the platforms (case studies), and evaluation stages within this work are completely portrayed.

CHAPTER 4

Automated Place Annotation with Multi-level Probabilistic Latent Semantic Analysis

4.1 Introduction

The personal daily life data in this research embrace beneficial temporal and trajectory details that can be exploited to discover invaluable knowledge about individuals. However, these data encapsulate inadequate semantic information such as place name, category, or the like that are highly valuable for the process of knowledge exploration. Therefore, this yields a need to ameliorate the trajectory data by conducting an automated semantic enrichment process. In general, semantic enrichment is utilised in conjunction with the process of movement analysis to investigate users' behaviour towards more valuable knowledge discovery and reasoning about the life pattern [9, 93, 120, 120, 156, 173, 173, 174, 236]. Nevertheless, the amount of research on semantic trajectory enrichment is not sizeable whilst various works can be found on the trajectory data and sequence of movements.

Semantic enrichment is a process of feeding contextual knowledge either dynamically or manually into the raw trajectories. Consequently, this information assigned to the raw data is called an annotation. An annotation can be attached to three different levels of detail – to the entire trajectory as a whole, to significant or specific parts of the trajectory (episodes), or to particular positions [23, 151, 204]. For instance, identifying a goal of a movement or a trip (e.g. vacation) is a trajectory-level annotation which holds a single value for the specific part of the trajectory. An example for a position-level annotation, which holds one value per position, is to assign the user's visited place to the exact position within the trajectory (e.g Museum, University, Restaurant, etc.). The annotation value can be a simple attribute such as a string or, in a more complex form, a list of attributes and links such as an object of arrays consisting of name, category, website, email, and the like.

Contextual information that can be used as annotation, nowadays, can be acquired by widely known online platforms such as Foursquare [1], Google Place [2], and Facebook [3]. These platforms provide Point Of Interest (POI), a complementary information related to various places and attractions. This information comes in the form of a POI that can be used as a powerful source for obtaining supplementary information, supporting mobility, and behavioural analysis. Therefore, each unlabelled data point within the trajectory data can be considered as an interesting point and identifying the relevant POI to this point can contribute to discovering what type of POI that motivates the stop [11, 15, 41, 82, 119, 120].

The goal of the automated place annotation, in this work, is to attach a multi-level semantic information into the users' unidentified trajectory data by determining pertinent POIs and computing the probability value of the nearest neighbourhood via an incremental probabilistic latent semantic analysis purely based on the historical location data and prior-knowledge of each individual person without requiring extra geographical information or any annotation previously shared by other users. The users can obtain the annotation information entirely based on

[1] https://foursquare.com/
[2] https://developers.google.com/places/
[3] https://developers.facebook.com/docs/graph-api/reference/location/

their own private data without sharing any information as in many commercial apps, which imply a significant reduction of potential ethical and security risks. The multi-level semantic information is referred to a list of annotation composed of information ranging from top-level annotation such as country or town to highly detailed annotation at fine grain such as place name. Thereafter, the result from the automated annotation contributes to the process of ranking significant events for personal life data in this research (Chapter 5).

The hypothesis for the automated place annotation is that the process of semantic enrichment needs to incorporate extra factors such as category probability, distance, time, and prior knowledge to achieve an adequate level of accuracy for annotating the unknown coordinate. This hypothesis is tested in the evaluation section and establishes evidence about the model.

The automated place annotation using multi-level probabilistic latent semantic analysis, in brief, checks and pre-processes the preliminary data from the corresponding databases, retrieves POIs, performs a density-based clustering, and determines the probability value of each candidate place by performing the latent semantic analysis incorporating a distance likelihood, user profile, place traction and category, altogether with the historical data. The process iterates after new points are added to the database in order to incrementally refine the probability values of the annotated places and hence improve the result. The output of the process consists of a multi-level information including the most probable candidate places, categories, street addresses, cities, and countries. Providing this list allows examination of the model and rectifying inaccurate annotation by the user during the evaluation process.

The main contribution of the automated place annotation is a novel multi-level probabilistic latent semantic analysis model, which exploits different factors such as distance, prior knowledge, and density clustering to automatically determine a list of foremost viable places including the name, category, street, city, and country. More specifically this model includes:

- Multi-level semantics to provide different layers of semantic information that can be used for different purposes such as sequence or pattern mining of spatio-temporal data.
- Probabilistic latent semantic analysis (PLSA) to compute the probability of the candidate places and their co-occurrence as a combination of autonomous multinomial likelihood by including the distance, frequency, place traction, and category voting.
- Bayesian computation of probability based on prior knowledge from user profile and personal conditions related to the place.
- Incremental computation to update the user profile and improve the result with respect to new entry data and convergent prior knowledge.

The remainder of this chapter is organised as follows. It first looks at the challenges and obstacles within the semantic enrichment of spatio-temporal data by means of POIs. Then, it describes the novel automated annotation model including data preparation, user profile, and clustering of data points. Next, the novel automated annotation model for determining the foremost probable place is depicted. And, in the end, the evaluation result based on real-life personal data is presented.

4.2 Challenges and Obstacles

To identify the foremost likely places for unknown points and add pertinent information to them without involving any labelled dataset, there is a need for semantic enrichment similar to the work discussed in the literature review. The process of semantic enrichment in these works rests on retrieving the POIs via measuring the distances, and number of check-ins and users. However, the following obstacles are involved within this process:

- **GPS accuracy**: The accuracy of the GPS data recorded via smartphones, particularly in indoor places, is low due to the lack of satellite signal and the device battery life management.
- **Dense area**: Query the nearby POIs in dense areas as a town centre would return a tightly packed list of places with almost identical distance, category, or address to the queried coordinate.

These obstacles can reduce accuracy and prevent the model from producing a sensible result, which does not comply with the automated annotation goal. A simple case of this situation is when the user works at an indoor office such as a university located in a town centre with a large number of POIs around, such as shops and restaurants. Querying the nearby POIs and calculating the probability values similar to the previous works shows that mimicking the previous methods cannot yield an accurate result including the place name and category due to the inaccurately recorded GPS data as well as the dense number of POIs. This results in an incorrect interpretation of the place which the user, in reality, has visited and hence an inaccurate annotation.

As a consequence, there is a need for a novel annotation to determine the most likely places as candidates by formulating a comprehensive model and including external factors in calculating the probability values. This enables the model to eventually assess the relevant places more attentively and returns higher quality results. To this end, automated place annotating with multi-level semantic using probabilistic latent semantic analysis that can provide annotation at five different levels is proposed, and this incrementally extends the process for new points and improved results. A hierarchical representation of semantics and details of the method is described in the following sections.

4.3 Automated place annotation model

According to the current challenges and obstacles in annotating the trajectory data, this thesis proposes an automated and incremental place annotation that is outlined as follows:

1. The input for this method is a tuple (p_i, t_i), where $i \in \mathbb{N} = 1, ..., N$, is a sequence of unknown place and time. Each p_i is associated to a set of GPS locations (coordinates) l_k where $k \in \mathbb{N} = 1, \ldots, N$. This is due to to the noise in the GPS sampling. Additionally, the number of GPS locations for p_i may incrementally increase over the period of time, particularly if the location is revisited by the target user (e.g. home place, workplace)

2. Make use of the Foursquare POI APIs to obtain a list of nearby venues to the GPS location as candidates for the annotation.

3. Define an initial user profile with the ability to update.

4. Determine the cluster of trajectory data by using a density-based spatial clustering [69, 183].

5. Define the probability value by using the Bayes theorem. [200, 212]:

$$\begin{aligned} P(a_i^j|p_i, t_i) &= \frac{P(p_i|a_i^j) \times P(t_i|a_i^j) \times P(a_i^j)}{P(p_i, t_i)} \\ &= \frac{P(p_i|a_i^j) \times P(t_i|a_i^j) \times P(a_i^j)}{\sum_{a_k} P(p_i|a_k) \times P(t_i|a_k) \times P(a_k)} \end{aligned} \tag{4.1}$$

$$P(t_i|a_i^j) \sim \sum_{p_i} n(p_i, t_i) \times P(a_i^j|p_i, t_i) \tag{4.2}$$

$$P(a_i^j) \sim \sum_{t_i} \sum_{p_i} n(p_i, t_i) \times P(a_i^j|p_i, t_i) \tag{4.3}$$

where a_i^j is the j^{th} candidate annotation for p_i amongst the candidate set $j = 1, \ldots, k$, $P(p_i|a_i^j)$ is the probability of the targeted place according

to its distance, $P(t_i|a_i^j)$ denotes the related probability according to the time and the category from the user profile histogram extracted from the prior-knowledge (questionnaire), $P(a_i^j)$ is the place traction (popularity) and its category. Also, $n(p_i, t_i)$ is the total number of data points. Note that a_i^j contains $(a_i^{j_n}, a_i^{j_g}, a_i^{j_s}, a_i^{j_t}, a_i^{j_c})$ which represents the name, category, street, town, and country of a targeted place, respectively.

The reason that $P(p_i|a_i^j)$ and $P(t_i|a_i^j)$ are independent is that they are the conditional probabilities solely based on the given annotations (a_i^j) which are fixed with no reference to time; hence their joint probability is equivalent to the product of their probabilities. This means that despite the annotations (a_i^j) being fixed, their geographical GPS coordinates do change independently of time and thus can appear marginally different each time, mainly due to the noisy and erratic GPS data with no relevance to time at any circumstances [4]. Therefore, the model considers the probability of the distance between the recorded GPS coordinate (unknown place) and the candidate POIs (annotation a_i^j) regardless of its time, and then uses the temporal information to obtain the likelihood of the annotation within the same time-slot from the user profile, generated and updated from the prior-knowledge.

In addition, one of the strengths of the model is the user profile, which is initially extracted from the prior knowledge and then incrementally updated by using the annotation results. Utilising the prior-knowledge together with the historical locations greatly contribute to calculating a more realistic probability result for the related annotations by giving more weight to the related place that the users tend to be at the particular day of the week and the specific time [5]. Furthermore, the iteration of the model refines the user profile continuously which results in a more sensible probability calculation

[4] An unknown place and its coordinates can appear slightly different (within a threshold of approximately 20 to 45 meters) each time the person is in the same place. As a result, the exact same place (e.g. workplace) may have a large number of encircled coordinates close to it, recorded at different times.

[5] For instance, if the user works in a university located in a town centre with highly dense POIs, knowledge of the user's personal geographical movement behaviour can be used to prioritise the university and highly-related annotations over irrelevant places, and significantly improve

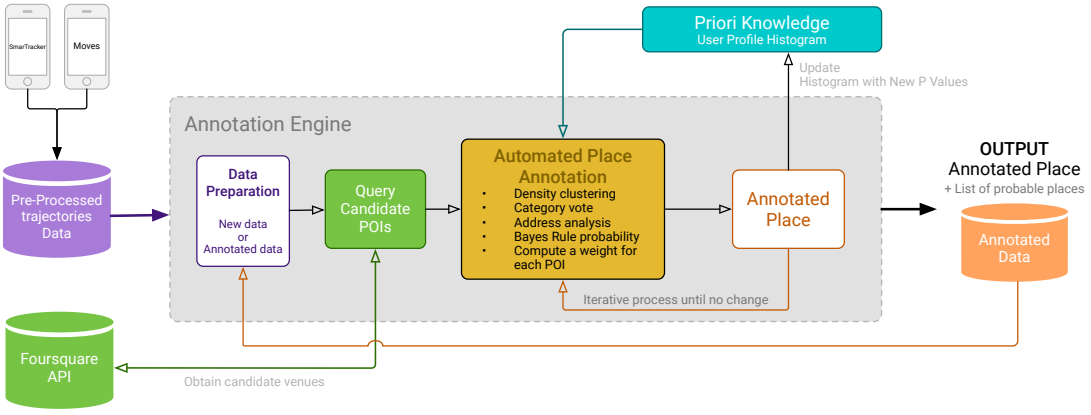


FIGURE 4.1: Automated place annotation architecture

purely based on the user historical places and prior knowledge. More details are given in Section 4.3.2.

- $P(p_i|a_i^j)$ is approximated with the calculation of the haversine distance between the average location of p_i denoted as $\bar{p}_i$ (mean of l_k) and the candidate place (annotation) a_i^j.
- $P(t_i|a_i^j)$ is a P value from a user profile based on the time and category of the places that the user tends to be.
- $P(a_i^j)$ is calculated by considering the assigned top category.

6. The model computes and iteratively updates the values until no changes are detected. The user profile and the place traction are updated upon a new point being added to the database.

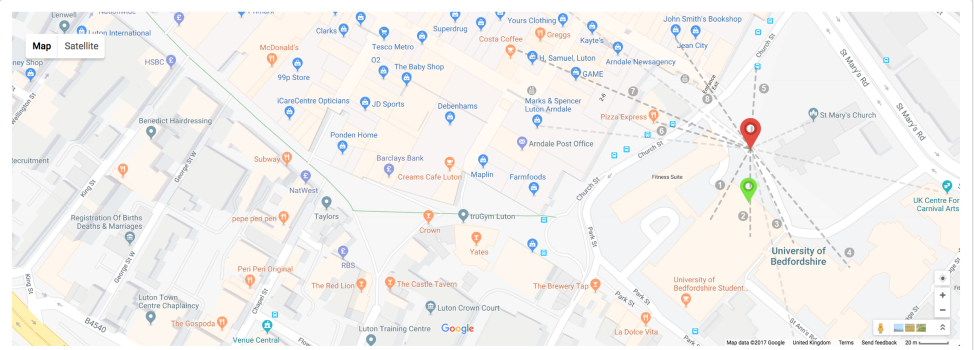
The automated annotation model is formed of four different components, namely, data preparation, POI retrieval, place annotation, and a direct link to prior knowledge. The architecture of the annotation module is illustrated in Figure 4.1.

The annotation model, in a nutshell, initiates the operation by fetching the preliminary data (p_i, t_i) from the corresponding databases. The data then undergoes pre-processing to create a unified flat-map structure and subsequently this is used

the annotation result compared to existing methods that estimate the annotations using only distance and dominant categories.

by the underlying part of the model. Next, the nearby places and their additional information are retrieved from the POI service by using the structured data. Meanwhile, the model generates an initial user profile histogram based on the prior knowledge provided by the users for use with the annotation algorithm. The process is continued by clustering the data points based on their significant density and storing them individually in the local memory for the upcoming calculation. And finally, the raw data, retrieved POI information, user profile, and clustered coordinates are all sent to the place annotation algorithm where the probabilities of the POIs $P(a_i^j|p_i, t_i)$ are calculated. The process iterates after the new points are added to the database in order to incrementally refine the probability values of the annotated places and hence improve the results. The model, at the end, returns a multi-level list of the most probable candidate places, categories, street addresses, cities, and countries and their probabilities.

The real case of using this model is shown in Figure 4.2. The red marker is the unknown place picked by the *SmarTracker* GPS application, and the green marker is the real position of the user. The date and the time of the place are recorded as Wednesday 01-March-2017 between 10:05AM to 12:51PM. According to the prior knowledge, the user works in the university during weekdays (Monday to Friday), morning and afternoon. The historical location data also, confirms that the recorded location is repeated during the weekday and certain hours of the day (approximately 8am to 6pm). Based on the gathered information such as POIs, user profile and historical location, the model calculates the pertinent probabilities and provides a list of candidate places, plus the street, town, and country, together with their probabilities sorted by the greatest P value. Moreover, the map in Figure 4.2 indicates the actual distance between the potential POIs (annotations a_i^j) and the recorded GPS coordinate (unknown place). It is clear that some of the irrelevant POIs (such as the bus stop, restaurant or church) are closer to the red marker (unknown place) and perhaps would be more likely to be the right annotation. But the model is explicitly giving less weight to them based on the user profile (prior-knowledge) during its calculation to provide a realistic enrichment result. Lastly, the weekly and hourly frequency of the unknown place



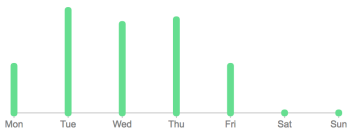
Cluster: 1, Coordinate: 51.878664,-0.410848, Assumption: Workplace
Address:

Name	Category	Date	Start	End	Probability
University of Bedfordshire - G321 1	College & University	01-Mar-2017	10:05:36	12:51:54	0.4523913770594671
IT Suite UOB 2	College & University	01-Mar-2017	10:05:36	12:51:54	0.3265803182026505 4
University of Bedfordshire - Luton campus 3	College & University	01-Mar-2017	10:05:36	12:51:54	0.10384059582182073
University of Bedfordshire - Business School 4	College & University	01-Mar-2017	10:05:36	12:51:54	0.07747154147492442
Church Street Stop C9 5	Travel & Transport	01-Mar-2017	10:05:36	12:51:54	0.025933350755085755
PizzaExpress 6	Food	01-Mar-2017	10:05:36	12:51:54	0.009297286043652266
Costa Coffee 7	Food	01-Mar-2017	10:05:36	12:51:54	0.0031923298103755542
John Smith's Bookshop 8	Shop & Service	01-Mar-2017	10:05:36	12:51:54	0.0005254474458389293
Iceland	Shop & Service	01-Mar-2017	10:05:36	12:51:54	0.0002928117122286281
Marks & Spencer	Shop & Service	01-Mar-2017	10:05:36	12:51:54	0.00016923410157280205
Tesco	Shop & Service	01-Mar-2017	10:05:36	12:51:54	0.00013567924516300187
The Mall	Shop & Service	01-Mar-2017	10:05:36	12:51:54	0.00009555561430166543
UK Centre For Carnival Arts	Arts & Entertainment	01-Mar-2017	10:05:36	12:51:54	0.000055401623122751125
St Mary's Parish Church	Professional & Other Places	01-Mar-2017	10:05:36	12:51:54	0.00001255369920996299
The Mezzanine Bar & Coffee Shop	Nightlife Spot	01-Mar-2017	10:05:36	12:51:54	0.000006517719874964090 5

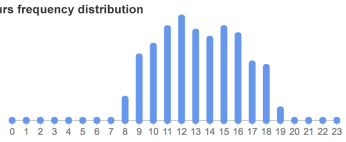
Street Address	Probability
the mall	0.4080578512396724
church street	0.16528925619834667
park street	0.09504132231404938
university bedfordshire	0.09297520661157005
park square	0.08471074380165272
vicarage street	0.08264462809917339
mary's road	0.0702479338842974
campus centre	0.0010330578512396695

County (town)	Probability
luton	0.8427983539094478
Greater London	0.15637860082304508
london	0.0008230452674897119

Weekly frequency distribution



Hours frequency distribution



Country	Probability
United Kingdom	0.9966666666666686

FIGURE 4.2: The result of annotating the unknown place (red marker on the map) compare to the real place (green marker) by using the automated place annotation

are shown. These frequencies are linked to the user profile and historical place information. More details relating to each part of the model are depicted in the next sections.

4.3.1 Data preparation

Trajectory data can be collected by various sensors. In this case, such data is obtained by means of smartphones with a GPS sensor and two individual tracking applications – *Moves* and *SmarTracker*. The Moves application is free and can be installed on iOS and Android while *SmarTracker* is developed for a research purposes with the MyHealthAvatar project[6] and can be installed only on Android devices.

The data from both applications have already undergone an initial pre-processing to remove outliers and also to detect the movements and stops by using the relevant advanced algorithms. The density of the data points collected by *Moves* is considerably low (10 GPS data points per day) as this application uses its own algorithm to minimise the number of GPS points within the certain range. In contrast, the data captured by *SmarTracker* is highly dense as it uses a high sampling rate (every 5 seconds) to record the GPS location. It is worth mentioning that the classification procedure of how the movements and stops are determined from raw trajectories, are not the scope of this research and are not covered in this thesis.

The structures of the pre-processed data from both applications (SmarTracker and *Moves*) are shown in 3.3 and in 3.4 respectively. The data, at minimum, includes latitude, longitude, and time-stamp. Although the structures of the data from the two applications, to some extent, are different due to their underlying architectures, this approach is able to deal with different data structures and creates a standard flat-map structure to be used by the rest of the process.

A list of requirements is defined that the data must meet to process the data points (raw places) within the proposed place annotation model. The list is composed of the following conditions:

[6] SmarTracker is an in-team development and its implementation process is out of the scope of this work

Algorithm 4.1: Data examination algorithm for annotation

```
input  : Preprocessed trajectories data
output : Valid data for automated annotation with semantic enrichment
Function (trajectoryData, conditions)
    trajectoryData ← removeDuplicates(trajectoryData) ;
    foreach object of trajectoryData do
        if object has no error/noise then
            if object has validCoordinate AND validTime then
                output ← standardiseFormat(object);
            else if object.time has specialFormat then
                output ← ConvertTime(object.time);
            else if object.coordinates has specialCoordinates then
                output ← ConvertCoordinate(object.coordinate);
            end
        end
    end
    return output;
end
```

1. The data need to carry a valid geographical coordinate. This allows querying the API place with the latitude and longitude coordinates of the user and obtaining an intact list of POIs.

2. The data is required to hold a correct time-stamp. This supports the model for more accurate results based on the prior knowledge provided by the user.

3. The data ought to be free from noise, error, and duplicate attributes.

These conditions are examined by the data preparation module in the proposed model before any further processing and the erroneous parts are deliberately dropped to prevent the model from producing an imprecise and poor result. The module, furthermore, restructures the input data to improve the performance of the algorithm in reading and writing. The procedure of examining the data in the annotation module according to the predefined conditions are shown in Algorithm 4.1. The structured data is subsequently used for retrieving the POIs, density-based clustering, and the annotation model.

4.3.2 User profile

The user profile accommodates the prior knowledge acquired from each individual. The prior knowledge, in this work, is the combination of user's lifestyle

and preferences, which evolves over time by being altered after each round of computation. This information, in the early stage, is obtained in the form of a small questionnaire regarding the user's personal daily life.

The questionnaire is designed to be non-specific and based on the daily life categories, e.g. Food, Travel & Transport, Professional, University & School. The rationale behind the questionnaire is to obtain broad information about the individual's daily life such as shopping places, work place, etc. in an easy and encouraging fashion. This increases the usability of the model and avoids impractical processing.

The questionnaire consists of five sections for the most significant categories that users deal with on a weekly basis, including shopping, food-related activity, outdoor & recreation, arts & entertainment, and profession. A complete questionnaire form can be found in Table A.2 in Appendix A.

All places within the POI services – in this case Foursquare – are classified into a number of categories to group similar places together categorically and reduce complexity. These categories are not fixed and can be vary in different POI services. The categories used in this work are retrieved dynamically via Foursquare's APIs – see Snippet 4.1 – and may update according to the POI service. More details about the Foursquare POI service is specified in the next section 4.3.3.

```
categories_list = [
    "Residence",
    "Professional & Other Places",
    "Shop & Service",
    "Food",
    "Travel & Transport",
    "Outdoors & Recreation",
    "Arts & Entertainment",
    "College & University",
    "Event",
    "Nightlife Spot"
]
```

Snippet 4.1: A list of categories retrieved dynamically from the Foursquare APIs

Based on the questionnaire, the user profile as part of the prior knowledge dataset is dynamically generated. The user profile is then used to create the initial profile histogram that is thereafter employed by the annotation model. The histogram generated consists of the common categories, day and time, weekday, and the number of category appearances during the week (frequency). The time of day is divided into three parts, namely, morning, afternoon, and evening to increase the usability and simplicity of the initial questionnaire. It should be mentioned that the data within the profile histogram can be subsequently amended or updated by new values from the model or by revising the user questionnaire. The structure of the questionnaire is shown in Snippet 4.2. Note that some of the categories in the questionnaire are empty but can be updated by the user or the annotation values accordingly.

```
var questionnaire = {
  "Arts & Entertainment": {freq:null, time:[], week:[]},
  "College & University": {freq:null, time:[], week:[]},
  "Event": {freq:null, time:[], week:[]},
  "Food": {freq:null, time:[], week:[]},
  "Nightlife Spot": {freq:null, time:[], week:[]},
  "Outdoors & Recreation": {freq:null, time:[], week:[]},
  "Professional & Other Places": {freq:null, time:[], week:[]},
  "Residence": {freq:null, time:[], week:[]},
  "Shop & Service": {freq:3, time:["evening"], week:["monday","friday"]},
  "Travel & Transport": {freq:null, time:[], week:[]},
};
```

SNIPPET 4.2: An initial user histogram structure with partially filled data

The structure of the user profile histogram is based on the weekdays. This means that there is an individual matrix for each day of the week. To accommodate the obtained information from the user questionnaire, a two-dimensional matrix $[24 \times 10]$ is created to fit the information within the 24 hours (i) and based on the categories (j). The structure of the matrix for each weekday is shown in Equation 4.4.

$$\textbf{weekdays histogram} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,j} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i,1} & a_{i,2} & \cdots & a_{24,10} \end{pmatrix} \tag{4.4}$$

The histogram matrix provides the probability value $P(a_{i,j})$ for specific categories (j^{th} column) during each hour (i^{th} row). The P value of $a_{i,j}$ is a float between 0 and 1 and the sum of each row (i) equals to 1 at all time. The process of creating the user profile histogram is shown in Algorithm 4.2. The P value is calculated evenly based on the frequency and the number of occurrences (Algorithm 4.2 – Line 7). The user profile histogram matrix is updated on each iteration and when the new data is added (Algorithm 4.2 – Line 12).

Algorithm 4.2: The user profile histogram algorithm

```
   input  : Questionnaire / current profile
   output : User profile histogram
 1 Function userHistogramMatrix
 2     if not histogramMatrix then create histogram matrix
 3         foreach w ∈ weekdays do
 4             matrix[w] ← matrix.zeros[24,10];
 5             foreach q ∈ questionnaire do
 6                 category(c) ← getCtg(q);
 7                 probability(p) ← calculateP(freq, wk) ;
 8                 updateMatrix(w, time, c, p);
 9             end
10             histogramMatrix←append(matrix[w]);
11         end
12     else update current histogram matrix
13         foreach pl ∈ placeList do
14             probability(p) ← getHighestProbability(pl);
15             placeWeekday(w) ← getWeekday(pl);
16             placeTimeSlot(time) ← getTimeSlot(pl);
17             placeCategory(c) ← getCtg(pl);
18             histogramMatrix←updateMatrix(w, time, c, p);
19         end
20     end
21     return histogramMatrix
22 end
```

Each part of the matrix can be updated accordingly based on the modified questionnaire or the model values for particular day, hour, or category. In total, the user profile histogram is formed of seven individual matrices with identical structure and an indication of the related weekday. The generic example of the final user histogram can be found in Snippet 4.3.

```
final matrix = [{ weekday: "Monday",
                  matrix: Array[24 x 10] },
                  {weekday: "Tuesday",
                  matrix: Array[24 x 10]},
                  {weekday: "Wednesday",
                  matrix: Array[24 x 10]},
                  .
                  .
                  .
                 ]
```

SNIPPET 4.3: Weekday histogram matrix

The next stage is to query the nearby POIs based on the identified coordinate point in the data. This process is depicted in the next subsection.

4.3.3 Retrieve Points of Interest

There are two available and free of charge services – with a certain rate limit – for acquiring POI information, Google Place and Foursquare. The Foursquare service is exploited in this research based on the following reasons:

- The popularity of the Foursquare service over Google Place which provides up-to-date venue information.
- The number of free requests per day known as the rate limit (5000 requests per hour in Foursquare[7] over 1000 requests per day in Google Place[8]).
- The ability to cache the information in Foursquare.
- Foursquare offers a competent three-level hierarchical category classification of the POIs (Figure 4.3) over the tag-based classification from Google; and
- Foursquare provides beneficial information such as checkins, and number of visits.

[7] https://developer.foursquare.com/overview/ratelimits
[8] https://developers.google.com/maps/pricing-and-plans/

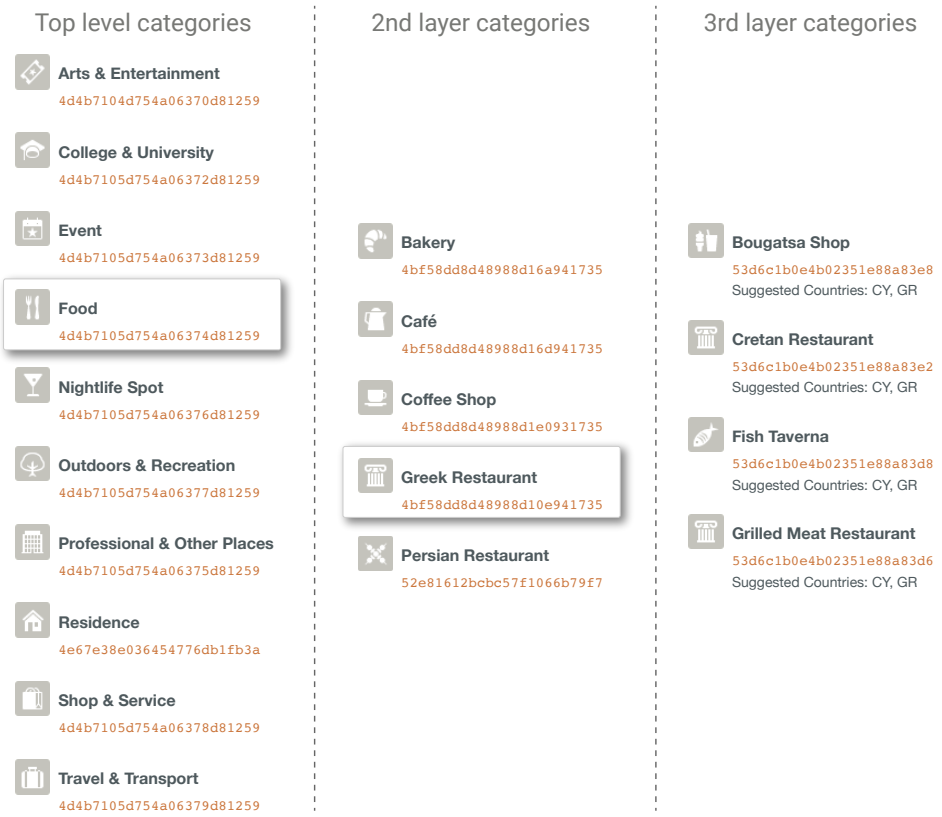


FIGURE 4.3: The hierarchical category provided by the Foursquare POI service.

The top-level category used in this research consists of 10 categories (see Code 4.1), each of which includes a number of subcategories and sub-subcategories. The subcategories are not included in the model but are provided with the additional information in the results.

Furthermore, Foursquare supports a complete RESTful API with useful options to search for venues such as radius, intent, and limit options. These options are used to optimise the POI retrieval process. The sample request to retrieve a list of nearby venues (or POIs) is represented in Snippet 4.4. Each request needs a valid coordinate in the form of `[Latitude, Longitude]`. Additionally, the result of the query is a list of potential places and their details according to the predefined option within the query. However, the return result does not include the haversine distance and the top-level category by default. To add these two attributes, the category name provided is mapped based on its ID to retrieve the

top level from the APIs while the distance is directly computed between each nearby place coordinate and the targeted place. The complete list is subsequently used in the annotation algorithm. The final result that the POI service returns for a nearby place (e.g. the University) is shown in Snippet 4.5.

```
var API_ENDPOINT = 'https://api.foursquare.com/v2/venues/search' +
        '?client_id=CLIENT_ID' +
        '&client_secret=CLIENT_SECRET' +
        '&v=20160815' +
        '&ll=LatLng' +
        '&limit=20' +
        '&radius=50' ;
```

Snippet 4.4: The REST APIs to retrieve the candidate venue

Additionally, in order to retrieve the venues in an optimum way and address the POI query limit, the preprocessed data including the coordinates are queried in one batch by 1) extracting all the valid and non-duplicate data points (lat, lng), 2) generating an identifier for each of them, 3) retrieving all candidate places according to their coordinates, and 4) keeping the list of results including the identifier in the local memory for use in the next step of the automated annotation model. By doing so, the model only makes one attempt and eliminates the duplicate queries and delay time for retrieving the information from the APIs.

Thus far the user profile histogram and a list of candidate places are made available by using the relevant algorithm. The next stage is to computing the P values by involving the user profile and retrieved POIs. In the next section, how the annotation engine employs these data to compute the probability and return a list of potential places is explicitly described.

4.3.4 The multi-level place annotation

As outlined at the beginning of this chapter, the automated annotation is inspired by the Bayes theorem to compute the probability value of the candidate places (annotations) retrieved from the POI service – Foursquare. This process and its

```
{
        "category": {
                "id": "4bf58dd8d48988d1fd941735",
                "cat_top_level": "College & University",
                "categoryName": "University",
                "pluralName": "Universities",
                "primary": true,
                "shortName": "University"
        },
        "contact": {
                "formattedPhone": "+44 1234 400400",
                "phone": "+441234400400",
                "twitter": "uniofbeds"
        },
        "hasPerk": false,
        "hereNow": {},
        "id": "4bb0e59df964a520e7673ce3",
        "location": {
                "address": "Park Square",
                "cc": "GB",
                "city": "Luton",
                "country": "United Kingdom",
                "haversine_dist": 59,
                "labeledLatLngs": {
                        "lat": 51.877853451990504,
                        "lng": -0.4113378201237996
                },
                "postalCode": "LU1 3JU"
        },
        "name": "University of Bedfordshire -- Luton campus",
        "referralId": "v-1493996190",
        "stats": {
                "checkinsCount": 3541,
                "tipCount": 10,
                "usersCount": 337
        }
}
```


Snippet 4.5: The structure of the data from the Foursquare POI service

underlying structure are extensively described in this section by breaking down the formula into pieces.

The formula (equation 4.1) is to compute a P value for each candidate place (annotation) $P(a_i^j|p_i, t_i)$ in connection with the distance $P(p_i|a_i^j)$, the user's profile $P(t_i|a_i^j)$, and the place traction $P(a_i^j)$.

The input to this equation is a tuple of unknown place (GPS coordinate) and time (p_i, t_i). However, due to the accuracy of GPS devices and the cyclic nature of the daily life data, the coordinates for a particular unknown place can be recorded marginally different. Therefore, this can, by some means, form a number of dispersed points repeated over time rather than a single point for the same

location. To handle such data and refine the result, these closely nearby points are clustered by means of an incremental density-based cluster algorithm and use the central point of the cluster instead. This process expedites the performance of the algorithms by eliminating redundant calculation.

$P(p_i|a_i^j)$ is formulated to provide a normalised result based on the distance of the unknown place (GPS coordinate p_i) and the retrieved venue (a_i^j).

$$P(p_i|a_i^j) = \left| \frac{1}{\log(\frac{1}{DS_{pa}})} \right| \tag{4.5}$$

Consequently, the distance between each candidate place (annotation) and the GPS coordinate DS_{pa} is calculated via the harversine formula. Haversine, in contrast with Euclidean, obtains the distance by considering a great circle between the given latitude and longitude on a sphere in which the result is more realistic – see equation 4.6. The haversine formula is used where there is a need to determine the distance between two points in the automated annotation procedure.

$$\begin{aligned} DS_{pa} &= 2R \arcsin^2 \left(\sin^2(\frac{\varphi_2 - \varphi_1}{2}) + \cos(\varphi_1) \times \cos(\varphi_2) \times \sin^2(\frac{\lambda_2 - \lambda_1}{2}) \right) \\ &= 2R \arcsin \left(\sqrt{\text{hav}(\Delta\varphi) + \cos(\varphi_1) \times \cos(\varphi_2) \times \text{hav}(\Delta\lambda)} \right) \end{aligned} \tag{4.6}$$

Where DS_{pa} is a physical distance between two coordinates, φ is latitude while λ is longitude of the unknown place and the candidate place (both in radian), and R is the radius of the Earth (as a sphere).

The term $P(t_i|a_i^j)$ is to determine the probability of a category's likelihood from the user profile histogram based on the time. In other words, this part contributes in assigning more weight to the category of viable places (e.g. university, resturant, etc.) that the user tends to be at within a certain time. This value is retrieved from the vector that contains all the categories within the particular time-stamp similar to the equation 4.8. For example, if the user tends to go to a cinema,

which belongs to the category of Art & Entertainment, on Friday nights, venues with the similar category are assigned more weight for the Friday evening matrix by using equation 4.7 and hence greater probability.

$$P(t_i|a_i^j) = \frac{frequencyCat_j \times durationCat_j}{\sum_{0}^{j}\Big((frequencyCat_j)(durationCat_j)\Big)} \tag{4.7}$$

The algorithm then acquires a related vector matrix according to the time and returns a relevant value based on the category. For instance, if the candidate place embraces a transportation category with the index of 3, the algorithm returns the third value of the vector within the certain hour (e.g 1am) $a_{hour,2}$ (considering the index starts from 0) [9].

$$profile_{matrix} = \begin{pmatrix} a_{0,0} & a_{0,1} & a_{0,2} & \cdots & a_{0,j} \\ a_{1,0} & a_{1,1} & a_{1,2} & \cdots & a_{1,j} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{i,0} & a_{i,1} & a_{i,2} & \cdots & a_{23,9} \end{pmatrix} \tag{4.8}$$

Selecting the specific *hour*:

$$x_{hour,category} = \begin{bmatrix} a_{1,0} & a_{1,1} & a_{1,2} & \cdots & a_{1,9} \end{bmatrix}$$

Another term is the place traction $P(a_i^j)$. The algorithm for this part employs a density-based clustering method to calculate and denote the prevalence of the categories within each cluster in harmony with distance and time. This algorithm utilises a density-based clustering DBSCAN with a predefined radius ϵ and minimum points min_{pts} to incrementally cluster the coordinates (Algorithm 4.3), then calculates the probability that the category belongs to the cluster with respect to the time information, and finally produces two individual probability matrices, "weekdays-weekend" and "weekdays". The former contains an aggregated $[2 \times 24]$ matrix for five days (Mon–Fri) and two days of the week (Sat and Sun) within 24

[9] The indexing system for all the matrix starts from zero. This means that the hour starts from 0 to 23 (contains 23:00 to 23:59:59), and categories starts from 0 to 9

Algorithm 4.3: Density-based significant clustering algorithm

```
input  : D, eps, MinPts
output : Cluster
Function DBSCAN(Coordinates, eps, MinPts)
    C = 1 foreach unvisitedPoint (P) in dataset (D) do
        visited ← mark P as;
        N = getNeighbors(P, eps);
        if size of N <MinPts then
            NOISE ← mark P as;
        else
            C = nextcluster;
            expandCluster(P, N, C, eps, MinPts);
        end
    end
end
Function expandCluster(P, N, C, eps, MinPts)
    add P to cluster C;
    foreach point P' in N do
        if P' is NOT visited then
            visited ← mark P' as;
            N' = getNeighbors(P', eps);
            if sizeof(N') ≥ MinPts then
                N = N joined with N';
            end
            if P' is NOT yet member of any cluster then
                add P' to cluster C;
            end
        end
    end
end
```

hours whilst the former includes a $[7 \times 24]$ matrix for all weekdays (Mon–Sun). The reason for producing two sets of matrices is that the result is highly dependant on the number of places within the cluster and a small amount of data in the early stage can lead to an extremely low probability within the process.

Each cluster has its own associated place and category traction voting score based on the time. The voting is calculated relatively by the following equation:

$$P(a_i^j) = \left(\sum_1^v \frac{C_v}{d_v} \right) \times \left(\frac{F_i}{\sum F_i} \right) \div \left(\frac{\sum day_i}{week} \right) \tag{4.9}$$

Where $P(a_i^j)$ is a voting score for the category (j), C_v is the number of check-ins to the venue with the category (i), while d_v is the distance of the same venue to the

Algorithm 4.4: The place traction vote within each cluster

input : Clustered trajectory data
output : Clustered data with a place traction and category vote

```
1  Function CategoryVote(data, allCandidatePlaces)
2      allClusters ← DBSCAN(data, eps, minPts);
3      foreach cluster in allClusters do
4          allCandidatePlaces ← venueLookup(cluster.parts.LatLng);
5          allTimes ← getTime(cluster.parts);
6          allWeekdays ← getWeekdays(cluster.parts);
7          CategoryFreq ← occurrences(allCandidatePlaces);
8          foreach venue in allCandidatePlaces do
9              Distance ← HaversineDist(cluster.LatLng, venue.LatLng);
10             checkins ← GetNormCheckin(venue.checkin);
11             voteScore ← ComputePvalue(Distance, checkins, categoryFreq);
12         end
13         voteMatrix ← store(voteScore based on  week-wknd, Weekdays)
14     end
15     return voteMatrix
16 end
```

cluster central point. F_i is the frequency of category (i) within the retrieved venues divided by the overall category frequency. $\sum day_i$ is the total number of days that the associated coordinates have occurred. For instance, if the coordinates have only occurred on Monday and Tuesday, $\sum day_i = 2$. Similarly, $week$ represents the fixed number of week days (seven days).

Each vote is stored in the voting matrix of $[24 \times 10]$ according to the time (24 hours) and the categories – see Snippet 4.6 for more information about the structure of each cluster and stored voting. This process is iterative. This means that new points can fit into the existing cluster or form a new cluster.

The location is the central point of the cluster. The parts are the index list of the datasets from which by looking up the index the actual place can be retrieved. The weekdays and times are related to the day of week and the time that the unknown places occurred within the cluster. The category vote is the final matrix. The algorithm that calculates the P values for place and category traction vote is shown in Algorithm 4.4.

The compound probability result of categories and related times for "weekdays-weekend" and "weekdays" matrices are portrayed in equation 4.10 and 4.11

individually.

$$m_{wk,hr} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{wk,24} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{wk,24} \end{pmatrix} \tag{4.10}$$

where $m_{wk,hr}$ is the weekday-weekend matrix based on 24 hours. The first row wk indicates the weekday (Mon–Fri) and the second shows the weekend (Sat–Sun) probabilities according to 24 hours.

$$m_{wkday,hr} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{24} \\ a_{21} & a_{22} & \cdots & a_{24} \\ \vdots & \vdots & \ddots & \vdots \\ a_{7,1} & a_{7,2} & \cdots & a_{wkday,hr} \end{pmatrix} \tag{4.11}$$

This matrix contains seven rows for the weekdays and 24 columns indicating 24 hours.

Lastly, the final algorithm makes use of the result of the aforementioned functions to compute the final probability value for a given GPS coordinate (unknown place) and returns a list of viable annotations in different level of details including the place name and category, street, town, and country with their related probability values. The Algorithm 4.5 represents the line of actions taken for providing a multi-level semantics.

The auto annotation algorithm starts with checking the existence of any auto annotation and candidate places list before creating and retrieving the required information (Algorithm 4.5 – Lines 3 to 19). Next, the trajectory data is being processed by the category voting algorithm to either create new or update the existing clusters based on the newly added data points (Algorithm 4.5 – Line 20). Subsequently, each unknown place (coordinates) in the trajectory data is being checked by considering the coordinate, time and weekday. If there is an exact match of existing annotation list, the algorithm assigns the same annotation list to the coordinate (Algorithm 4.5 – Line 26), otherwise the algorithm retrieves the related

candidate places (based on the $LatLng$), finds the cluster of unknown coordinates, and creates a vector matrix based on the user profile and the coordinate's temporal information. It then computes the probability value of address, town, and county based on the unknown coordinate's clusters (Algorithm 4.5 – Lines 32 to 37). Next, the algorithm calculates the probability for each candidate place (object) and appends it to the annotation list[10]. The algorithm by default selects the candidate place with the greatest P value as a most likely annotation – place name (Algorithm 4.5 – Line 50). This process is iterated (10 times as default) until no changes can be found between the annotation including its P value from the last iteration and the current one. The algorithm, in the end, stores the result on the server as a complete auto annotation list.

An example of the annotation result is shown in Snippet 4.6. The result comprises the GPS coordinate (lat, lng) of the unknown place, temporal information such as start and end time, a unique identifier, a local id (if applicable) and a list of country, town, and street probabilities together with the complete candidate places list (known as annotation list). Each candidate place encompasses rich semantic information such as the top level and low level place category details, the haversine distance (between the coordinate and the candidate place), temporal info, candidate place name, calculated probability value, and complementary information about the place.

4.4 Evaluation and Benchmark

An extensive evaluation was conducted in order to validate and study the accuracy of the annotation model by using six different datasets with known ground truth, three of which were collected by the Moves application and the other three acquired by SmarTracker. Both datasets comprised a period of 4 weeks, 6 weeks, and 8 weeks. The data includes the trajectories (place coordinates and timestamps)

[10] The list contains a number of individual candidate places (can be adjusted between 10 to 50 places) for a single unknown coordinate with complementary information and also the street, town, and country probability list

```
country_probability:Array(1),
town_probability:{
    city:"luton"
    probability:0.9345
},
street_probability:Array(7),
annotations:{
    category_details:"University",
    category_top_level:"College & University",
    distance_meter:14,
    end_time:Mon Mar 27 2017 12:40:07 GMT+0100 (BST)
    name:"University of Bedfordshire -- Luton Campus Library"
    probability:0.8943,
    start_time: Mon Mar 27 2017 12:00:39 GMT+0100 (BST)
    venue_info:{}
},
identifier:"51.56921380951998,-0.27571810185650736",
local_id:undefined,
location:[lat,lng],
name:"Unknown",
start_time:"2017-03-27T12:00:39+01:00",
end_time:"2017-03-27T12:40:07+10:00",
timestamp_millisecond:1490612439000
```

SNIPPET 4.6: Annotation result for the an unknown place

collected during the user's daily life and underwent a complete pre-processing procedure.

Given the limitation in collecting the real-life data for this research, the collected data from *Moves* and *SmarTracker* were derived into four different categories based on the geographical location. Dataset 1 belongs to a person who lives in London and his work place is in Bedfordshire during the weekdays. Dataset 2 belongs to a person who lives and works in Bedfordshire. Dataset 3 belongs to a person who lives in Hertfordshire and travels to his work place in Bedfordshire during the weekdays. Dataset 4 belongs to a person who lives in Milton Keynes and works in Bedfordshire.

The semantic enrichment process was performed for each dataset (comprising large number of data points - unknown places' trajectories) on a weekly basis in order to allow the model to update the user profile histogram along with historical information and improve the results. The results were compared to the ground truth that had been made available by the participants. Additionally, the participants were asked to review the result and validate a list of top potential places with high probability scores followed by flagging incorrect annotations.

Algorithm 4.5: The automated annotation algorithm

```
input  : Trajectory data
output : List of annotated trajectory data
Function autoAnnotation (trajectoryData, candidatePlaces, autoAnnotationList)
    if no autoAnnotationList then
        autoAnnotationList ← List();
    end
    if candidatePlaces then
        check candidatePlaces against trajectoryData;
        if new coordinates with no candidatePlaces then
            placeList ← retrieveVenue(coordinates);
            candidatePlaces ← append (placeList, coordinate);
            updateOnServer(candidatePlaces)
        else
            continue
        end
    else
        candidatePlaces ← List();
        placeList ← retrieveVenue(coordinates);
        candidatePlaces ← append (placeList, coordinate);
        storeOnServer(candidatePlaces)
    end
    voteCtg ← categoryVote(trajectoryData, candidatePlaces);
    foreach unknownPlace in trajectoryData do
        time ← getTime(unknownPlace);
        weekday ← getWeekdays(unknownPlace);
        if unknownPlace has annotationList in autoAnnotationList then
            if (time & weekday) = annotationList (place,time,weekday) then
                autoAnnotationList ← assign(this.annotationList);
            else
                GoTo line 31;
            end
        else
            annotationList ← List();
            candidatePlaces ← venueLookup(unknownPlace.LatLng);
            cluster ← lookupCluster(voteCtg.index);
            vectorMatrix ← createVector(profileHistogram, time, weekday);
            P-streetList ← compute(candidatePlaces.location.address);
            P-townList ← compute(candidatePlaces.location.city);
            P-countryList ← compute(candidatePlaces.location.country);
            if length of candidatePlaces ≥ 1 then
                foreach object of candidatePlaces do
                    haversineDistance ← computeDistance(object.LatLng,
                      unknownPlace.LatLng);
                    placeTraction ← frequencyPvalue(unknownPlace);
                    P-value ← computeWeight (vectorMatrix[object.category],
                      haversineDistance, placeTraction[object], voteCtg[cluster]);
                    object ← append(P-value);
                end
                annotationList ← append(candidatePlaces, P-streetList, P-townList,
                  P-countryList);
                normalise all P values;
            end
        end
        topPvalue ←findMax(P-value) in candidatePlaces;
        placeName ← (candidatePlaces).indexOf(topPvalue);
        if iteration < max-iteration AND (placeName & P-value) ≠ last iteration result then
            updateProfileHistogram(time, weekday, placeName.cat, P-value);
            iterate();
        else
            stop iterate();
            autoAnnotationList ← append(placeName, annotationList);
        end
    end
    return autoAnnotationList;
end
```

This way, valuable information were gathered regarding the result and incorrectly assigned places. The result of the evaluation for the multi-level annotation is depicted in the following paragraphs.

As discussed, the result of the annotation contains a list of top 10 probability values in different levels such the name, category, street, town, and country of a targeted place. The method was examined by looking at each level individually and obtaining to what extent the calculated result could be accurate.

The result of the multi-level examination for a_i^j contains $(a_i^{j_n}, a_i^{j_g}, a_i^{j_s}, a_i^{j_t}, a_i^{j_c})$ is shown in Figure 4.4. The analysis shows that the annotation model finds the correct country $(a_i^{j_c})$ on the first instance with 100% accuracy. The accuracy of determining the correct town $(a_i^{j_t})$ is 93% in the first instance and 100% within the top two instances. The street address $(a_i^{j_s})$ reaches a high level of accuracy within the top three instances whilst the first instance returned a poor accuracy due to the nature of the POIs around the work place to which they were all allocated within the shopping mall. The accuracy of obtaining the precise category $(a_i^{j_g})$ of the place reached the highest value of 100% within the first top two instances whilst the model achieved 86% validity on the first instance (top 1). The place name was the challenging part of the evaluation as the rate of error was high due to the nature of POIs along with accuracy of the collected data. The result of analysing the place name $(a_i^{j_n})$ shows that the automated annotation model was able to reach an accuracy of 69% within the first instance in a list. The rate rose to 76% for the top two suggested places and continued to increase to 86% within the first top three suggestions. According to Figure 4.4, the annotation reached the highest validity of place name within the top seven suggested places.

Figure 4.5 shows the overall accuracy in calculating the probability value and obtaining the correct multi-level annotation within the list of top 10 suggestions. The overall accuracy of the model based on the multi-level result is 72% within first instance and rises to 94% for the top two instances.

The model and the implemented algorithms were tested in JavaScript to examine the performance and efficiency. All the tests were conducted on the Chrome web

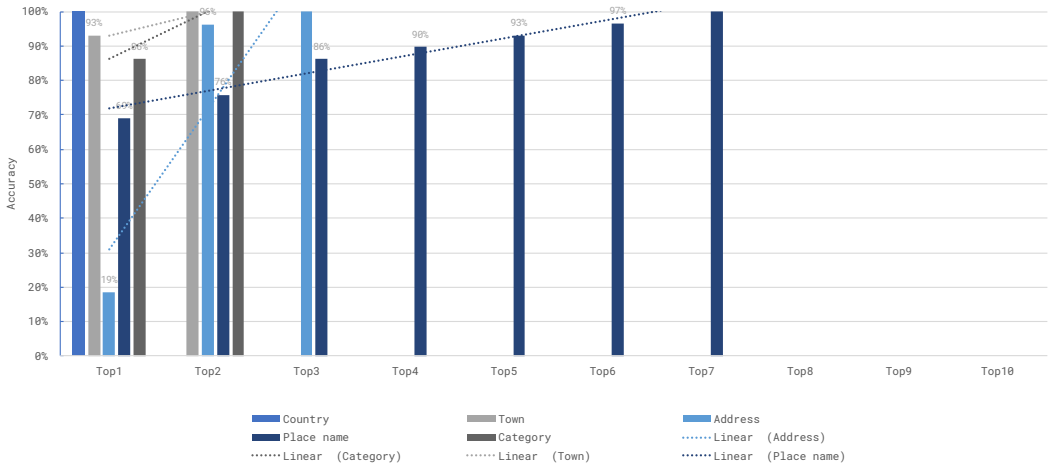


FIGURE 4.4: The result of the multi-level annotation evaluation

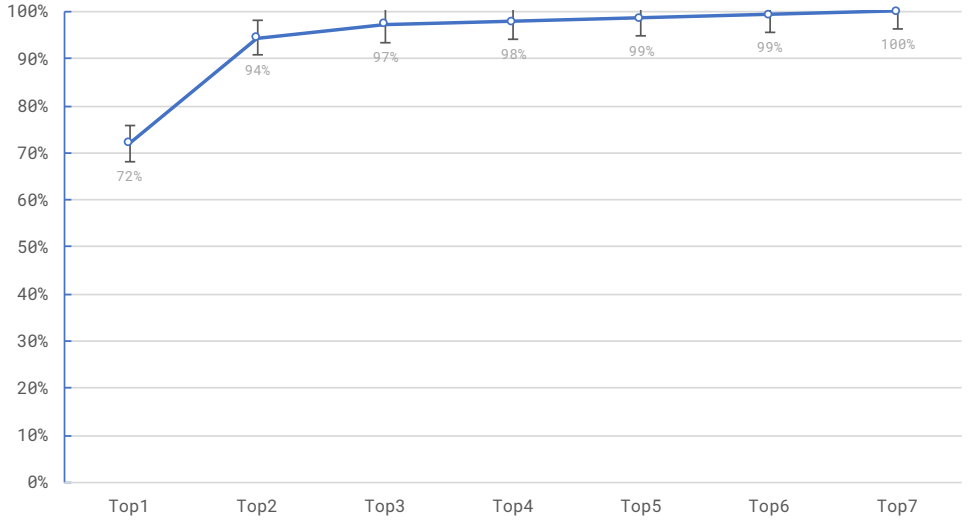


FIGURE 4.5: The overall accuracy of the multi-level annotation

browser v58, MacOS 10.12.4, Intel Core i7 with 32 GB memory and internal graphic card. Two types of tests were carried out on the automated place annotation, performance and efficiency test. The performance test is referred to loading different size of datasets and measuring the performance together with the accuracy of the model. The efficiency test is examines the implemented algorithm by obtaining a normalised value of achievable iterations per second. The former benchmark test is written as a native part of the implementation while the latter employed is the jslitmus.js[11] library to obtain the results.

[11] https://github.com/broofa/jslitmus

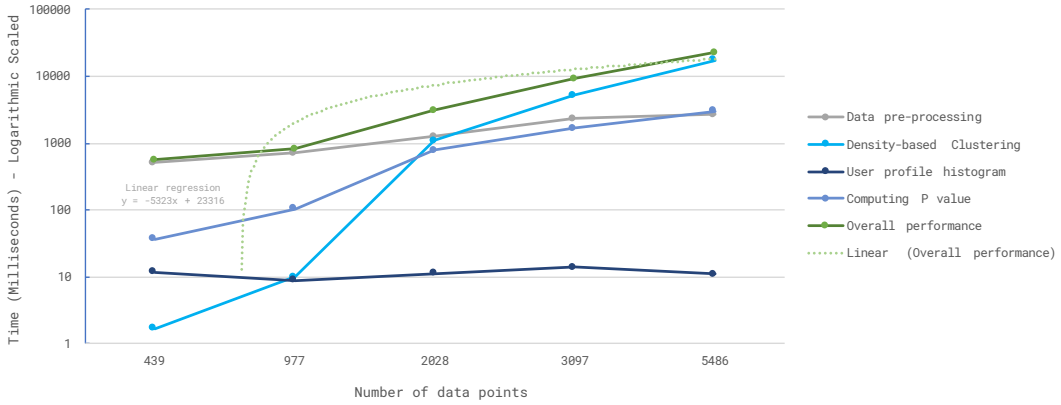


FIGURE 4.6: A benchmark result of the annotation model and its components with a logarithmic scale time

The performance test is conducted by using five real-life datasets of different data points (each of which contains at least 5000 data points – unknown coordinates), running the annotation model according to the pipeline, and measuring the completion time of annotation model including the incorporated components. The test includes the overall performance, data preprocessing, density-based clustering, user profile histogram, and P value computation time. Figure 4.6 shows the overall result of performance test based on the completion time and the number of data points within each datasets in which the completion time grows by increasing the number of data points. The density-based clustering performance has noticeably risen, particularly when the number of data points expands. This is due to the structure of the density-based algorithm that needs to form the cluster based on the predefined distance and minimum points. The performance of the density-based algorithm is not initially prominent but dramatically improved within the iteration and the following round of annotation as the new data points are either added into the existing clusters or partially updated.

The annotation algorithm performance, as shown in Figure 4.7, is $O(n \log n)$ and highly affected by the density-based clustering performance. The algorithm is also assessed to see the impact of increasing the number of data points on each part. Figure 4.8 illustrates how the number of datasets affects the performance of each part individually. Once more the clustering algorithm is shown as a higher pick

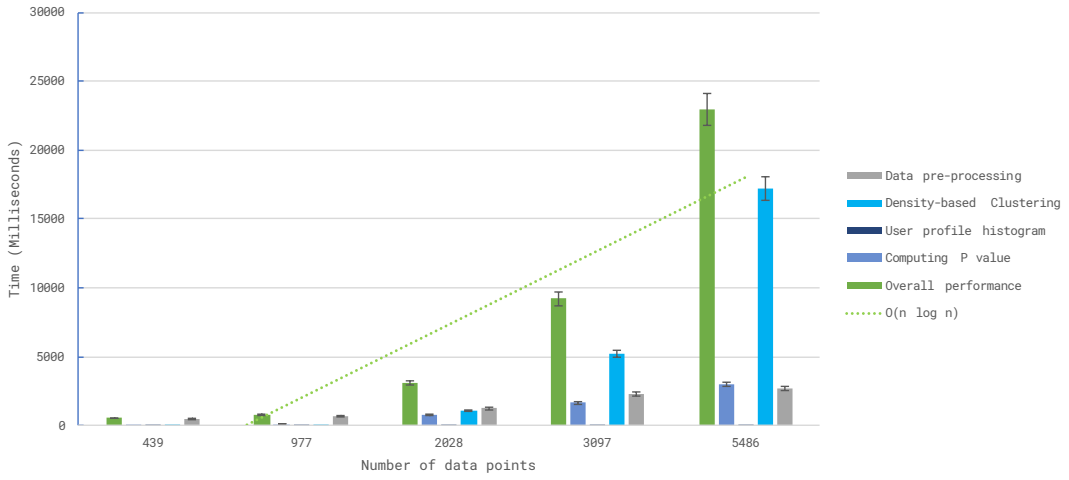


FIGURE 4.7: Benchmark result for individual part of the annotation algorithm

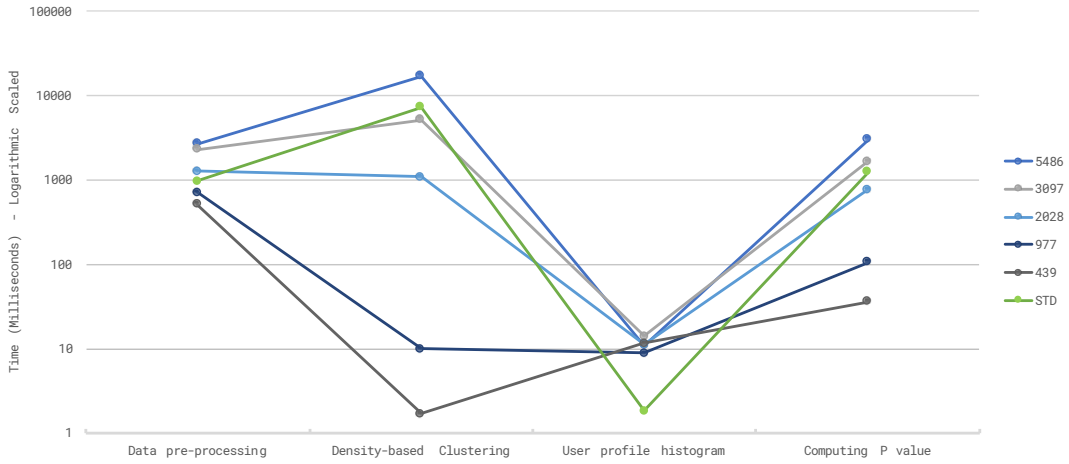


FIGURE 4.8: The effect of increasing the data points on each component of the automated place annotation

within the rest of the components. This yields the need to enhance the clustering algorithm by reformulating the algorithm, which can be addressed in future work. In addition, the efficiency of the implemented algorithm in JavaScript is analysed and the result is shown in Figure 4.9 with a logarithmic scale.

Moreover, the ontology is defined (based on the Renso et al. [172]) in order to facilitate the process of determining the user's frequent places such as home, work place, etc. The ontology indicates the places that the user has a long and frequent stay overnight, the weekend, or weekdays. For instance, the definition of the work place in the ontology encompasses the place that the user stays during the

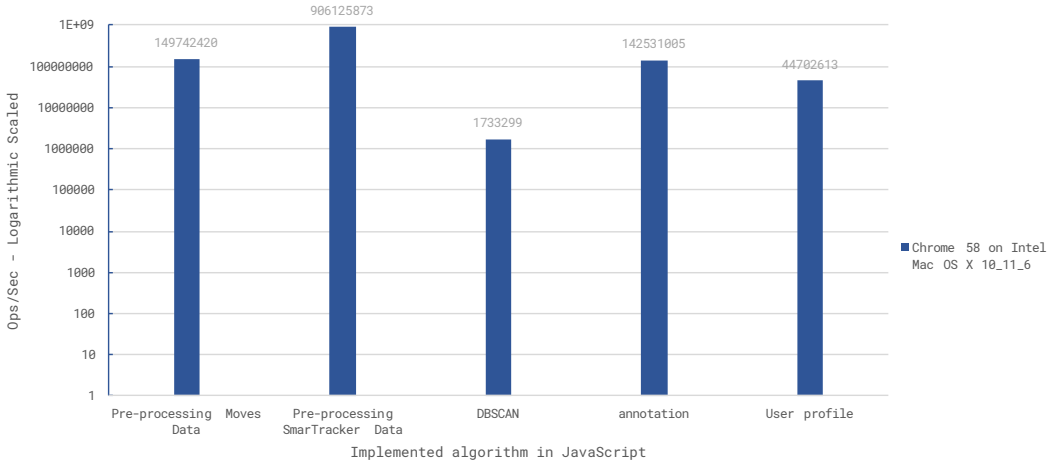


FIGURE 4.9: The efficiency of the implemented algorithm in JavaScript based on the iterations per second.

day frequently – possibly on a daily basis. However, the ontology definition can vary according the user's prior knowledge histogram. It means that if the user mentions that the working hours are on a night shift, then the ontology needs to be dynamically adjusted to the place that the user spends time during the night as a work place.

Examining the annotation method shows that the algorithm returns practical probability scores by incorporating the user profile histogram, density-based clustering, place traction, and category voting. The accuracy of the method is measured by analysing the results against the ground truth. The result shows 79% accuracy within the moderate venues. However, the accuracy may be reduced within the dense area that contains a great number of candidate venues. Moreover, the performance of the proposed model and its implemented functions are assessed in JavaScript. The result shows a rational $O(n \log n)$. Nevertheless, it indicates that the performance drops by increasing the number of data points. Furthermore, using an ontology contributes to identifying the legitimate places as home or work place.

4.5 Chapter Summary

In this chapter, the multi-level automated annotation model with latent semantic enrichment was extensively described. This model is composed of data preparation, the user profile histogram, venue retrieval via POI services, and an annotation model to compute the probability values and provide a multi-level annotation list of country, town, street, place name, and category. Each part of the model has been described in detail followed by an example to show how the algorithm comprises the prior knowledge, the venue search result, and iteration technique to determine the list of best places according to the time and place histories. The model has been discretely evaluated via using six datasets with known ground truth and achieved an overall of 72% accuracy based on the first top suggestion whilst the accuracy rose to a tolerable rate of 86% for the top two suggestions and 100% within the top seven suggestions in a multi-level annotation list. In addition, a complete benchmark was carried out to validate the efficiency of the implemented algorithms for this model. The benchmark results show that, as expected, the algorithm runs with $O(n \log n)$ complexity time and by increasing the number of places, the performance is reasonably decreased.

CHAPTER 5

A Multi-Significance Event Ranking Model

5.1 Introduction

This research is intended for an effective exploration of personal daily life data via a novel visual analytics approach. This requires an extended model that embraces a trivial event filtration along with an automatic and highly customisable significant event extraction. However, event detection with fixed and predefined parameters may not be of interest from the user's perspective with different life style and preferences. According to the literature review, extracting events without involving factors such as frequency, occurrence, and user point of interest can lead to impractical information extraction and hence ineffective knowledge discovery. For instance, an event that takes place frequently every day can be less important while an event occurs rarely within a certain period can be more valuable and hence interesting to users. To address this, this thesis introduces a novel significance ranking model that comprises an extensive data-mining algorithm and allows for openly customising the definition of significance according to user preferences to provide tailored significant events.

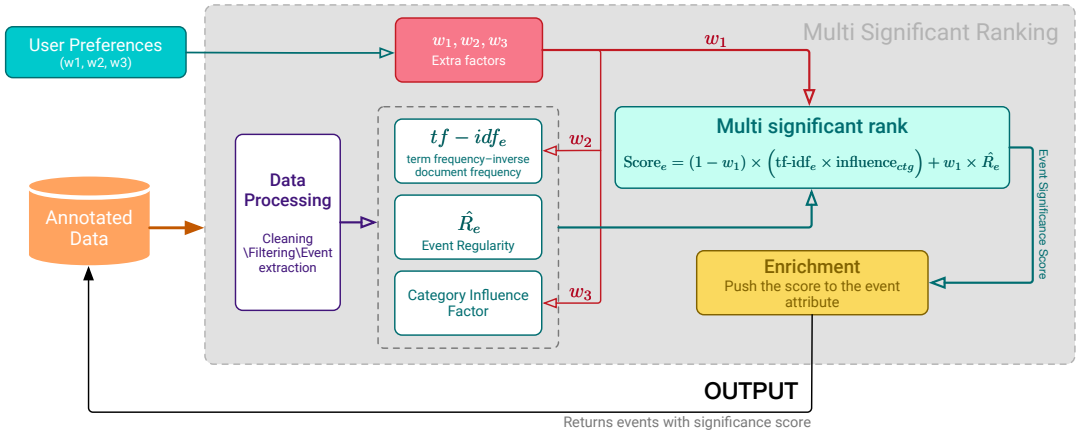


FIGURE 5.1: The multi-significance ranking model overview

The proposed event ranking model is the core of event extraction in this research as it can set aside trivial data and bring concealed information to view. The definition of significance is described by employing a number of factors based on user interests and preferences. The model exercises these factors within its calculation to form the empirical ranking score. These factors are composed of an occurrence frequency, variations, and category influence factor. Over and above that, these factors are customisable based on the user's preferences and articulated via three different weights by users. The overall process of the ranking model including the data processing, fixed factors, and final significant score calculations is illustrated in Figure 5.1.

The contribution of this chapter is a novel multi-significance event ranking model, which calculates the significant scores for the events in a personal history according to user preferences and allows users to efficiently identify key events over a selected period of time based on their personal preference settings, including the preference for event category, occurrence frequency, and regularity.

The remainder of this chapter is organised as follows. The process of data processing is described including an event detection model and significance ranking model based on the aforementioned factors with generic examples. Next, the accuracy and performance of the model is demonstrated in the evaluation section and progress is concluded in the last section.

5.2 Event Detection Model

In this section, the event extraction and its procedures are described. First, the data preparation for the extraction stage is explained and then, focus moves to the extraction process by providing the associated flowcharts and algorithms. The section ends by presenting the event extraction outcome used by the ranking module.

The personal daily data, as mentioned before, consists of a massive number of segmentations such as time-stamps, activities, location information, track points, and the like that shape events. This data may include errors, missing points, and low level semantic details. To handle this, there is a need to identify and refine the issues by means of automated place annotation, data pre-processing, and event extraction techniques. The automated place annotation in Chapter 4 works toward enriching the semantics while the data pre-processing is used to delete the errors or unneeded attributes during the course of flat-mapping the data, which results in higher performance in retrieving the massive amount of daily data. Furthermore, to amplify the performance of reading and processing the large-scale personal data, the events are extracted and the original order of this data restructured to an optimised form.

5.2.1 Data pre-processing

The process of data cleaning followed by restructuring is particularly designed to be run on-the-fly despite the pre-processing in the annotation model. This process is performed each and every time the data changes or loads on the client to 1) retain the original data untouched, and 2) to comply with user privacy needs.

The pre-processing of data, within the ranking model, includes three main parts: 1) removing the erroneous or incomplete parts of the dataset, 2) filtering unnecessary attributes, and 3) validating the dataset. The cleaning validates the data with the main requirements – see Figure 5.2. This process can be similar to check if the

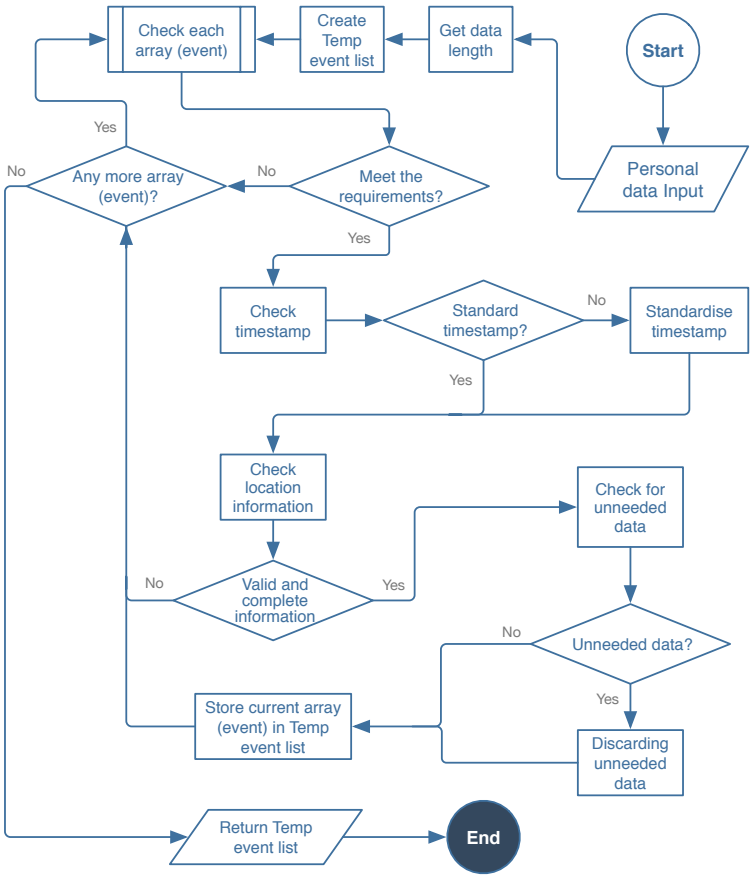


FIGURE 5.2: Flowchart for data cleaning and filtering

data has a valid timestamp (start and end times), available location information, and the like. If any of the arrays (events) do not match with the requirements or do not contain the necessary event's attributes (e.g. time, location, etc.), the data cleaning function disregards the entire array of the recorded activity or place to prevent the model from breaking from a hidden problem. In addition, the filtering aims at simplifying the data by dropping the unnecessary attributes which the visual analytics methods dose not use. And, the validation evaluates the process to ensure that the data are processed correctly and the outcome is free from any missing points or error. Figure 5.2 shows a flowchart of the process.

5.2.2 Event extraction

This process includes identifying potential events and passing them on to the ranking model for further analysis. This process, similar to the other parts, is completed on-the-fly and hence, is required to be fully optimised for handling large-scale personal daily life datasets.

The extraction process is initiated by finding the potential events based on our event definition in which each place and its corresponding actions as an event are considered. The potential events must encompass the minimum requirements such as timestamps information, place name, place category, and geographical location. In addition, as the data is enriched by the automated place annotation and contains a number of potential places and categories, the extraction algorithm is required to select the foremost place and the category. To this end, the algorithm compares the annotations in accordance with the probability value and determines the place names and categories with the highest probability. Subsequently, the process continues through procuring and storing all the possible events in memory to support faster retrieval of data by the ranking model.

A brief outline of the event extraction process is depicted in Algorithm 5.1. The input data is the events with rich semantic information that are extracted and enriched according to the predefined requirements[1]. Initially, the algorithm creates an empty event list, checks each array (which represents a day), iterates to extract potential events within each day, and stores them in the temporary event list as a flat map entry. Next, all the events (in the temporary list) are assigned unique IDs and dates before the final validation. Lastly, the temporary list is appended to the main event list created in the beginning of the process and subsequently passed on to the ranking algorithm for further calculation. It is important to mention that the raw data stays untouched during the event extraction.

[1] An activity that takes place at a physical location (including valid GPS coordinate) at a particular date and time

Algorithm 5.1: Event extraction algorithm

```
input  : Pre-processed and enriched data imported from the application
output : Events data (event)
Function detectEvents(data, requirements)
    event ← list();
    for i in data do
        check each elements of data[i];
        tempEventList ← list();
        for j ← 0 to j < data[i].length do
            potentialEvent ← extractPotentialEvent(data[i][j]);
            date ← getDate(data[i]);
            if potentialEvent meet requirements then
                potentialEvent ← select-placeName(potentialEvent);
                potentialEvent ← standardise(potentialEvent.timestamp);
                tempEventList ← append(potentialEvent);
            else
                continue;
            end
            tempEventList ← removeDuplicate(tempEventList);
            foreach element in tempEventList do
                assignID(element);
            end
        end
        validateEvent(tempEventList);
        addDate(tempEventList, date);
        event ← append(tempEventList)
    end
    return event;
end
```

An example of the extracted events is shown in Snippet 5.1. Note that the place annotation remains within the data in order to change the selected annotation name or category by the user based upon the calculated list.

5.3 Event Ranking Model

The event consists of an activity that takes place at a physical location on a particular date and time. The data includes a semantic name and category annotation and is analysed based on hours to generate the event matrix. The variety of event frequencies required by this model is calculated from this event matrix.

```
    country_probability:Array(1),
    town_probability:Array(3),
    street_probability:Array(5),
    annotations:Array(10),
    identifier:"51.877937, -0.411429",
    local_id:12Y0234E22HA2349,
    location:{
        address: "Park Square, Luton, UK",
        postCode: "LU1 3JU",
        town: "Luton",
        country: "UK",
        coordinates: [Lat,Lng]
        },
    name:"University of Bedfordshire -- Luton Campus",
    category:"School & University",
    start_time:"2016-03-01T08:30:03+00:00",
    end_time:"2016-03-01T17:51:23+00:00",
    timestamp:1488326403031
```


SNIPPET 5.1: An event extracted by the algorithm

A day event matrix D_{ij}^{d} is created to record the activities that took place on day d, where i ranges from 0 to 23 representing the hours of the day and j is one of the top-level categories, respectively. The created matrix is similar to equation 5.1.

$$D_{ij}^{d} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,j} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i,1} & a_{i,2} & \cdots & a_{24,10} \end{pmatrix} \tag{5.1}$$

To complete the event matrix from the input data, the events are extracted based on their category types, and start and end times. The event takes a float value between 0 to 1 in the matrix D_{ij}^{d}. For instance, if the activity (within the particular category) started at 09:30 and ended at 10:15, then the corresponding entries are the float value of 0.5 for hour 9 ($i = 9$) and 0.25 for hour 10 ($i = 10$), respectively. Subsequently, an overview for month matrix m is generated in harmony with the daily matrix as follows:

$$M_{ij}^{m} = \sum_{d} D_{ij}^{d} \tag{5.2}$$

where the variables are as defined in equation 5.1. Next, by using the above values, σ_j^m can be computed, which represents the measurement of the overall j activity in the categories at month m in 24 hours:

$$\sigma_j^m = \sum_{i} \sum_{d} D_{ij}^{d} = \sum_{i} M_{ij}^{m} \tag{5.3}$$

Subsequently, the overall level of each activity (in categories) over the year Y_j^y can be computed by aggregating the calculated month values σ_j^m as follows:

$$Y_j^y = \sum_{m} \sum_{i} M_{ij}^{m} = \sum_{m} \sigma_j^m \tag{5.4}$$

Similarly, the overall measurement of activity j at year y for all activities is calculated as follows:

$$\sigma^y = \sum_{i} Y_{ij}^{y} \tag{5.5}$$

In order to provide an activity accumulation for day, month, or year, the activity category j is aggregated according to its monthly value, followed by normalisation as follows:

$$CIMP\_MH_j^m = \sigma_j^m \tag{5.6}$$

$$\sigma_j^m = \frac{\sigma_j^m - \min(\sigma^m)}{\max(\sigma^m) - \min(\sigma^m)} \tag{5.7}$$

where $sigma_j^m$ represents the normalised severity degree of the category of activity j in month m. Similarly, the severity for each activity category j at year y can be calculated according to its yearly value, followed by a normalisation:

$$CIMP\_YR_j^y = \sigma_j^y \tag{5.8}$$

$$\sigma_j^y = \frac{\sigma_j^y - \min(\sigma^y)}{\max(\sigma^y) - \min(\sigma^y)} \tag{5.9}$$

Two lists M_{month} and M_{year} are formed to accommodate all the monthly and yearly severity hours, respectively. These lists are created by scanning each day in D_{ij}^d and obtaining the maximum value within each hour. As a result, both lists include all of the hours during which the important activities occurred. This result is explicitly used for one of the use cases in this research – LifeTracker in Chapter 7.

Furthermore, in order to calculate a significance ranking score for the events, the model considers three customisable factors: the category, frequency and regularity of the events. These factors include an initial default value but can be altered by users according to their preferences through the user interface and interaction described in Chapter 6. The importance of these factors is expressed in three weighting coefficients in the model:

- w_1: The weight of importance for event regularity
- w_2: The weight of importance for event frequency
- w_3: The weight for event category as a point of interest

The mathematical calculation of significant event ranking is depicted in equation 5.10.

$$S(e|d) = (1 - w_1) \times \left(\text{tf-idf}_e \times \text{influence}_{ctg}\right) + w_1 \times \hat{R}_e \tag{5.10}$$

Correspondingly, the ranking score is calculated by incorporating the following three components:

1. **Event tf-idf score**: To suppress trivial events and extract significant ones, the $tf - idf$ concept from text mining is borrowed and endowed with event mining semantics [27, 125, 170]. The frequency and inverse event frequency scheme are used to score the events according to their frequency and uniqueness, using the following equation:

$$\text{tf-idf}_e = \left[\log(1 + \frac{\sum t_i^e}{d \in D})\right]^{w_2} \times \log\left(\frac{|D|}{1 + |\{d \in D : t_i^e \in d_i\}|}\right) \tag{5.11}$$

   where tf_e, as the event frequency consists of the $\sum t_i$, is the number of times that the event has occurred within a time period, $d \in D$ is the total number of terms (events) within the selected time period, and w_2 is the customised weight of the event frequency.

   The inverse event frequency idf is computed as the total number of days in the selected period (e.g. year or month) against df_e, which is the number of days that the event has occurred in the period. Within the equation, the numerator D refers to the total available days or cardinality of D and can also be interpreted as $D = d_1, d_2, , d_n$ where n is the number of days in the selected time period of events. The denominator $|d \in D : t \in d|$ indicates the total number of days in which term t, here known as event, has occurred within the time period (the $d \in D$) which enforces the event to be in the current time period space only.

2. **Regularity of event**: $tf - idf$ weighting represents the global frequency of the events but cannot reflect the magnitude of changes of the data along the time axis. A rare but sudden change, such as moving house, changing work or workplace, is highly valuable in personal event mining but cannot be discovered efficiently by frequency-based $tf - idf$ measures due to their exceptionally low event frequency. To uncover theses sudden changes, the

regularity of the events is modelled by using the variance of their occurrence. More specifically, the variance is modelled based on the change of monthly occurrence in the previous 12 months.

$$\sigma_e^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \tag{5.12}$$

where x_i is the number of occurrences in the month (i), $\overline{x}$ is the average occurrence, n is the total number of months involved in the calculation for variance. In this work, n is set to the last 12 months as this is a reasonable length of time period to estimate the regularity of an event as users are more interested in event regularity in recent years rather than in earlier history. Based on the variance, the regularity $\hat{R}_e$ of the event is determined as follows:

$$\hat{R}_e = \left[\frac{\log\left(\frac{|D|}{1+|\{d \in D : t_i \in d_i\}|}\right)}{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} \right] \times \left(\frac{1}{1+|\{d \in D : t_i \in d_i\}|}\right) \tag{5.13}$$

Written as:

$$\hat{R}_e = \left(\frac{idf_e}{\sigma_e^2}\right) \times \left(\frac{1}{df_e}\right)$$

where (idf_e) is the inverse event frequency, (df_e) is the number of days that the event has occurred in the period, σ_e^2 is the calculated variance. In essence, the above empirical calculation of the regularity is in favour of infrequently but regularly occurring events by involving (idf_e) and (df_e) as the numerators in the equation.

3. **Category influence factor**: as the daily tracking data is semantically enriched with category labels, the category information plays an important

role in event pattern mining. However, categories should not be treated equally in weighting. The category scoring follows the $(tf - idf)$ scheme. But here instead, the duration of all occurred events in the same category is considered [2] rather than event occurrence, and the score is weighted by w_3, which is the weighting coefficient for the event category:

$$\begin{aligned} \text{influence}_{ctegory} = \log(1 + \frac{\sum Cat_{dur,i}}{Cat \in D}) \\ \times \log\Big(\frac{|M|}{1 + |\{m \in M : Cat_i \in m_i\}|}\Big) \times w_3 \end{aligned} \tag{5.14}$$

Where the $(\sum Cat_{dur,i})$ is the category's duration of the occurred events while the $(Cat \in D)$ is the total duration of the entire events' categories both within the time period. Moreover, the term $(|M|)$ indicates the number of available month and the denominator $(|\{m \in M : Cat_i \in m_i\}|)$ denotes the total number of available months that the category has occurred within the selected time period. Ultimately, the multi-significance ranking score of each event is calculated by exploiting the aforementioned event $(tf - idf)$ score, regularity of event, and category influence factor together with user adjustable weightings coefficients (w_1, w_2, w_3), defined as follows:

$$\begin{aligned} S(e|d) = (1 - w_1) \times \Bigg[\Big(\log(1 + \frac{\sum t_i}{d \in D})\Big)^{w_2} \times \log\left(\frac{|D|}{1 + |\{d \in D : t_i \in d_i\}|}\right) \\ \times \log(1 + \frac{\sum Cat_{dur,i}}{Cat \in D}) \times \log\Big(\frac{|M|}{1 + |\{m \in M : Cat_i \in m_i\}|}\Big) \\ \times w_3\Bigg] + w_1 \times \left[\left(\frac{idf_e}{\sigma_e^2}\right) \times \left(\frac{1}{df_e}\right)\right] \end{aligned} \tag{5.15}$$

[2] Considering the *duration* rather than the *occurrence* of categories has a great impact on the influence factor and can regularise the weight between them. For instance, the influence factor for a workplace category, which normally occurs once a day with a long duration (e.g. 7 - 8 hours) and repeated every day, cannot be the same as the category including short events duration (e.g. 1 hour) with similar or higher occurrence.

Algorithm 5.2: Significant event ranking algorithm

input : Extracted events (flatData)
output : Events (flatData) with added significance ranking score

Function significantRank(flatData, externalFactors)
  $[w_1, w_2, w_3] \leftarrow$ extract(externalFactors);
  foreach *event* **in** flatData **do**
    simEvent $\leftarrow$ findSimilarEvents(*event*);
    eventOccurrenceDay $\leftarrow$ findEventOccurrence(*event*, *day*);
    eventOccurrenceMonth $\leftarrow$ findEventOccurrence(*event*, *month*);
    eventDuration $\leftarrow$ GetEventDuration(*event*);
  end
  dfDictionary $\leftarrow$ calculateIDF(eventOccurrenceDay, eventOccurrenceMonth, eventDuration, flatData *.length*);
  forall *j* **in** flatData **do**
    eventTermFrequency $\leftarrow$ calculateTermFreq(flatData [*j*]);
    $\mathbf{tf - idf_e} = (\log(1 + \mathbf{eventTermFrequency}))^{w_2} \times \log(\text{number of available days} \div \text{dfDictionary}(\text{flatData}[j].name)$;
    eventRegularity($\hat{R}_e$) $\leftarrow$ computeRegularity(flatData([*j*]));
    catInfluFactor $\leftarrow$ influFactor(flatData[*j*].*category*, eventOccurrenceMonth, w_3);
    significantScore $= (1 - w_1) \times (\mathbf{tf - idf_e} \times \text{catInfluFactor}) + (\text{w}_1 \times \text{eventRegularity}(\hat{R}_e))$;
    flatData[j] $\leftarrow$ append(**significantScore**);
  end
  return flatData;
end

The implemented algorithm for the significant event ranking is shown in Algorithm 5.2. The result of this model is a significance score that is pushed into each event as a float value between (0.0) to (1.0). Subsequently, events are stored as flat data to increase the performance of retrieval for the visualisation. This assures that the data comes with the least complexity and unnecessary attributes whilst the original data is kept unaltered. The final look of the extracted events and their calculated scores is shown in Snippet 5.2.

5.4 Evaluation and Benchmark

The significant event ranking model is examined to determine the quality of the results and the performance using real-world datasets. To this end, the initial benchmark was carried out to create the ground for future methods. The

```
    country_probability:Array(1),
    town_probability:Array(3),
    street_probability:Array(5),
    annotations:Array(10),
    identifier:"51.877937, -0.411429",
    local_id:"12Y0234E22HA2349",
    location:{
        address: "Park Square, Luton, UK",
        postCode: "LU1 3JU",
        town: "Luton",
        country: "UK",
        coordinates: [Lat,Lng]
        },
    name:"University of Bedfordshire -- Luton Campus Library",
    category:"School & University",
    start_time:"2016-03-01T08:30:03+00:00",
    end_time:"2016-03-01T17:51:23+00:00",
    timestamp:1488326403031,
    significance: 0.220391
```

Snippet 5.2: An annotated event with significance score

benchmark was conducted by this author's research centre's own implementation as well as the JsLitmus.js[3] JavaScript library. Subsequently, the quality of the ranking model was evaluated by asking the participants to examine the significance score of their own events produced by the model. Moreover, the model was assessed once more within the provided visual analytics tools such as MyEvents in this research.

However, as mentioned in Chapter 1, the limitation of time and having only a small number of real-life datasets with at least two years daily activities has influenced the process of evaluating the ranking model in this research. Therefore, the model was evaluated by using a limited number of datasets which may be, to some degree, controversial from the experts' standpoint. Acquiring further personal life data can push the boundaries and help to get more credible results by assessing the multi-level significance ranking model further.

The benchmark includes assessing the model and its implemented algorithms in JavaScript in terms of performance and efficiency. The benchmark is conducted on

[3] https://github.com/broofa/jslitmus

the Chrome web browser v58, Mac OS 10.12.4, with Intel Core i7 CPU with 32GB memory and onboard Intel 4000 GPU. Similar to the automated place annotation, the performance and efficiency are conveyed by measuring the maximum iteration of the implemented algorithms per second and also measuring the completion time for different sizes of datasets.

The result of the benchmark shows that the time complexity of the data processing algorithm including the event extraction is $O(n)$. This means that by increasing the number of days N the execution time grows linearly – see Figure 5.3. Subsequently, the significance ranking model outperforms the event extraction with a complexity time of $O(\log n)$ – see Figure 5.4. The result indicates that by increasing the number of days for calculation, the execution time expands logarithmically.

Correspondingly, the maximum operations per second of the implemented algorithms in the model manifest an outstanding performance despite an immense data size and complexity of the pseudocode. Figure 5.5 portrays operations per second of the data processing and the ranking model separately over a number of tests.

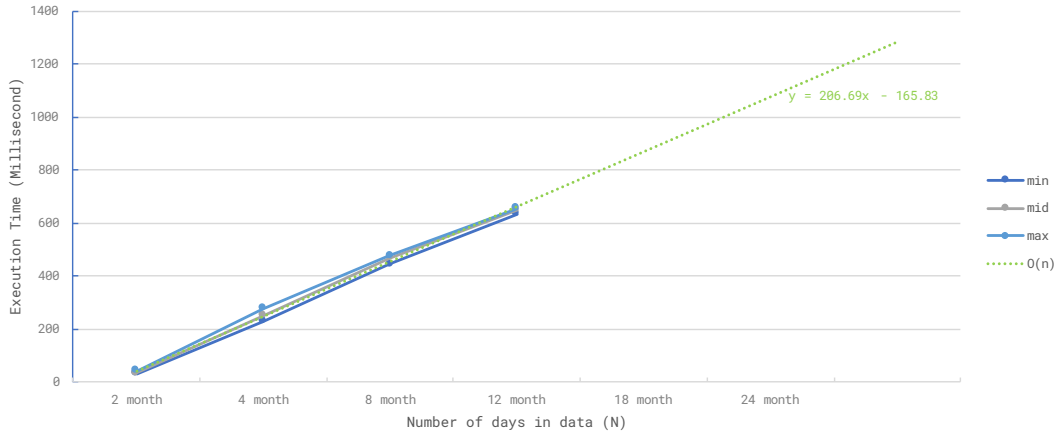


FIGURE 5.3: Data processing time complexity and performance including the data cleaning and event extraction

The quality of the model is independently assessed by applying the significance ranking model to all individual real-life data in accordance with user preferences. The algorithm provides a list of top 20 significant events with calculated ranking

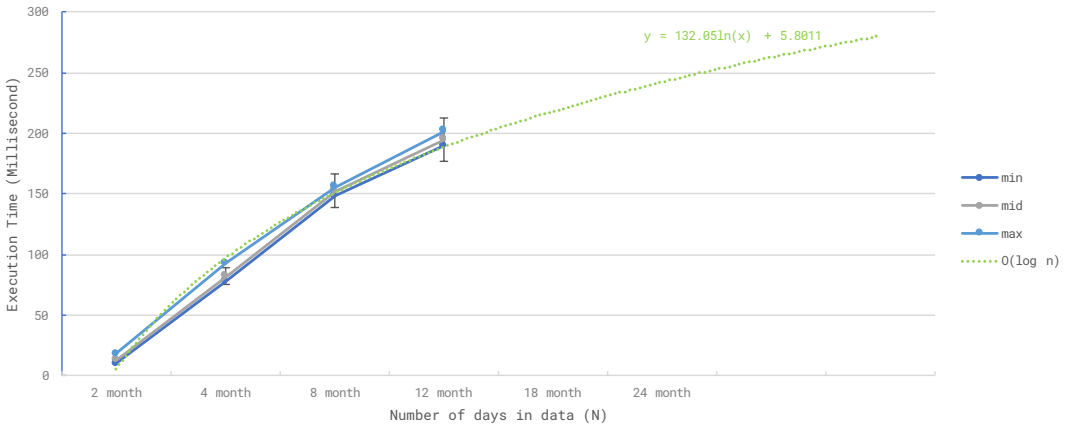


FIGURE 5.4: Multi-significance ranking model time complexity and performance

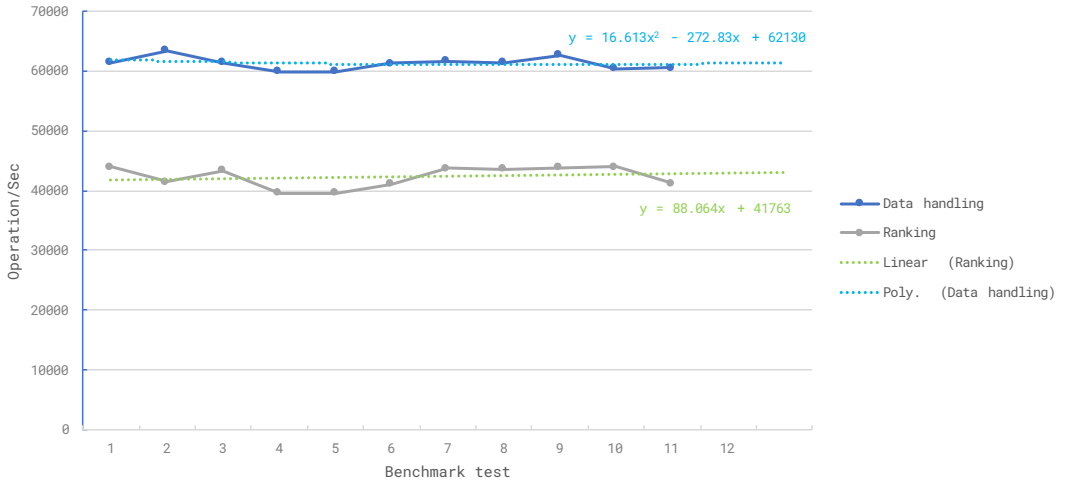


FIGURE 5.5: Efficiency of the data processing and multi-significance ranking model based on operations per second

scores in five separate instances by using different settings of preferences, such as regular places or a particular category as a point of interest. The participants were, then, asked to evaluate the calculated scores for each personal event based on their subjective assumption of importance by giving a rank score ranging from 1 to 5 to each event. For instance, the participants could provide a subjective rank of (4 or 5) to an event (e.g. related to travel, food, health, etc.) with the calculated score of (0.9) which indicates that the model provided a sensible score for a fairly significant event from their views. The score from the participants can be the minimum (not significant at all) or the maximum (very significant).

In total, each participant ranked 100 selected events of their own (20 events in 5 separate instances). The result of the evaluation by each individual is shown in Figure 5.7 by aligning the events [4] on the x-axis and the participants' ranking score on the y-axis.

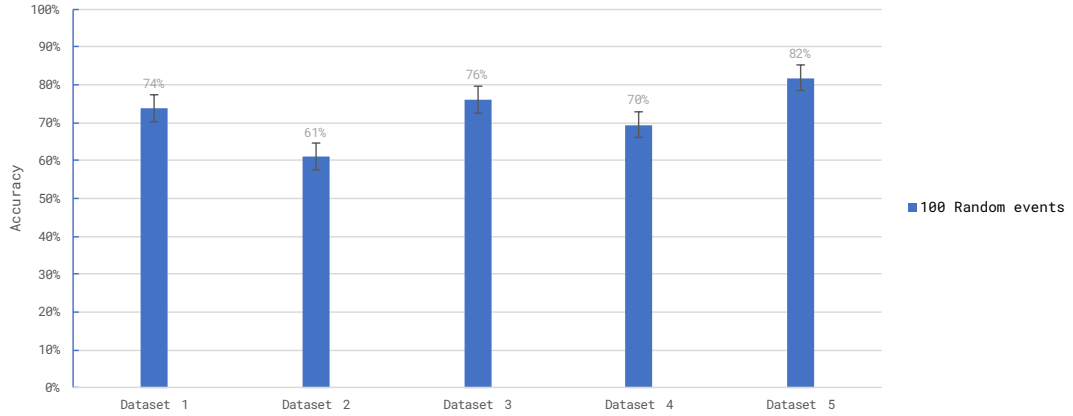


FIGURE 5.6: The accuracy of the results from the multi-significance ranking model

According to the result analysis, the significance ranking model can achieve an overall of 72.43% (SD=0.07) accuracy in determining significant events. The accuracy for each dataset are shown in Figure 5.6. The accuracy is calculated by considering the mean value of the given score to each event and its significance value for each participant.

5.5 Chapter Summary

This chapter introduced a novel multi-significance ranking model, one of the major contributions of this research (**C3**) that can be used in line with the personal daily life domain and identified significant events based upon user preferences. This model requires a set of fixed and customisable factors as coefficients to extract significant events. The model is composed of three different parts, namely $tf - idf_e$, event regularity, and category influence factor, all of which are considered in calculating a sensible score for each event. In addition, the model employs three

[4] The events' names are anonymised to comply with ethical requirements

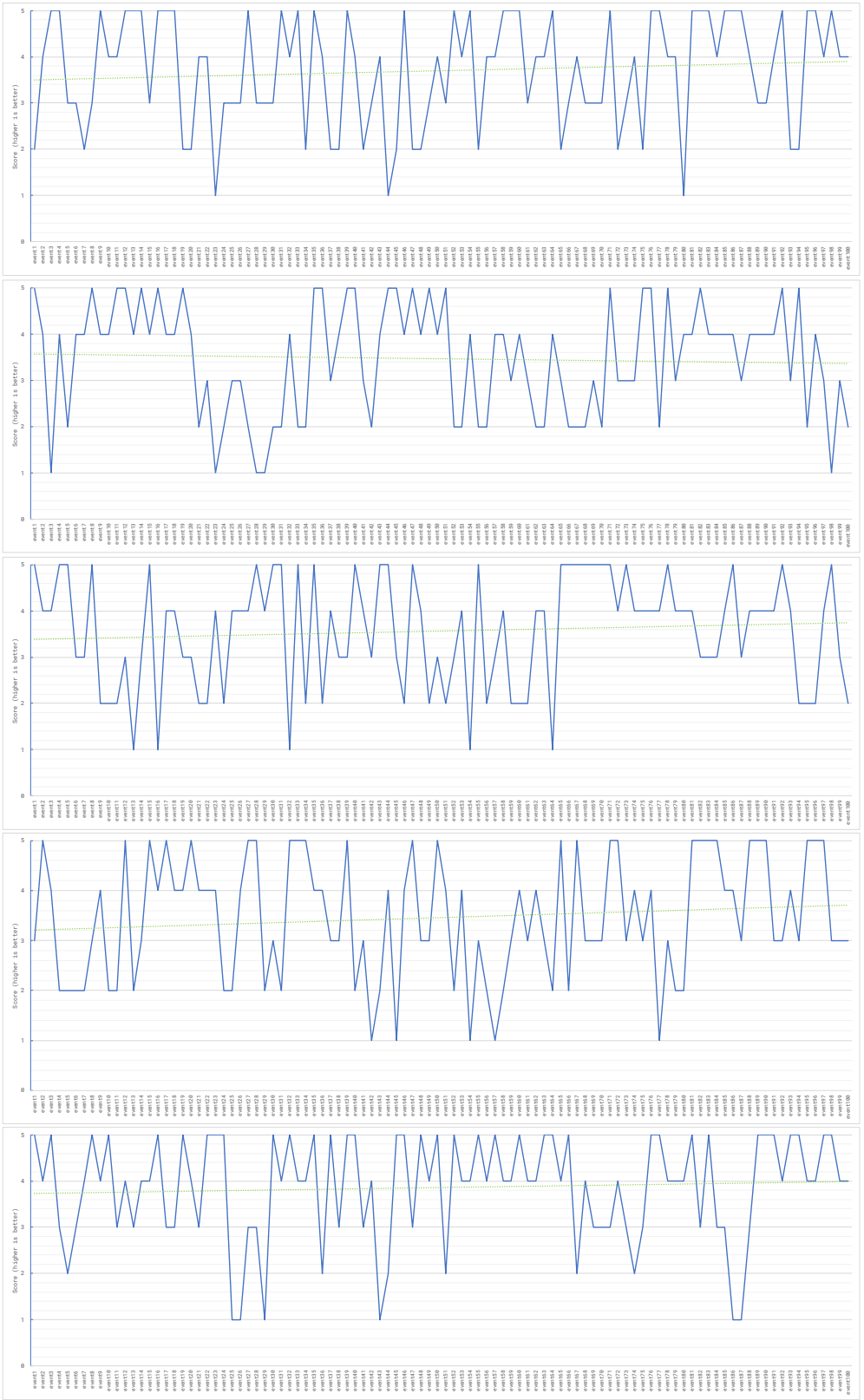


FIGURE 5.7: The quality of the results from the multi-significance ranking model in detail

external factors w_1, w_2, w_3 to reflect the user preferences within its significance calculation. Correspondingly, the model was assessed to determine the accuracy and time complexity of the implemented algorithms. The result of evaluation indicates that the ranking model time complexity is logarithmic $O(\log n)$, while the event extraction performs with $O(n)$. Moreover, the quality evaluation of the model, despite a limited number of real-life datasets, shows that the model delivers an adequate and acceptable significance extraction of 72.43% (SD=0.07) accuracy for the real-world data based on user preferences. In Chapter 7, this ranking model was used to interactively encode the personal data in relation to the extracted significant events and enable the users to understand their data based on their own grounds and preferences.

CHAPTER 6

Visualisation Design Components

6.1 Introduction

Visual analytics is highly dependant on interactive information visualisation to be able to deliver meaningful details of the analysed data and allow for acquiring adequate knowledge. The majority of the available scientific approaches nowadays attain this goal by means of either designing a new component or combining current visual analytics techniques. Despite the number of scientific visualisation approaches within the community, the number of approaches that target personal data – particularly the life logging data – is inadequate. This yields the need for an effective visualisation to deal with such data and facilitate the process of knowledge discovery.

In this chapter, the interactive information visualisation is completely developed. The method is described for how the extracted knowledge from the data mining models can be effectively visualised by looking into the four goals of effective visual design – ability to convey important information, control eye movement, capture attention, and evoke intended emotions – during the design process. We conduct

a design study and review the literature to obtain the key requirements and users needs, assess the design, and enhance the design using working prototypes.

6.2 Design Goals and Requirements

To design a collective visual analytics approach that can be adopted for different purposes within the area of personal daily life, strategies are employed to ensure that the design of the user interface is effective, intuitive, and perceptible. In addition, highly interactive tools are offered that can involve users within the process of self-knowledge discovery and visualisation. The design endeavours to strike a sound balance between the level of automation and user control, allowing users to effectively seek for desired information based on their own preferences with the support of the proposed data mining techniques. Moreover, this design must be optimised for use by non-expert users towards gaining meaningful understanding of different aspects of daily life with a minimal level of learning.

To fulfil the above objective [**OBJ8**], the key requirements are identified within the best practices for personal data [**OBJ7**] through literature review (described in Chapter 2) such as [53, 117, 186, 220, 229] and user requirement analysis (A). Each requirement indicates in what way the approach needs to accommodate the user needs towards achieving the main goal of interactive visualisation. The identified requirements are listed below:

- **Narrative Structure**: a good narrative visualisation should initially attract users, persuade them to interactively explore further, and conclusively provide an outcome to a clear goal. The settings, hence, must be well defined as they are crucial to ensure that users can adequately understand the visualisation with only basic details. The key point in narrative visualisation is to avoid ambiguity and provide multiple views with ample interactivity to facilitate the process of understanding, learning, and subsequently, knowledge discovery.

- **Consistency**: Consistency is essential in the visualisation design, particularly, when multiple views, faceting, filtering, and constructive interaction are involved. Each of these can distinctively represent various dimensions of the data. It is important to make sure every part of the visualisation has consistency with the established setting.

- **Stereotypes**: The visualisation can stand out and reveal unusual relationships or interesting facts and compel the users to re-assess a hypothesis, make a decision or take an action, and ultimately look for interpretation by means of interactions.

- **Analytical Methods (semi-guided exploration)**: The visualisation should be backed by compelling analytics methods to be able to deal with large-scale personal life data. The analytical methods accelerate the process of knowledge discovery by addressing the multi-dimensional personal life data. It is also important to extend the analytical capacity by offering a modification based on the user's preferences. Hence, the method should support the ability to filter or change the setting with respect to user choices.

- **Appearance and Visual Representation**: The visual representation of the data should follow a set of fundamental principles:

    1. *Practical fundamentals*: keep clarity, focus, and simplicity within the visualisation
    2. *Executive fundamentals*: quality and accessible exploration at any time with scalability in mind
    3. *Cognitive fundamentals*: accelerating insight, expediting attention, simplifying mental progressing, and helping memory.

- **Provide an overview**: Users need to be able to obtain general overview of the data at any exploration level within a selected period of time. The overview should be comprehensive and trigger the attention to any pattern or anomaly before a user drills down for more details in the data.

- **Multi-level view**: The approach should facilitate the visual understanding by providing various levels of views. This means that the process represents data in different fashion to allow users to delve into the data and gain meaningful understanding in an optimal way.

- **Search and control**: Visualisation should provide a search environment to accommodate the idea of searching by natural language and also help users to get initial insight during the search process. Some means of control should also be made available to allow users set targets and preferences via a graphical user interface and also to comprehend the available options within the analysis.

- **Events visualisation and highlights**: Space and time information in a chronological order should be made available in visualising the personal data. Moreover, it is important that the visualisation is capable of highlighting a set of or particular data points.

- **Filtering**: Visualisation is required to handle trivial data points to avoid overplotting and still be able to represent them on demand.

- **Correlation**: Showing the possible associations between the data points during the exploration allows users attain a deep level of information.

- **Contextual information**: Providing additional information in order to allow users to recognise their points of interest for the exploration within the process of knowledge discovery and delving into details.

- **Support multivariate and heterogeneous datasets**: The visualisation should support time series of heterogeneous sources.

- **Interaction**: This allows users to delve into the data based on their interests and carry out further exploration to gain better understanding of data. It can be by:

    1. Dynamic multi-foci: drilling down for more details
    2. Summary view to perceive the trend over the time

3. User-defined data abstraction
4. Maintaining the history of the exploration process

6.3 Visual Encoding

The visual encoding is a key part in visualisation. Employing suitable shapes, colours, glyphs, and the like leads to a successful visual design and notably increases the effectiveness of visual analytics approaches. This section describes the visual encoding design based on the nature of personal daily life and user requirements.

Here, the concept of preattentive processing is used to facilitate the process of visual understanding. Preattentive processing happens naturally prior to human conscious attention. This processing can be triggered by a certain shape or colour. A simple example to show how this process works is to count a certain number – here, 5 – within a bag of several numbers in Figure 6.1. To count the number in the Figure 6.1(a), it is necessary to scan the available numbers sequentially whereas in the Figure 6.1(b), you only need to scan the red digits within the rows of numbers. The reason behind reading Figure 6.1(b) more easily is that the colour is preattentively processed; hence, it enhances the response time and facilitate the process of understanding the distribution of the specific number. According to Ware [220], the study of preattentive processing shows that the response time to perceive an object that is preattentively different from its neighbourhood – with certain distractors – is greatly lower than the non-preattentive items. The study adds that the difference between the response time to the preattentive and non-preattentive items increases by increasing the number of distractors. Hence, the main reason to use preattentive processing within the information visualisation can be concluded to expedites the process of understanding information with many features.

Ware [220] classifies the features that can be preattentively processed based on colour, form, spatial position, and motion. In this work, a number of these features

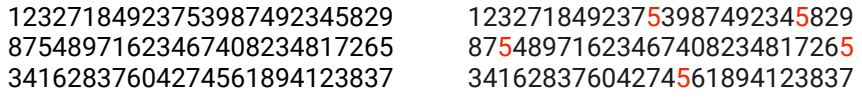
are applied to encode the information effectively and assist the user to get better understanding (Figure 6.2). The combination of the following features are applied to each visual component in this work:

- Shape
- Size
- Colour
- Colour intensity and Opacity
- Enclosure
- Addition
- Motion
- Highlighting
- Glyph

These features can be combined in different ways for use within the visualisation, similar to many works in this domain such as [25, 177, 233, 235].

Shape: circles along with lines envisage the daily life data in the majority of the visual components in this chapter. The circle is used to show places or physical activities as it is an optimum shape to show the great number of items in a single interface. Lines are selected to show duration and association as they can be easily interpreted by human cognition for that purpose.

Size: is an effective feature to show the difference between a number of encoded items within the visualisation canvas. The size can be appointed to duration, probability, weight, etc. However, using different sizes within a large-scale dataset might lead to some degree of uncertainty as the process of judging the size becomes problematic when involving the user's verdict. Therefore, it is vital to use the



(a) Counting the 5 in this table requires scanning all the numbers sequentially

(b) To identify the 5 in this table, scanning only the red digits is sufficient

FIGURE 6.1: Preattentive process example

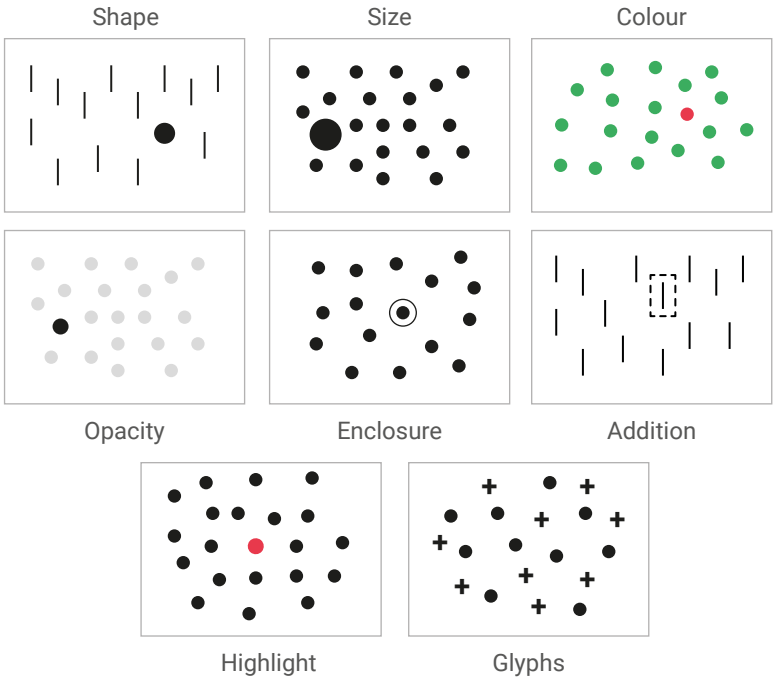


FIGURE 6.2: Preattentive features used within the process of visualisation in this research

size in conjunction with the other features in order to deliver an uncomplicated encoding.

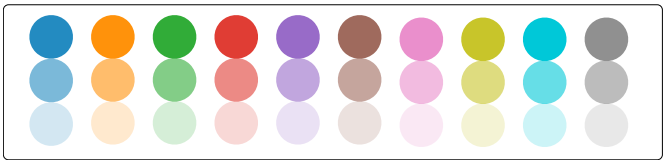
Colour: is the main part of the preattentive process. Using distinct colours in line with daily life activities is one of the key challenges in envisaging the activities, including the places and movements. Chung et al. [49], Ware [220], West et al. [225] argue that using more than eight colours is not suitable for the preattentive process as it is difficult to distinguish the differences. However, at least 10 colours are needed here based on the data, so the successor colour schemes are adopted from D3[1], Google[2], and ColourBrewer[3], which can be fit this design study. Four sets of distinct colours were evaluated to determine the most suitable and distinguishable set by conducting a user study in this research – see Figure 6.3. The result of this user study is documented in the design evaluation in this chapter. The study shows that majority of the users can differentiate *colour scheme set A* with different intensities. Correspondingly, these colours were tested against black as well white

[1] https://github.com/d3/d3-scale
[2] https://material.io/guidelines/style/color.html
[3] http://colorbrewer2.org/

FIGURE 6.3: The four-colour scheme selected for the user study in this implementation, with application of different intensities and opacities

backgrounds with exactly the same settings to determine on which background the colours are most perceivable – see figure 6.4. Based on the result, the eligible colour scheme that met requirements was selected – see Figure 6.5.

(a) The colours against the dark background suffer loss of contrast which prevents the low intensity colours from being clearly recognisable



(b) The colours (particularly the low intensity) can be perceived better against the white background

FIGURE 6.4: The set of 10 distinct colours against the dark and bright backgrounds

Colour intensity and Opacity: are used to visually filter out series of objects within the visualisation. For example if the data consists of various places and the user ought to see more important places then a lower colour intensity together with lower opacity can be applied to the trivial places. As a result, this increases the user's focus on the preferred information. Moreover, this feature can be used

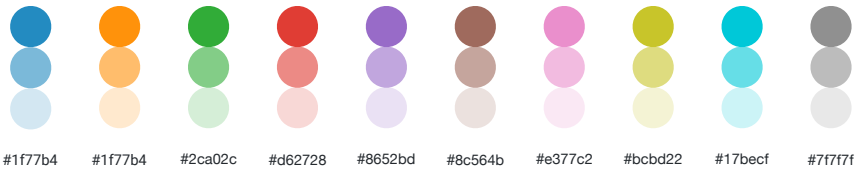
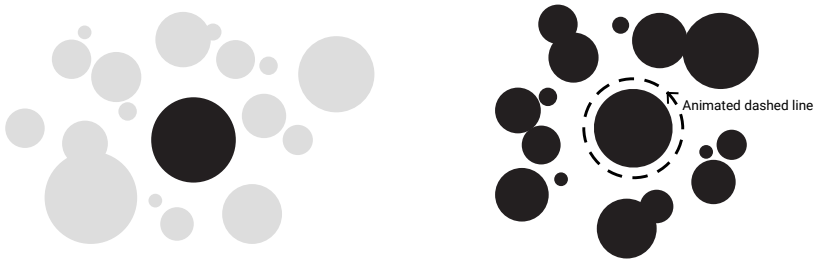


FIGURE 6.5: The colour scheme used in this implementation

during the interaction phase when the user seeks additional information regarding the particular visual object within the process of exploration and knowledge discovery.

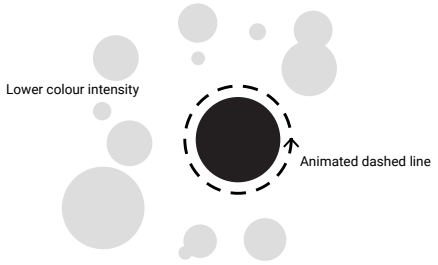
Highlighting: is the most favourable way to make a particular part of the information stand out. This feature can address simply the preattentive processing in the homogeneous graphical display (e.g. highlighting the text on paper). However, highlighting within the visually complex environment – with numerous shapes, colours, and textures – can become challenging. For instance, the targeted object can be highlighted by: 1) reducing the opacity or colour intensity,in which may lead to a tentative highlighting when the visualisation includes numerous contrasting colours (Figure 6.6(a)), and ; 2) using a transition or an addition such as a dashed line to trigger attentiveness (Figure 6.6(b)), but it may not be visually compelling within a dense area of visual objects. To tackle this issue and provide an effective highlighting process, a mixture of visual features is used such as lower opacity, transition, and addition (dashed line) – see Figure 6.6(c). By doing so, the users benefit from the more comprehensible visualisation with the bold features highlighted. It is worth mentioning that the highlighting process used in each part of the research follows the same principles but these may be slightly recast to deliver a best result.

Glyphs: Using well-defined glyphs can facilitate the effective visual communication due to their ability to employ a different visual channel, i.e. shape, colour, texture, size, location, and orientation Sanyal et al. [185], Wittenbrink et al. [230]. To this end, a number of glyphs were designed in accordance with the nature of the personal daily data and the essential design principles defined by Chung et al. [49]. During the course of designing the glyphs, we consider visual orderability,



(a) The object stands out by reducing the opacity of the rest of the visible objects

(b) The animated dashed line shows which object is highlighted



(c) The combination of opacity, animation, and dashed line makes the highlighting significantly effective.

FIGURE 6.6: Highlighting process using motion, line, and opacity

channel capacity, searchability, learnability, and attention balance. Orderability corresponds to the ability of sorting the glyphs perceptually by establishing a uniform rule, while searchability refers to the ability of looking for a particular item amongst the others within the visual encoding. Learnability assists the user to understand and memorise the meaning of the glyphs with minimum need for a glyph legend. The attention balance corresponds to preventing the encoding from unbalanced attentiveness amongst the others by using moderate shapes and colours.

Ten glyphs were designed, reflecting the personal life place categories and four glyphs for showing the recorded physical movements – walking, transport, running, and cycling. Three features are utilised in designing the glyphs, namely, colour, size, and position. The glyphs benefit from single circle shape, different sizes, consistent colour scheme, and set of icons[4] related to the category. Figure 6.7 shows the glyphs.

[4] Icons made by Freepik from www.flaticon.com - (cc by 3.0)

FIGURE 6.7: Glyphs designed to reflect the movement and place categories

6.3.1 Visualisation components

The process of designing the visual components within this work are explicitly described in this section. These components are developed in the light of preattentive processing and continuously evolved and enhanced according to the results from the evaluations. All the components are formulated to be customisable and employed by different techniques towards achieving various goals within the area of personal daily life. This allows the visualisation tools to be adopted for different types of knowledge discovery while their effectiveness remains.

The visual components are evaluated at every stage of this work to ensure they are logically effective in carrying the message and can meet the requirements. These components, subsequently, are deployed and evaluated in three experimental, integrated visual analytics tools (platform) in Chapter 7.

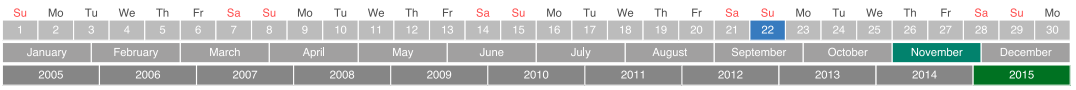
Linear and multi-layered timeline

The time-oriented approach aims to effectively support multifaceted tasks and visualise multivariate temporal data with an adequate flexibility and analytical support in promoting exploration [7, 10, 202]. One of the main components to

support the navigation of such data, is a timeline [64, 164]. The timeline allows laying out data linearly based on their associated times. This component can be also utilised in line with the other dimensions of data to envisage trend and pattern over the time period [57, 122]. It is fair to say that, in visualising the temporal data, the use of a timeline is inevitable.

The timeline is presented as a vertical or horizontal axis that shows the data according to their associated time information. In this research, two interactive timelines are implemented – the linear timeline and the multi-layered timeline. These two types can be plugged into different data visualisation approaches to project the temporal data. The main timeline designed (inspired by Dachselt and Weiland [57]) for exploring the purpose of temporal exploration is not only linear, instead, it incorporates three different layers – year, month, and day – representing the time at different scales - see Figure 6.8. This timeline supports selecting the time range in a semi-guided way. In each layer, the corresponding year, month, or day appear as an interactive tile which can be selected to change the time range and narrow down the visualisation result correspondingly. In addition, the tiles within the timeline have the ability to be highlighted in order to represent any indication about the selected period.

(a) Different layers of the multi-layered timeline – day, month, and year layer



(b) Multi-layer timeline used for selecting the range of time

FIGURE 6.8: Multi-layered timeline structure

The timeline is designed to be highly interactive by performing the following functions in order to derive the data at different time periods:

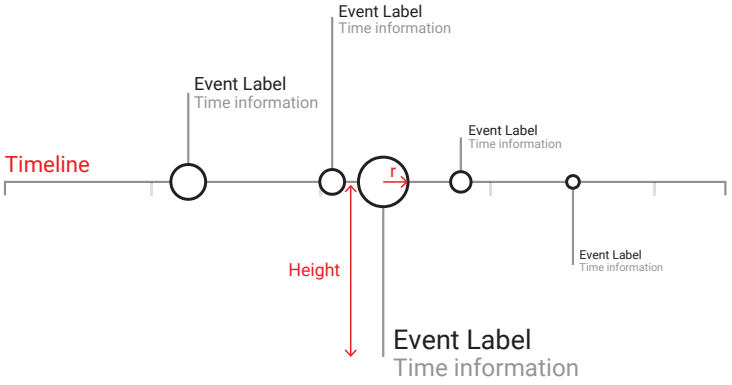
- **YearMode** via year selection – a year selection can be made by clicking on the year layer on the timeline. By this, the month layer is activated and the associated time axis is set to the YearMode to display all the corresponding temporal data across the 12 months of the selected year.
- **MonthMode** via month selection – a month selection can be made by clicking on the month layer of the timeline following the selection of the year. By this, the day layer is triggered and also the time range is limited to the selected month.
- **DayMode** via day selection – a specific day can be selected by clicking on the day layer following the selection of the month and year. The timeline can pass the day to the corresponding visual components for envisaging the related data.

Storyline

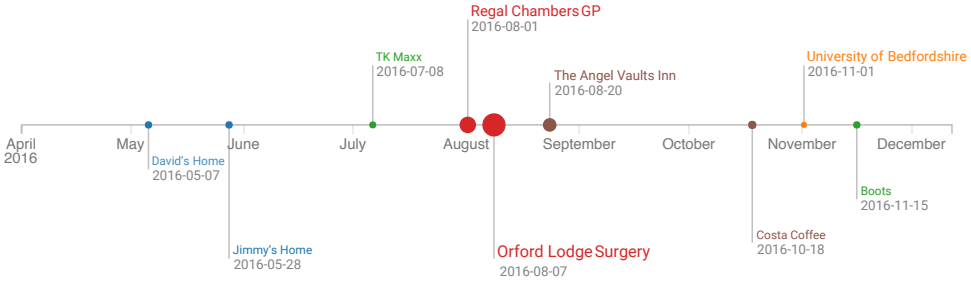
Storyline is another form of encoding temporal data over the linear timeline [202, 204]. This form, similar to the other components, employs a set of visual properties such as colour, size, shape, and height to encode the time-oriented information by considering the preattentive process. This component is designed to represent particular events and their affiliated information that is obtained via the data mining process and selected by the user during the process of exploration and knowledge discovery by means of different visual attributes, namely, circles of different radius, different size of labels, colours, and various lengths of lines over the timeline – see Figure 6.9. For instance, a significance ranking score of the event can contribute to the scale of the radius in a circle to allow the visualisation to emphasise events with main importance. Furthermore, labels are used to give a brief hint of events by displaying the name and time information. The colour identifies the associated category to which the event belongs. The length of a line can be the combination of event duration and its significant score in which the more significant event stands out further. All the events are laid out over the timeline according to their time. The lines and labels which are connected to

the circles can be vertically positioned either on the upper or lower part of the timeline to avoid overlapping of the labels.

Correspondingly, the interactions in the storyline is designed to facilitate the process of discovery or evocation. The interaction allows the individual to zoom, pan, and get additional information by means of an informative tooltip – see Section 6.3.1.



(a) The radius and height of each event circle are reflected by the significance ranking score and the duration. Labels are of different size according to the importance of the event.

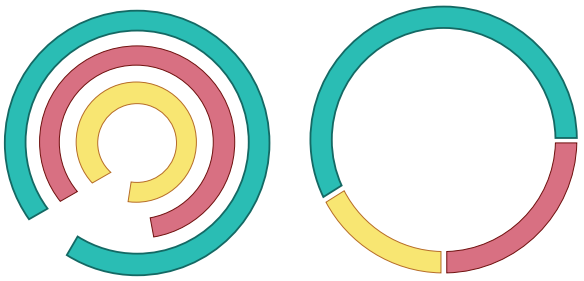


(b) A real example of storyline with added colour scheme.

FIGURE 6.9: General structure of the storyline and a real-world example

Circular-based visualisation

To create a hybrid layout that can facilitate a better grasp of the daily activities based on the continuous (ordered) and cyclic structure of the daily life data, a circular layout is defined that can accommodate the cyclic formation of the two main sets of daily life activities – movements and places. The circular layout



(a) The stacked layout can fit the data with same position

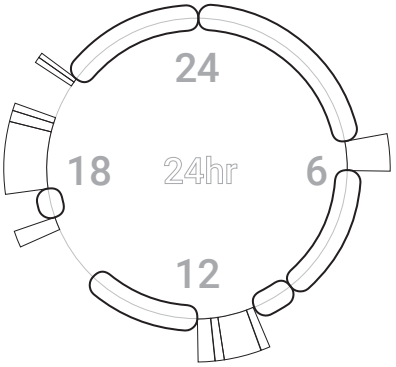
(b) The sequenced layout is suitable to show the sequences but not the repeated items

FIGURE 6.10: Two potential circular layouts to visualise the information

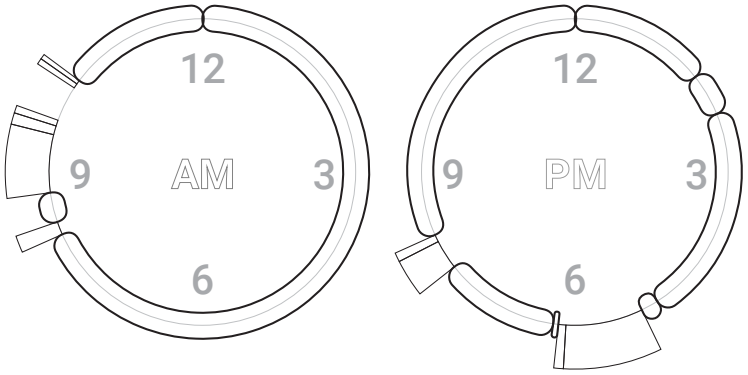
(stacked or sequential), according to [63, 150], is suitable for visualising the time-series ordered structures (Figure 6.10(a)). Another reason to select the circular layer is that users are familiar with the traditional clock metaphor. This results in an effective visualisation and hence gains more valuable knowledge. However, using solely the sequenced form can lead to visual clutter whilst the multiple series of data with different time intervals need to be accommodated in the same layout. Therefore, a mixture of stacked and sequential circular layouts are exploited to display the sequenced data on two individual rings without any interference.

A combination of stacked and sequential circular layouts inspired by [52, 63, 190, 239] is used to create the circular-based visualisation for representing the daily activities. To divide the activities into movements and places, two sets of arcs with different positions and size of radius are employed. The places are shown with narrow arcs while the wider arcs are allocated to movements. By this, the movements and places are differentiated, which prevents the users from any misinterpretation.

The challenge is that all the daily activities including movements and places occurring during the 24 hours can be laid out along a single circle (24 hour) or two individual circles (AM/PM) based on their classification, start and end times. These two layouts were created and evaluated to determine the more effective and natural way of such representation; the evaluation results can be found in Section 6.4).



(a) The single ring with 24 hours of activities



(b) The dual rings with 12 hours of AM and PM

FIGURE 6.11: Two different circular layouts for representing daily activities

The first layout was a single ring containing 24 hours in which the activities are placed based on their timestamp information. The second layout was a combination of two 12-hour rings (AM/PM) representing all the activities individually, similar to the traditional clock metaphor - see Figure 6.11. Based on the evaluation result, the dual rings layout (AM/PM) was employed for the circular-based visualisation. This layout provides a more comprehensible way of representing the data in accordance with human perception. Moreover, the human brain, to a great extent, is familiar with the traditional clock and position of hours.

Several techniques are employed in this approach to deliver an effective visual representation of daily activities. The circular layout uses a movement/place classification, a consistent colour scheme, a relevant set of glyphs and an informative tooltip. The circular layout classifies the daily activities into places and movements.

The process of encoding the activities on the circular layout is based on the start and end times. The movements are places along the outer radius of the circle with the greater width whereas the places are laid out on the radius with the narrower width to show the difference between movements and places. The position of movements and places are calculated in the same manner. As each circle represents 12 hours of a day, the activities between 00:00 hr and 11:59 hr are grouped together for the AM circle and the rest are associated with the PM circle. The start and end times of each activity are converted to a decimal value and the length of the arc is computed by determining the difference between the start and end times. The length of the arc within the circle with known radius (r) is calculated by the equation 6.3.

$$T_{converted} = T_{hour} + \frac{T_{minutes}}{60} + \frac{T_{seconds}}{3600} \tag{6.1}$$

$$L = |T_{end} - T_{start}| \tag{6.2}$$

$$\begin{aligned} \frac{L}{circumference} = \frac{\theta}{2\pi} = \frac{degree}{360^\circ} \\ \theta = \frac{L \times 360^\circ}{circumference} \end{aligned} \tag{6.3}$$

By knowing θ, the end point can be positioned and the arc is completed – see Figure 6.12.

The colours used in the circular layout are consistent with the colour scheme. However, if the data comes with undetermined places, the circular layout shows them by only an outline to notify the user. Moreover, a set of related glyphs for movements and places is borrowed to add adequate means to support users to apprehend the visual encoding with closely packed activities. The design of this method is shown in Figure 6.13.

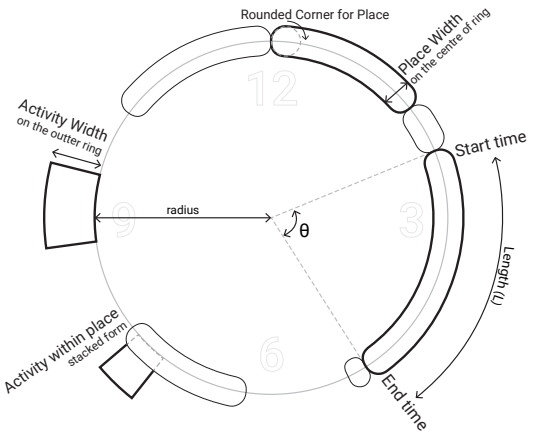


FIGURE 6.12: Structure of the circular layout

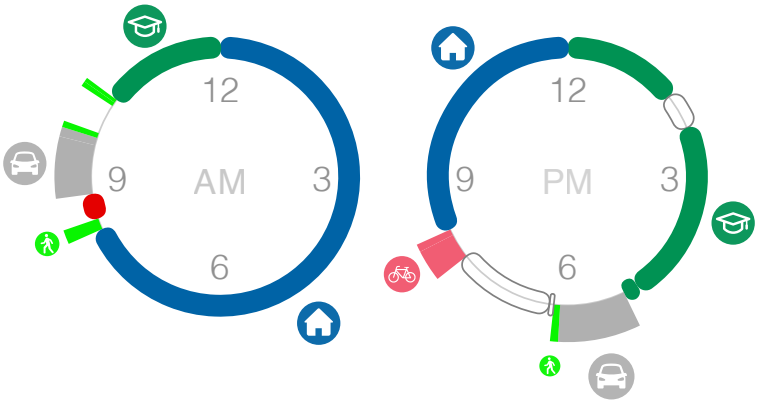


FIGURE 6.13: The circular-based layout shows daily activities by employing the colours and glyphs

This approach, similar to the others, offers interaction to accelerate the ability of perceiving the relationship between the data in depth, and subsequently gaining new insights. This interaction is shown in more detail with the designed platforms in Chapter 7.

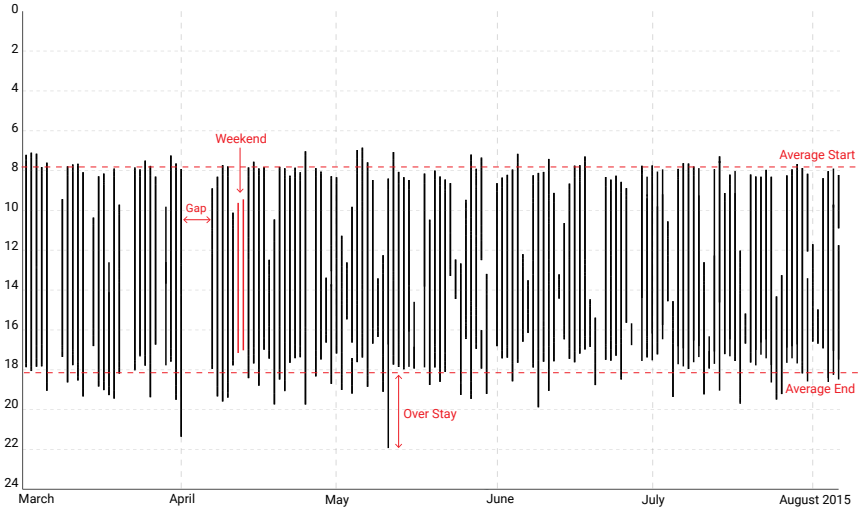
24-hour event visualisation

Personal daily life data embrace a great number of events including movements and places that can be used to study individual behaviour, determine a particular pattern, and discover hidden knowledge. These temporal data are concurrently sequential (linear time information, e.g. from 2013 to 2017) and cyclic (repeated

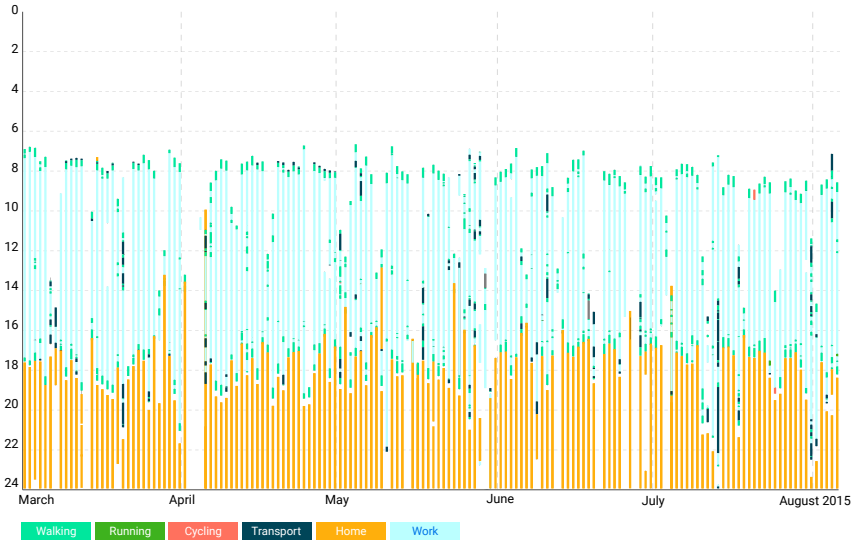
cycle of days, weeks, months, seasons, and years). Hence, providing a meaningful representation of this data with such characteristics requires robust settings. Time can be, similar to many approaches, shown on the x-axis by means of the timeline, whilst the other dimensions can be represented by using the y-axis or different visual properties such as colour, size, and shape within the visualisation medium. However, using only a single timeline may not cover all the characteristics of the data (e.g. cyclic features). To this end, the 24-hour event visualisation is designed so that it represents the events in a grid structure according to their time and duration. The horizontal axis (x-axis) denotes day, month, and year, while the vertical axis (y-axis) indicates the 24-hour time of day. According to the literature review, in many previous works [23, 85, 115, 151, 190, 219], a straight line is a natural choice to represent an event. The line can indicate start and end times (by position), the duration (by length) and the other dimensions (by width).

Using the line to only show the event occurrence based on its duration is a common practice and preattentive in the visualisation context. This feature can be used to show the cyclic feature of the data against the 24 hours, week, month, and the like. In this research, the same concept is used to show solely the events by laying out the line on the timeline (x-axis) and the 24 hours (y-axis) to show the occurrence linearly and cyclically, respectively. Additionally, the same colour scheme is used for the events in accordance with their category classification. The result of visualising the event by such a concept is shown in Figure 6.14. This figure represents two different views of the data: 1) work-related events and 2) the overall events of an individual within the particular time range. By looking at the visualisation in Figure 6.14(a), inherently, the general pattern of the working style can be determined, e.g. the individual works during the weekends; and there are some anomalies such as over-staying at work, or a gap that can be a translated as a holiday. The visualisation in Figure 6.14(b) shows the overall events in a more complex fashion but employs the colour scheme and interaction to support more dimensions and increase the focus on the desired activities.

Furthermore, many previous works such as [115, 219] have employed width as an added feature to a line to show the significance of events or instances. Although

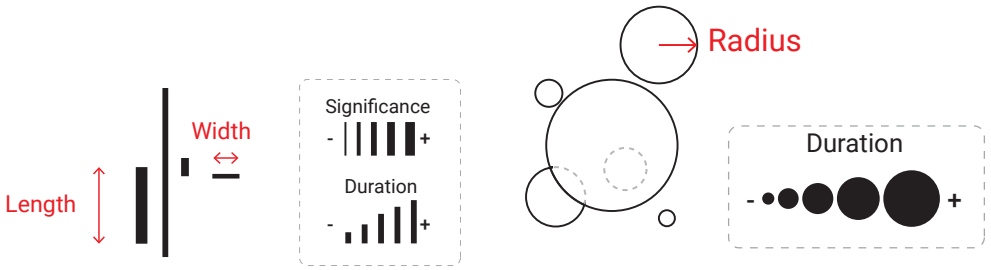


(a) The work related events in the grid structure are used to show the working pattern of an individual within a five-month time range.



(b) All the events in the grid structure are shown within a five-month time range by using a colour scheme to classify the different categories of places and movements.

FIGURE 6.14: Example of representing the events via a line in a 24-hour grid-based layout.



(a) The events are represented by employing lines. The length and width of the lines stand for duration and significance of events, respectively. Note that the short event with a highest significance cannot be recognised among the other events.

(b) The circle can be used to show the events. The radius of the circle represents the duration of the event. A problem arises when the event is lengthy and covers other events.

FIGURE 6.15: Representation of the temporal data via line and circle individually

the use of width in such a representation is reasonable, a very significant event with a short duration may appear visually less significant than a normal event with long duration – see Figure 6.15(a). For instance, the significant event with the duration of 20 minutes can be perceived as less important than the normal event with two-hour duration.

In addition to the lines, some previous works [25, 60] use a circle to represent events, where the radius is used to show the duration and the position of the circle is used to represent the time of occurrence – see Figure 6.15(b). However, it is controversial that using the radius of circle to indicate the length of events leads to overlaps and hence diminishes short events. Moreover, as this type of representation could not envisage the duration and significance of the temporal data at the same time in the previous work, they only show one dimension at a time with the ability to switch to the another dimension.

In order to demonstrate the events and their significance within the 24-hour event visualisation, there is a need for an uncomplicated visual attribute that can be perceptually perceived by humans and causes no overlap. This visual attribute is required to indicate the significance, time of occurrence, and the duration of events all in a most favourable way. To this end, a novel visual attribute was designed by considering the previous works as a base. This visual attributes is composed of a circle and a line in which a circle indicates the significance according to its

radius, and a line shows the duration by its length and scaled value of significance by its width mapped onto the 24-hour axis.

This attribute can effectively envisage the significance factor and the duration of the temporal data – events – without any misinterpretation. This means that the short duration event with a high importance is now encoded correctly. Furthermore, to avoid potential visual clutter, the events, in the first instance, are represented solely as circles with different sizes that point to the their significance score. The size of the circles are scaled ordinally ranging from 5 pixels to 25 pixels based on the minimum and maximum significance score to prevent the visualisation from visual clutter caused by the high-score events. The circles are placed along the grid timeline according to their temporal information – x-axis for the day and y-axis for the time of occurrence. The circle centre point is positioned in the middle of the event line which is obtained by the start and end times of each event $t_{start} + ((t_{end} - t_{start}) \div 2)$ (Figure 6.16).

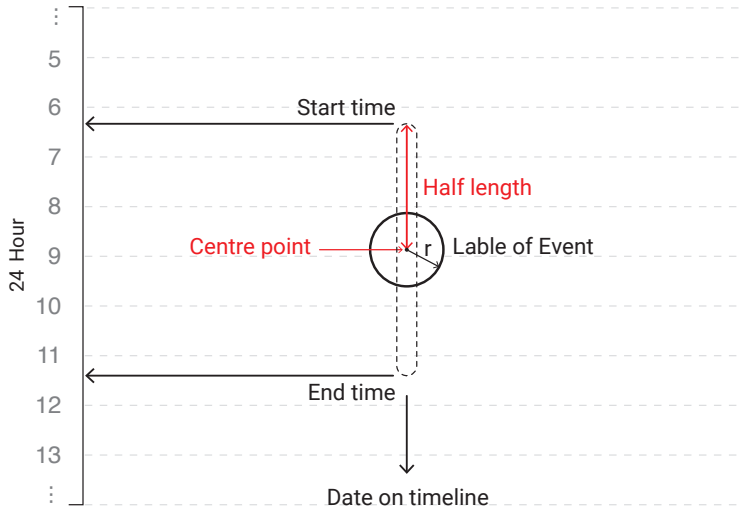


FIGURE 6.16: An event is shown as a circle. The circle's centre point along the y-axis is half of the length (start and end times). The position on the x-axis is according to the date.

The event visualisation is designed with a high level of interaction capacity. The interaction aims at performing additional functions and providing supplementary information. An example of the interaction would be providing the length of the selected event, showing association between the events, selecting or highlighting

the event for more details, and the like. The interaction can be triggered by either mouse click or mouse hover on different parts of the visual properties, such as event circles or timeline. Figure 6.17 illustrates the overview of the interaction. More details on the interaction are described as follows:

1. Show the duration of the event and the other events within the same category by revealing the line. This feature is used when the user intends to get more information regarding the events. The width of each line is scaled to reflect the significance but not exceed a certain value – Figure 6.17(a).

2. Highlight the event according to the established highlighting method earlier in this chapter together with increasing the size of the radius by 20%. This distinctly indicates the selected event – Figure 6.17(b).

3. Fade out the events with dissimilar category by reducing the opacity and colour intensity. This feature intensifies the focus on the current event and its relation within the same category – Figure 6.17(c).

4. Perform additional actions based upon the requirement by providing an event-listener function. For example, a click on the event can open a new panel and show the geographical location.

5. Provide an informative tooltip that consists of additional details regarding the event by considering mouse click or mouse hover action – Figure 6.17(d).

6. Reflect the multi-layered timeline to indicate the date or month of the event occurrence by highlighting the corresponding tiles. This feature would be significantly beneficial by showing the similar events occurrences. For example, in Figure 6.17(e) the occurrence of the food-related event can be seen on the vertical timeline with the highlighted tiles in green.

7. Display an association between the desired event and the other events in the same categories by establishing a link. The association link can be solid or dashed-line with different colour intensity and width in relation to the

location, place name, and distance – see Figure 6.17(e). The encoded link follows the following settings:

(a) *Solid or dashed line*: The line would be solid if the event name is an exact match and the location is the same. This means that the event with the solid-line linkage is the same event or place that took place before or after. The dashed line is drawn if

   i. the event has the same name but the geographical location is different; or

   ii. the event name is not the same but the geographical location is within close range of the selected event.

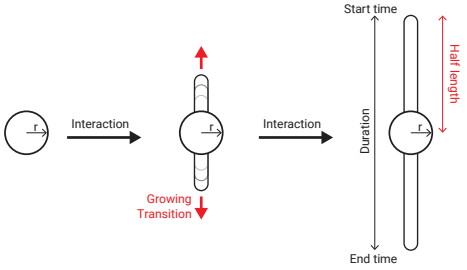
(b) *Line width*: The width represents the distance similarity of the selected event and the other in the same category. For instance, if the place is very close to the selected event (short distance) then the line has greater width than the place with a longer distance.

(c) *Colour intensity*: Colour intensity is added to extend the value of distance. The intensity is reduced by increasing the distance between two events.

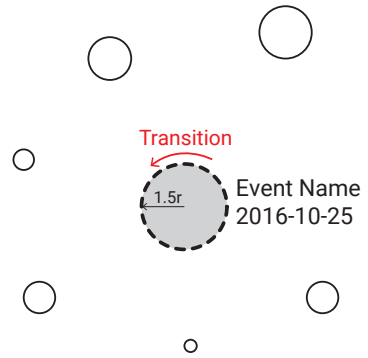
Bubble chart

This visualisation method is designed to show the level of physical activities including walking, running, cycling, and transportation by using different size of circles over the time similar to [25, 60, 190]. In this method once again the timeline is employed to project the time-related information along with the activities. The physical activities consist of various data points such as step count, duration, calories burned, sleep, and duration. To represent such data, an uncomplicated and interactive component has been designed that can be used by non-experts.

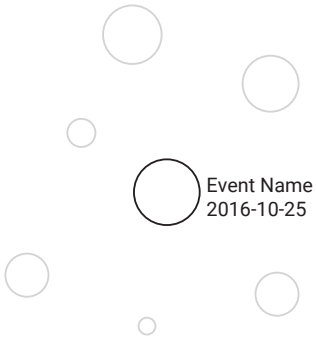
This component is formed of different groups of circles (hereinafter bubble) over a timeline, each of which belongs to one of the four activity classifications. Moreover,



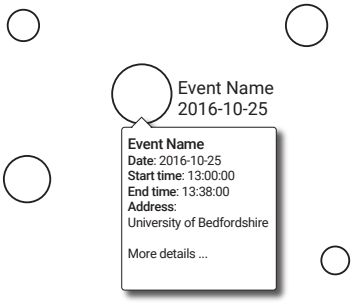
(a) The event is shown as a circle with different radius as a significance indication. Further interaction reveals the length of the event by showing the line. The circle is always located at the middle of the event duration.



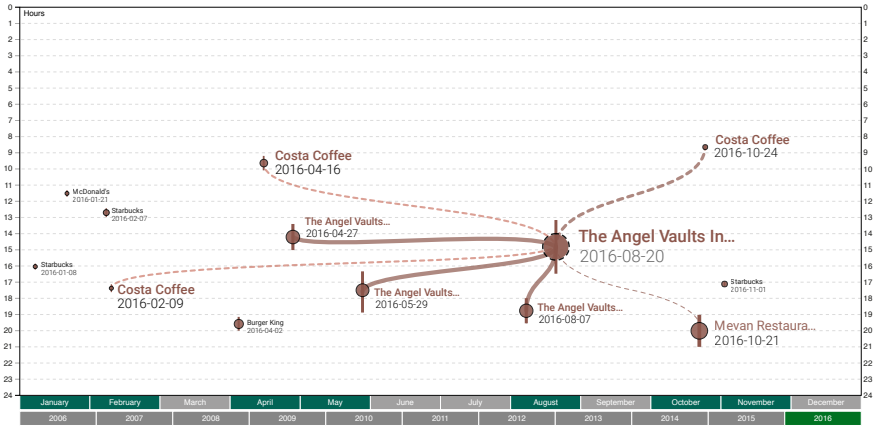
(b) A mouse click triggers the highlighting function. This function increases the size of the circle, reduces the opacity of the unrelated category events, and adds a transitional addition to the event.



(c) The events, which are not in the same category, are faded out by reducing the opacity but can be seen for the comparison purposes.



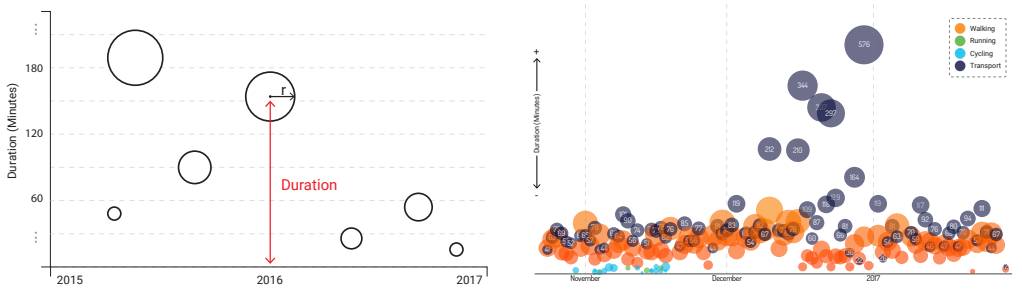
(d) The informative tooltip provides supplementary information regarding the event. The layout of the tooltip varies but needs to follow certain criteria.



(e) The interaction in the 24-hour event visualisation is triggered via mouse hover or mouse click and can reveal the duration of the event, highlight the event, display the association linkage between the event, and show a simple pattern of occurrence over the timeline

FIGURE 6.17: Grid design of 24-hour event representation

four distinct colours and glyphs are assigned to each physical activity. The activity is shown based on its date (position on x-axis timeline), duration (height on y-axis), and step count (radius of circle) – see Figure 6.18. However, as the transportation activity comes with no step count information, a duration is assigned as a default value for its size and also as a notation in the circle.



(a) Each physical activity is located on the timeline according to its timestamp (x-axis) and the duration of the activity (y-axis). The size of the circle stands for the step count.

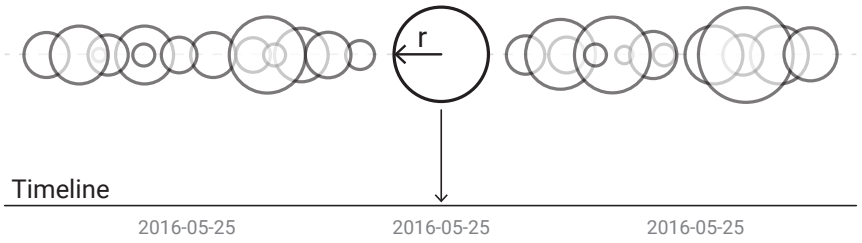
(b) Activities are distinguishable and the individual can perceive the basic pattern of the data. The interaction can reveal more information regarding the activities.

FIGURE 6.18: Bubble chart structure and a visualisation of real daily life data

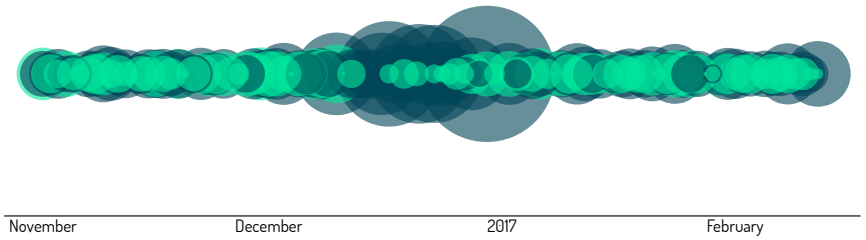
The main goal of this method is to support individuals in understanding their physical activities by looking into the level of activity over time. This can motivate the individual to delve into the details and grasp valuable information regarding the physical activity pattern of their own life – see Figure 6.18(b). Moreover, the designed interaction for this method extends its practicality by providing more details via zoom and pan, informative tooltip, and data filtering.

Besides, the form of the chart can be transformed into a linear layout by laying out all the activities along the single axis and keeping the same setting as before. This form of visualisation can contribute to identify the dominant activities over the time by taking advantage of the colour accumulation. The colour accumulation together with some degree of opacity support the linear layout visualisation of daily activities. The radius of a circle represents the duration of the related activity and the horizontal position shows the date of occurrence over the timeline. Figure 6.19(a) illustrates the simple wire-frame of the visualisation.

However, the key challenge in this form of visualisation is the overlapping of the lengthy activities with bigger radius. To tackle this issue, the data is further



(a) All the activities are represented as circles with different colours and radii according to their type and duration



(b) The result of visualising everyday walking and transport activities over the linear timeline

FIGURE 6.19: Linear bubble chart with colour accumulation technique

analysed during the visualisation process to rearrange the encoding circle based on their duration and regardless of the time of occurrence. This means that the circle with longer duration and hence a greater radius is arranged to appear behind the smaller ones – see Figure 6.19(b).

We evaluated this method to determine its effectiveness based upon on a real-world example. The evaluation result can be found in the evaluation section in this chapter.

Statistical charts

The need to visualise the aggregated information has been felt during the process of the design study. The users were mainly enthusiastic to look into their own data as a whole to determine concise information in a statistical form. This statistical information can be significantly important to increase the general knowledge of the individual. For instance, picturing a distribution of the shopping-related activities

during the week, month, and year or average spending time for shopping can provide valuable information concerning the user's shopping behaviour. This also can persuade the user to explore the data in more detail and hence, amplify the process of knowledge discovery.

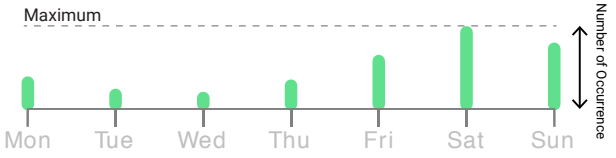
To this end, similar to many visualisation approaches [54, 101, 126, 209, 218], a set of uncomplicated statistical visualisations was designed that can support non-expert users to apprehend the hidden knowledge within the data via a simple and easy to understand layout – see Figure 6.20. Here, three different graphs are used as an exemplar to show the design and the meaning of this type of visual component. In Figure 6.20(a), a number of bars with different lengths are used along the weekdays to show the distribution of the related activities such as shopping or particular place. The length of each line represents the number of overall visits in the particular day.

Figure 6.20(b) uses similar lines but in different directions to show the linear distribution of time and duration of visits along the 24-hour axis. The length of each line stands for the duration of the visits. As a place can be visited a number of times during the 24 hours and over months, a colour accumulation concept is used to show the dense times over the x-axis. In addition, an overview of the average time and duration of visit is provided via different colours on top of all the layers.

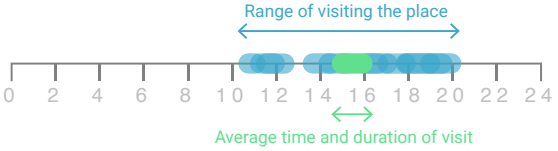
Lastly, Figure 6.20(c) envisages the frequency of visiting a particular place over the year. This chart uses a normal interpolation to link the value of each month to its neighbourhood. For instance, the three peaks is recognisable in this figure. These peaks can be compared to each other to determine the peak of the visits.

Smart Legend

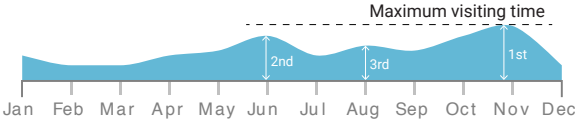
In general, visualisation methods use divergent colours, shapes, and sizes as visual properties to represent the different values and associations of data. Therefore, to assist users to understand the meaning of encoded information, using a legend – a



(a) The distribution of visiting a particular place based on the weekday. The number of occurrences is scaled based on the maximum overall value for each weekday.



(b) The time and duration distribution along 24 hours of the day on the x-axis. The blue and green colours show the overall distribution and average distribution, respectively.



(c) The overall visits to a particular place. The peaks of visits during the year are clearly recognisable.

FIGURE 6.20: Visualisation of the statistical analysis results

gear to provide the meaning of the values and association within the visualisation – is inevitable. There are a number of ways to provide the legend similar to many previous works in this domain such as [60, 119, 162, 211, 231, 232]. Figure 6.21 shows an abstract view of the common static legends.

As discussed earlier in this chapter, the number of colours that can be easily distinguished by a human in visualisation is limited. Consequently, employing a limited number of colours without providing the meanings can lead to misinterpretation. Therefore, a novel dynamic legend has been designed to assist users by providing the meaning of the colour scheme associated with data points interactively. This

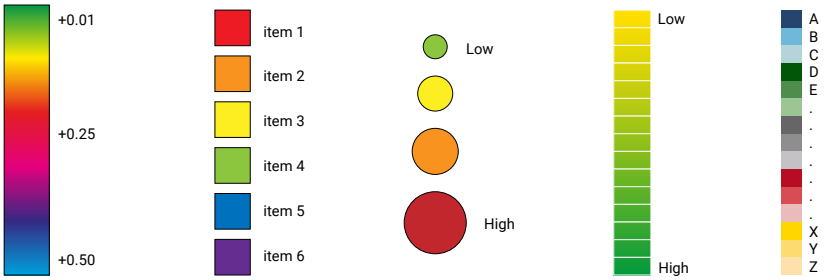
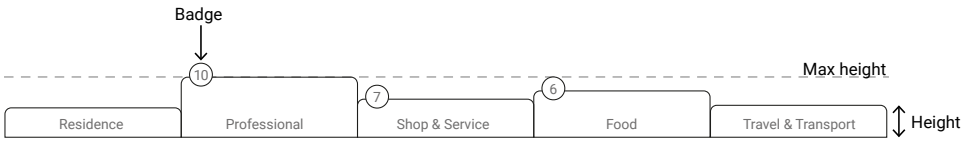


FIGURE 6.21: Different static legends that are used in visualising data

novel legend utilises a set of visual properties such as colour, height, badge, and interaction – see Figure 6.22. The interaction works toward providing supplementary information and extra options for further exploration. The interaction is formulated to embrace a list of associated items, show an indication of the number of items included in the relevant category with a distinct colour, filter out the rest of the colours and categories to increase the focus, and specify the influence factor.

The legend, in this research, is directly used in the 24-hour event visualisation to represent the meaning of each colour assigned to the related data points (Events). The legend dynamically represents the available categories based on their availability within the selected time range. This helps to provide only the available categories rather than representing the fixed number of categories.

(a) The static legend structure that can be used to provide the colour association.



(b) The dynamic legend that provides additional information using badge, height, and interaction.

FIGURE 6.22: Static and dynamic legend differences

This legend is located at the bottom of the visualisation. It helps to keep track of the visualisation whilst trying to interact with the legend. The designed legend provides the following features(Figure 6.23):

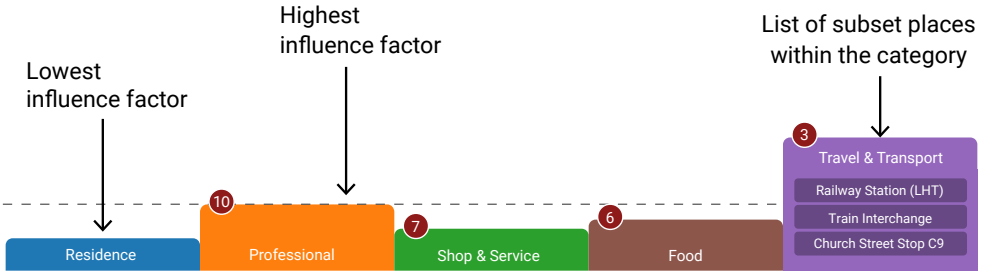


FIGURE 6.23: Dynamic legend with the colour-coded categories, different heights to show the influence of each category, badges, and a list of the included subset

- the categories of the events in different colours,
- the importance of each category using different heights, and
- a list of top-ranked places within each category.

The interaction adds more value to the smart legend by empowering the user to perceive the visualisation in a different class. The interaction in the smart legend offers the following options:

- Highlight the pertinent categories within the visualisation.
- Filter the rest of the unrelated categories.
- Provide abstract information about the top-ranked subset places within the selected category.

Informative tooltip

The informative tooltip is a valuable part of the interaction to provide instant information that is not included within the visualisation in visual analytics. However, design of the highest standard tooltip is a key challenge within the visualisation context. In general, the tooltip comes in different forms depending on the type of the information (e.g. text, image, combination of text and image), but not much in the way of practical standard can be found on designing an

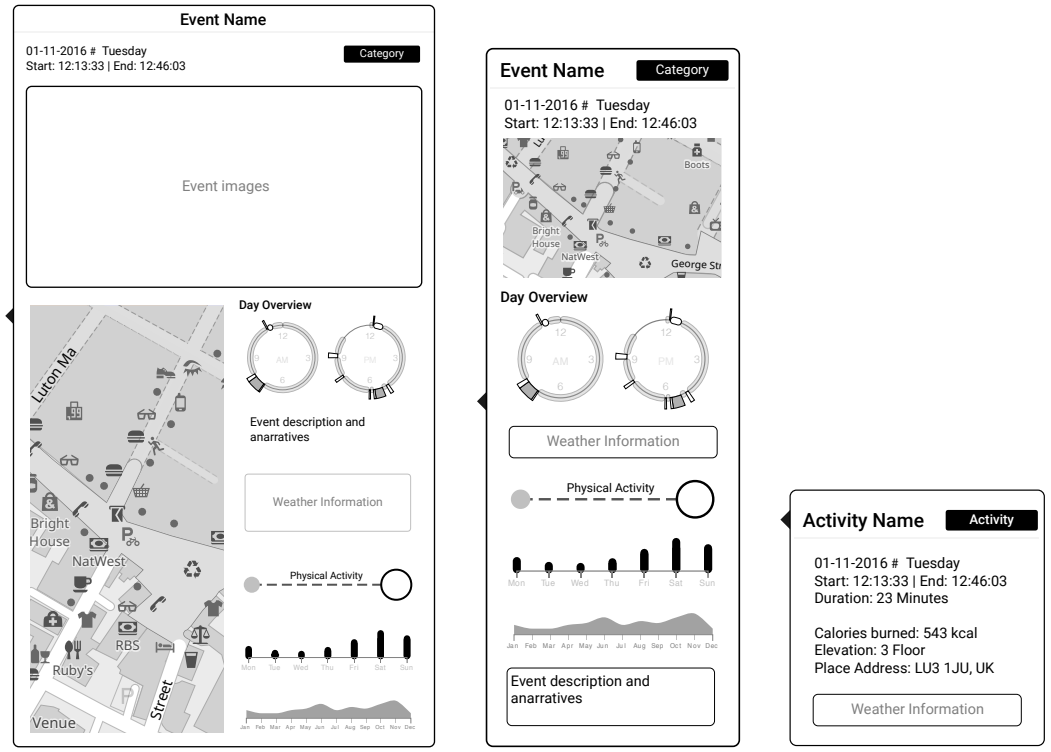
effective tooltip. In this section, the fundamentals of effective tooltip are reviewed to design a compelling and informative timeline.

The tooltip is an individual visual component that can assist in providing supplementary information on demand. A set of characteristics is defined within the design process of the tooltip to equip the visual analytics approach with an effectively informative approach. The following are the predefined requirements for the tooltip:

- *Succinct*: only include the pertinent information that supports the visualisation.
- *Compact size*: compact size that can prevent the visualisation from developing clutter.
- *Locality*: the tooltip is required to be placed within a certain distance from the object.
- *Noticeable*: stands out amongst the other visual properties.
- *Unimpeded*: does not impede the visualisation.

An innovative graphical layout has been designed for the informative tooltip with respect to the aforementioned requirements to accommodate the need for providing additional information within the visualisation of personal daily life. This design is universal and can be adopted or modified by other works towards providing an effective tooltip.

The design includes three different layouts in line with the requirements. The process of designing the layout, in general, includes formulating the frame size, the number of colours, position, and the arrangement for each tooltip. For instance, the frame size is highly dependant on the amount of information that needs to be displayed. If, for instance, there is a need for a memory recall process, the amount of information would be higher to facilitate all aspects of the corresponding purpose – see Figure 6.24.



(a) The large-sized tooltip provides complete details including the day overview, affiliated images, weather information, and related statistics regarding the event to support the user for evoking memory.

(b) The medium-sized tooltip can show sufficient information in the compact form.

(c) The small-sized tooltip can be employed to show the summarised textual information regarding the event or physical activity.

FIGURE 6.24: Tooltip layout design for three different purposes

Figure 6.24(a) illustrates the tooltip frame and its layout. This tooltip is suitable for displaying a large amount of information that the user needs, for instance, to initiate the process of memory recall. Nevertheless, using such a tooltip for the other purposes that aim to facilitate the process of data exploration and knowledge discovery is not recommended as it can impede the visualisation and hence lead to a serious overlap.

The medium (Figure 6.24(b)) or small (Figure 6.24(c)) tooltip are suggested for exploration and knowledge discovery purposes as their layout is compact without any complexity.

All these tooltips are used in different experiments within this work. In the next chapter, the use of each tooltip is shown in action.

6.3.2 User interface (UI)

Personal data has numerous dimensions that can be visualised at the same time closely to each other to enable the user to explore different dimensions together with the relations amongst their data. A well-designed interface for this visual process using a range of techniques, enables users to comprehend the visualisation and perform tasks more effectively.

To this end and according to a large number of approaches with successful user interfaces [10, 53, 71, 202, 204, 209, 214], a unique user interface has been designed that has been adopted in the platform in Chapter 7. The user interface in each platform is slightly different but follows the same principles. In this section, the design principle [51] of the user interface is describe and the details are provided in the next chapter under each exemplar platform.

The design principles that have been used to create an effective and intuitive interface for this research are listed below. It is worth mentioning that these principles are equally employed in the process of designing the visual components in this work.

- **Mental model**: to design the interface according to the concept of daily life and similar tracking, health and fitness applications.
- **Metaphors**: to represent familiar interface concepts to support users to understand the interface of a new knowledge discovery experience.
- **Explicit and implied actions**: to support 1) an explicit action which clearly indicates the result of executing an action for an object, and 2) an implied action which uses visual context or cues to convey the result of an action.
- **Direct manipulation**: to help users feel that the process is controlled by them via displaying the impact of every action instantly.

- **User control**: to expect users to initiate and provide suitable actions within the process of exploration and knowledge discovery.
- **Feedback and communication**: to inform users regarding the initiated process by establishing a continual conversion and providing a sense of control.
- **Consistency**: to prevent users from being forced to learn new ways of performing tasks, or new definitions of colours and categories within different parts of the approach.
- **Aesthetic soundness**: the visual and interaction behaviour need to be consistent with the predefined knowledge discovery purposes by using subtle and unobtrusive elements, standard controls, and anticipated behaviours.

A range of requirements are established to reflect the design principles within the process of designing the user interface for this approach. These requirements are used to form the general user interface. In the next chapter, the user interface is adjusted according to additional requirements of the techniques to boost effectiveness.

UI requirements:

- Use a single view user interface which can be adjusted according to the screen size. This can support users to explore, filter, narrow down, and interact with the visual components in a single window without leaving the page.
- Use a juxtaposed view to show different dimensions of the data closely in a single browser window.
- Use a grid-based layout to accommodate different juxtaposed elements.
- Use neutral colours within the interface that have no impact on the visualisation. This condition assures that the interface embraces the visual component with no visual disturbance.

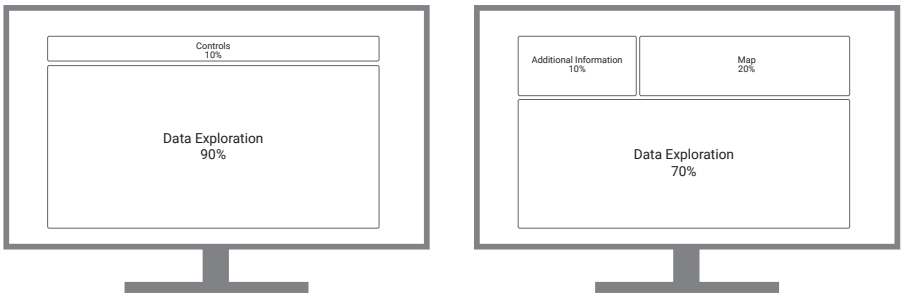
- Use perceptible feedback to inform users about the actions taken and to communicate the status of the data mining and visualisation processes.
- Use simplicity and clarity in the design of the interface. This contributes to its consistency and aesthetic soundness.
- Use well-crafted, inconspicuous graphics, and standard text and controls.
- Provide control over exploration of the data by means of explicit and implicit actions.

The generic user interface layout is formed of a number of sections in a grid-based environment to accommodate the corresponding components. The primary parts are, namely, controls, search, geographical map, and visualisation frame. The layout can be altered to address particular needs of the visualisation.

The interface approach is required to fit all the visual components in a single view with respect to the size of display. Using a single view including the visual components, improves the productiveness of the approach and facilitates the process of knowledge discovery as there is no switching or exchange between any tab or window.

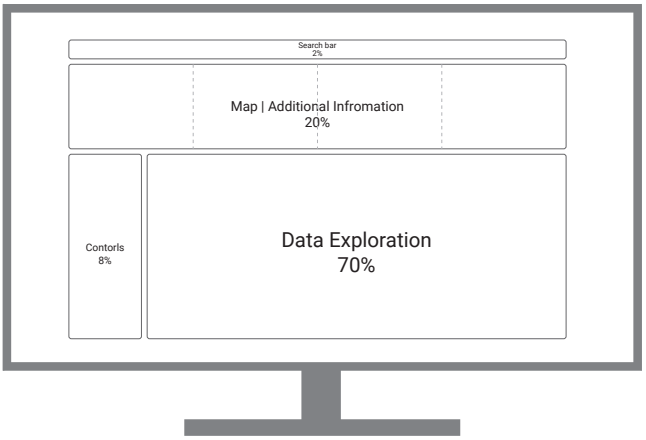
The main section in the proposed interface is the visualisation frame. Giving great weight to the visualisation can amplify the users' performance during the process of exploration and knowledge discovery. To present metadata within the visualisation which holds supplementary information about a certain data point, using a universal method such as an informative tooltip, is common. Although this method is effective, it can lead to overlap or hiding some part of the visualisation if the metadata is large. To prevent the interface from impeded view, a juxtaposed view is mostly used to represent different dimensions of the data. The highlighting feature is employed within the visualisation to disclose related data.

Figure 6.25 shows three different layouts that are used in this work. Each of these interfaces reflects the design principles and requirements, and the layouts will be discussed in the next chapter.



(a) The layout is uncomplicated and accommodates a large visualisation frame.

(b) The interface with three different parts to provide extra information, display the geographical map, and represent data.



(c) The majority of the interface is assigned to the data exploration according to the web application metaphor.

FIGURE 6.25: Different set of layouts with respect to the design principles and requirements

The user interface comprises a number of components, namely, control panel, search bar, geographical map frame, text-based list, modal frames, and data exploration frame. Each of these components is utilised to accommodate the essential needs in the visual analytics approach. The rationale of designing the components together with a brief description are depicted in the following.

Search box

The search box is designed to allow event queries or filtering using a range of keywords for named entities in natural language. The search box fits the need for

a simple contribution from non-expert users who are keen to make queries about events or narrow down their exploration to certain points.

The idea of providing a search box is studied via examining different layouts during the user-centred visualisation design. Employing a search box has been identified as intuitive and can support non-expert users to filter the data and focus on a particular part based on their interests.

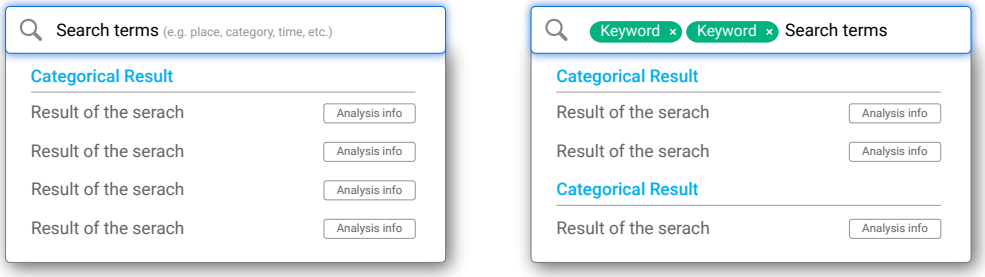
The search box is equipped with a smart assistance feature such as smart search, and multiple entities, and instant hints to enable users to search constructively as using no means of hints or suggestions can lessen the instinctiveness of the exploration process. A new categorical auto-suggestion functionality has been incorporated to assist the search process. The auto-suggestion is adapted from the Twitter Typeahead JavaScript library [5] with a flexible and powerful suggestion underlying structure.

The search box also provides brief information about the suggestion term during the search in a form of hints or clues. This can help users to gain initial knowledge before narrowing down to their particular search term. In addition, the search box allows for entry of multiple named entities incrementally as the search within the daily life context can be formed of a set of keywords. These entities or any other search preferences can be entered in an arbitrary order to facilitate the search experience – see Figure (6.26).

Control panel

The level of control over the data exploration varies. Many approaches are available using different types of controls in this context. For simple exploration, the provided controls are limited to a limited number to avoid confusion and misinterpretation whilst for more convoluted data exploration, more controls are available and progressively so to convey a deeper exploration.

[5] http://twitter.github.io/typeahead.js/



(a) Search bar with auto-suggestion ability provides a close match to the search term along with an instant fact or information.

(b) The incremental search together with a suggestion helps users narrow down the result based on multiple entries.

FIGURE 6.26: Search box appearance and functionality

As this approach targets novice and non-expert users, a generic control panel has been designed with limited controls within the user interface. The control panel is provided as an alternative means to the search box, in which the users can set the parameters according to their preferences. For instance, the control panel can be deployed to select a particular time range or duration within the process of data exploration (Figure 6.27(a)).

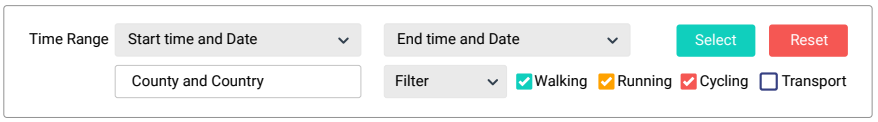
The control panel and its parameters are designed to be directly linked to the data mining model and visualisation (e.g. semantic enrichment and significance ranking) to reflect the settings on the underlying structure. Additionally, the control panel provides an instant visual feedback following any alteration to its parameters to satisfy the design requirements. Figure 6.27 shows three different set of controls which can be employed over different data explorations in the visual analytics approaches.

The effectiveness of using the graphical interface – control panel – to adjust the setting is proven in the design study. The result of the evaluation is depicted later on in this chapter.

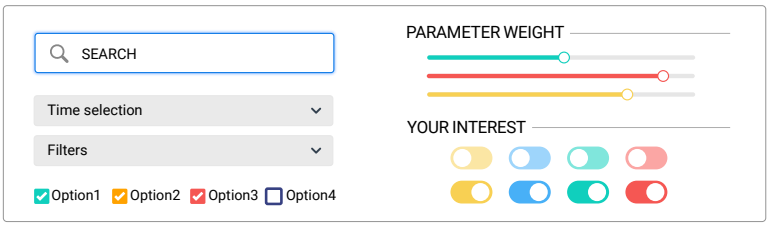
Geographical map frame

Personal daily life comes with spatiotemporal information. This information provides a simple overview of the user's location and can help them to contemplate

(a) A basic control panel that provides a selection of sensor, date, and duration using a drop-down menu in a non-complex way.



(b) A control panel with more control over the process of exploration using a drop-down menu together with a tick box to facilitate the process of filtering.



(c) An advanced control panel with a greater number of controls.

FIGURE 6.27: Three different layouts for the control panel to reflect the process of exploration.

the movements and stops visually on the geographical map and support them in the process of data exploration and knowledge discovery. Thus, an individual frame is assigned to display the spatiotemporal information on the interface. The size of the frame can be varied based on the number of data points. For instance, to display an overview of the data, the map should present a broad width and certain heights whereas to show a single point the size of the map frame can be compact. Different frame sizes are shown in Figure 6.25.

Text-based list

The text-based list takes part in the process of exploration by providing a familiar list to the users. According to the literature review, providing the information in this form is highly beneficial as users can directly view the provided information.

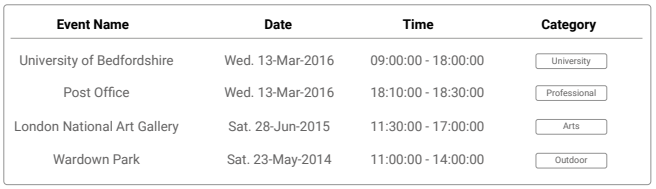


FIGURE 6.28: A text-based list that shows the user activities

The text-based list is one of the options that can be exploited based on the requirements of the platform – see Figure 6.28.

Modal dialogue

The modal dialog is a window positioned on a top-level layer of the interface to overlay the entire content. This form of dialog can increase the focus on certain facts or findings about the data with no interference to the visualisation. The modal dialog is used here to emphasise the results or any additional fact about the personal data. The modal dialog is based on the Bootstrap framework [6] and designed to be highly interactive. The modal dialog is used in one of the exemplar platforms in Chapter 7, section 7.3.

Data exploration frame

The data exploration is designed to be the main part of the interface as it is the foremost goal of this visual analytic approach. It accommodates the visualisation methods and offers an interactive exploration. Thus, the majority of the interface (e.g. 70% of overall display viewport) is appointed to this part to give a great weight to the visualisation and amplify the users' performance during the process of exploration and knowledge discovery. An example of the data exploration can be found in Chapter 7.

[6] https://v4-alpha.getbootstrap.com/components/modal/

6.4 Design Evaluation

The design evaluation was conducted in line with the evaluation methodology in Chapter 3. The design of each visual component in this research was assessed several times and enhanced accordingly based on the goals and requirements. Within this chapter, this allows determination of crucial weak points in line with the designs before establishing any experiments.

The evaluation was conducted by 20 volunteer participants aged between 25 and 56 – 14 males and 6 females – with different, high education backgrounds and normal vision without any colour deficiency. The evaluation was conducted via a standard PC and standard 21-inch desktop monitor. The participants were asked to answer a set of ranking questions ranging from 1 (low) to 5 (high), Likert scales with the scale of 1 (Strongly disagree) to 5 (Strongly agree), and multiple choice questions based on the working prototypes for each visual component.

In general, all the visual components achieved an adequate overall score. The evaluation of the component divulges a number bottlenecks that were identified and addressed accordingly. The following subsections present the results.

6.4.1 Colour scheme and highlighting

Four sets of colour schemes were evaluated to determine the foremost set in terms of distinctiveness intensity and opacity for visualising the daily movements and activities. The result in Figure 6.29 shows that **set A** can be distinguished in different intensity to a greater degree compared to the other sets by 70% of the participants.

Moreover, the participants were asked to rate the three highlighting methods that were introduced earlier in this chapter. The result is shown in Figure 6.30. A blend of opacity, transition, and addition for highlighting a particular item in a visualisation with a large number of visual properties was selected by

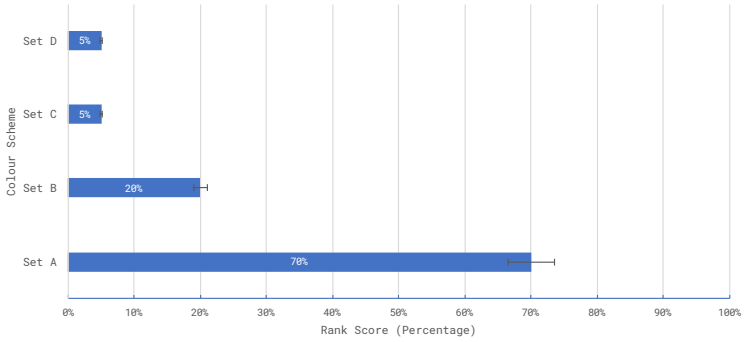


FIGURE 6.29: Ranking result of four sets of colour schemes

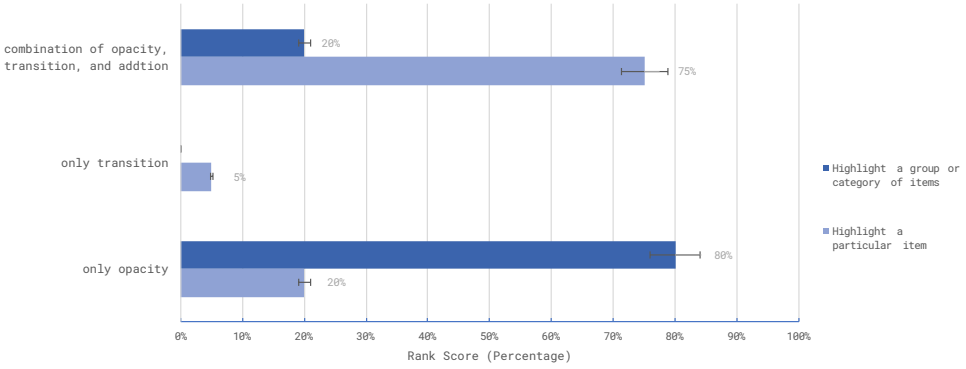


FIGURE 6.30: Result of the highlighting methods for a particular item or a group of items

75% of the participants. Nonetheless, 80% of the participants believed that for highlighting a certain category of activity, employing opacity would be sufficient. The effectiveness of the highlighting method is cross-checked in evaluating the 24-hour event visualisation.

6.4.2 Glyphs

The participants were asked to answer a set of ranking questions regarding learnability, attention balance, searchability, orderability, and channel capacity. The result indicates that the glyphs are highly usable by the participants without much learning and any misinterpretation by reaching an overall 76% (SD=0.08) score. More specifically, the design of the glyphs received 88%, whereas the channel opacity scored 61% on account of using the glyphs within a high number of visual

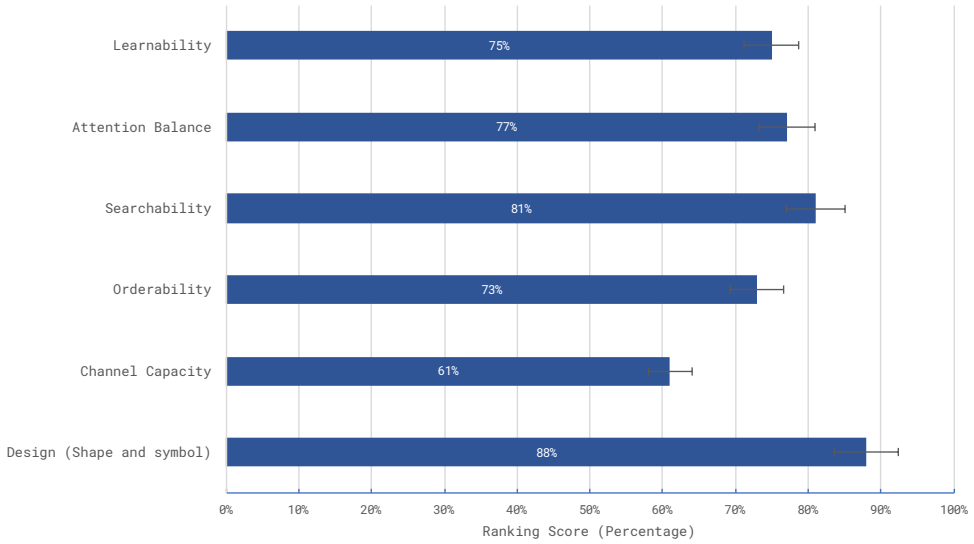


FIGURE 6.31: Glyphs evaluation result

attributes which leads to visual clutter (Figure 6.31). This issue can be addressed by defining a logic to show the glyph only for more important activities and also employing different sizes of the glyphs according to the importance or duration.

6.4.3 Multi-layered timeline

The multi-layered timeline is evaluated in accordance with its design, comprehensibility, interaction, and functionality. The overall ranking score for this component is 76% (SD=0.04), according to the result. More details about each term are illustrated in Figure 6.32. The highest rank is related to the term comprehensibility of the timeline with 83%. The result shows that the timeline is well designed and can sufficiently support the visualisation.

6.4.4 Smart legend

This component was assessed in terms of design, comprehensibility, provided interaction, functionality, and colour scheme. The Smart legend achieved 77% (SD=0.05) of the total score in which the design and comprehensibility were

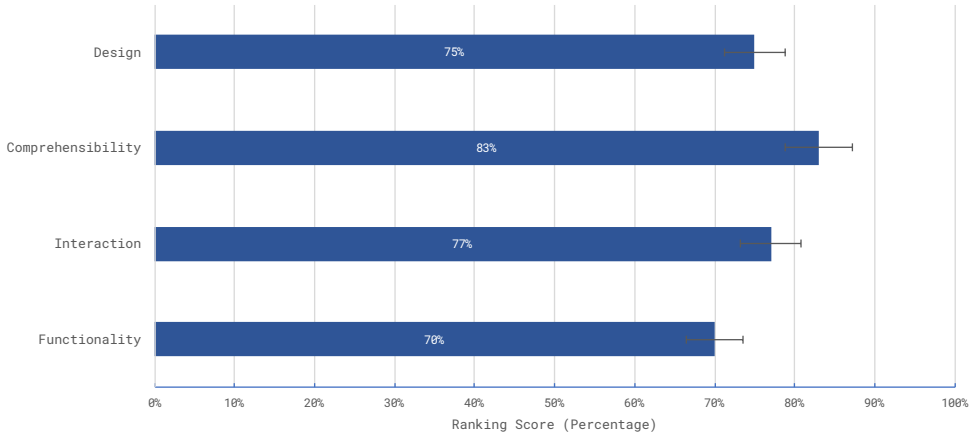


FIGURE 6.32: Timeline design evaluation result

the highest and lowest rank amongst the evaluation, respectively. The design attained 85%, whilst the comprehensibility reached 69% (Figure 6.33). Although the score is not critically low, the smart legend needs an extra description during the training as there no similar legend is used before and the participants were perceptually not familiar with this component.

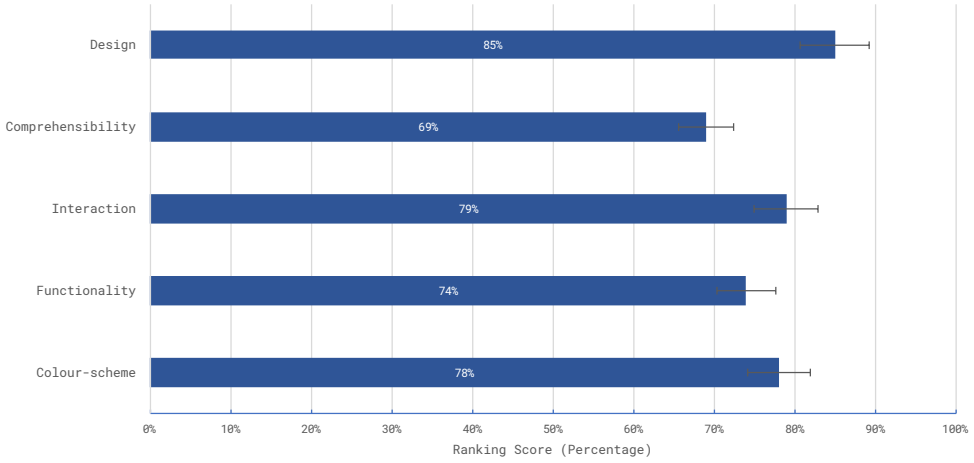


FIGURE 6.33: Smart legend design evaluation result

6.4.5 Tooltip

The design of the tooltip is evaluated with reference to the requirements, namely, succinctness, size, locality, noticeability, and overall design. The result in Figure 6.34 indicates that the design of the tooltip fulfils the requirements by scoring 80% (SD=0.04) in total.

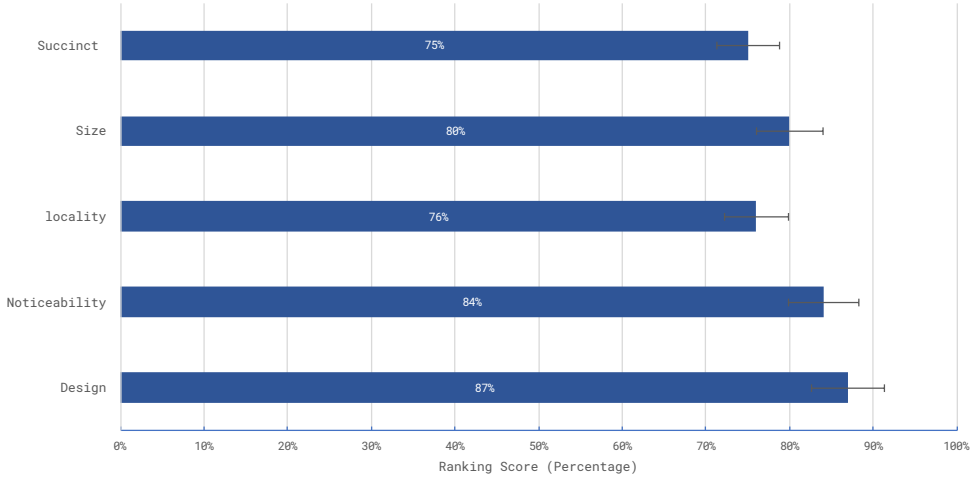


FIGURE 6.34: Tooltip design evaluation result

6.4.6 Circular-based layout

The circular-based layout was evaluated according to the implemented prototype. The participants were first asked to select one of the proposed layouts (A or B) by using a multiple choice question and then answer the ranking questions with regard to their trial session. The result of the analysis indicates that 75% of the participants preferred layout B (AM/PM) over layout A (Figure 6.35), as this layout is designed with consideration of the concept of the traditional clock metaphor of which the users are, to a great extent, familiar. Moreover, the circular layout design attained the overall score of 74% (SD=0.06), whilst the comprehensibility of Layout B outperformed Layout A. The provided interaction received a score of 78%, while the visualisation, the colour scheme used in the

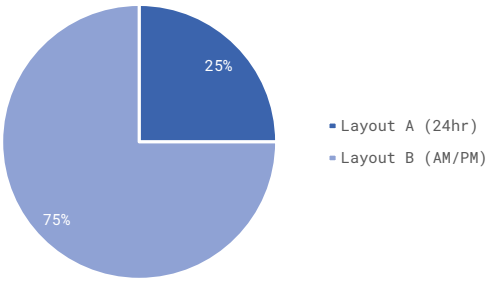


FIGURE 6.35: Participants' choice of layout between 24 hours and AM/PM

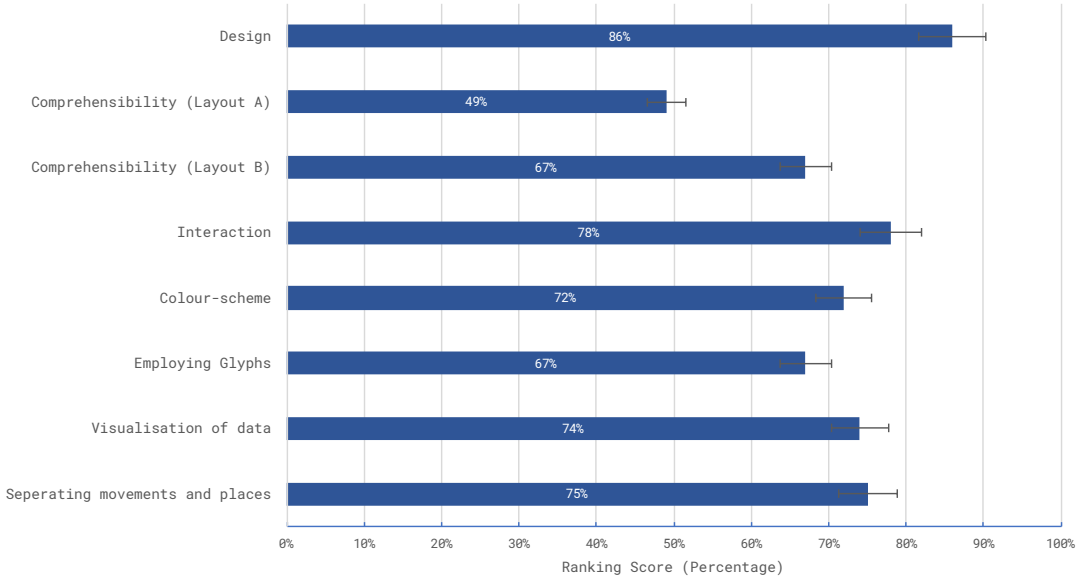


FIGURE 6.36: Evaluation result of the circular-based layout in detail

layout, and separating the movement and place into two individual rings scored 74%, 72%, and 75%, respectively. However, employing glyphs within the circular layout was ranked 67% due to the channel capacity that is revealed in the evaluation of the glyphs. The detail of the evaluation is illustrated in Figure 6.36.

6.4.7 24-hour event visualisation

This module was evaluated to determine the effectiveness of the design. The participants were asked to answer 12 Likert-scale questions with 1(Strongly disagree)

to 5 (Strongly agree) together with 12 ranking questions ranging from 1 (low) to 5 (high) regarding the event visualisation. The questions can be found in Table A.3. The result of the evaluation shows that the 24-hour event visualisation achieved an overall score of 80% (SD=0.05). Figure 6.37 illustrates the result in detail. The highlighting and visualising of the events were cross-checked in this evaluation and as expected attained 84% and 90%, respectively.

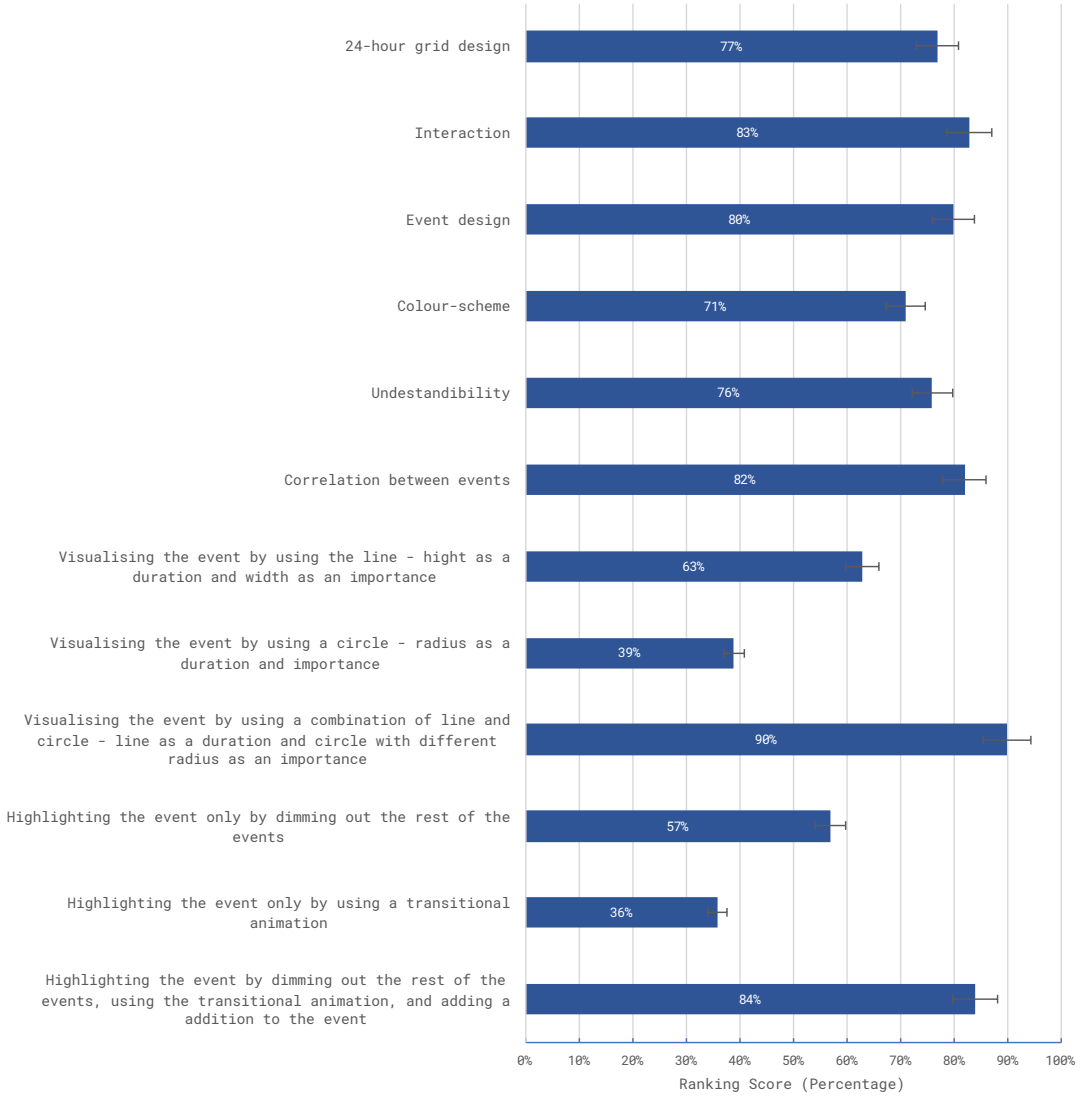


FIGURE 6.37: Ranking result of the 24-hour event visualisation in detail

The analysis of the Likert questionnaire (Figure 6.38) shows that the method provides a compelling interactive visualisation with a 24-hour grid style. Encoding

the event via a combination of circle and line can, from the results, effectively encode the importance along with the duration of the events at the same time.

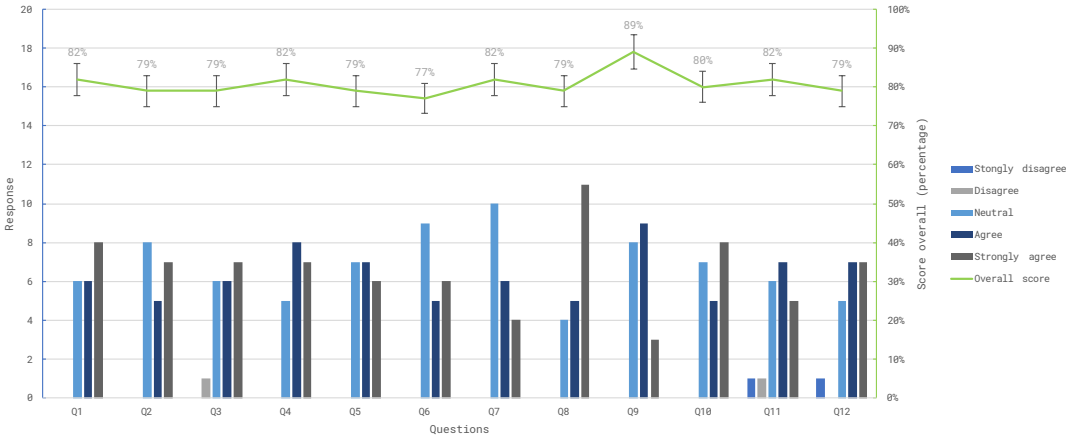


FIGURE 6.38: Result of the 24-hour event visualisation Likert questionnaire in detail

6.4.8 Bubble chart

The bubble chart evaluation shows that this type of visual component can effectively show activities but not adequately when the number of activities grows dramatically. The bubble chart in total achieved 70% (SD=0.06) of the overall score. The visualisation and design were scored 67% and 69%, respectively, owing to the cluttered view caused by the great number of activities (Figure 6.39). However, by using interaction which gives a filtration option, this method is highly ranked as a comprehensible method that can show an individual's activity pattern effectively.

Despite the fact that the visualisation can produce a cluttered overview, the result of the Likert questionnaire shows that this method can reveal the activity pattern effectively via its two different modes (Figure 6.40).

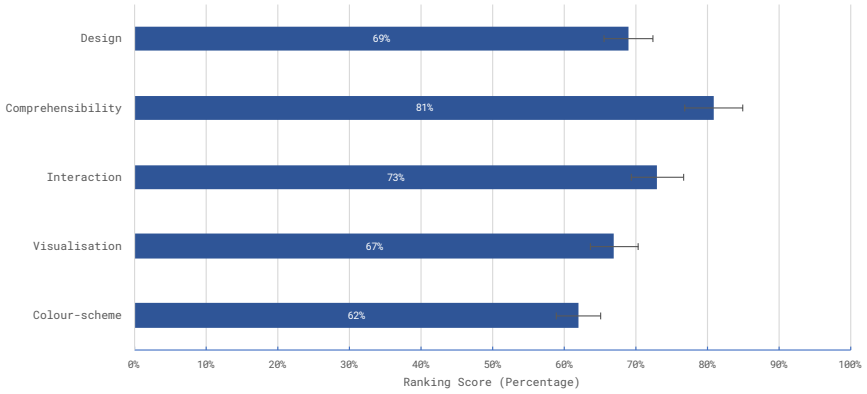


FIGURE 6.39: Result for the bubble chart design

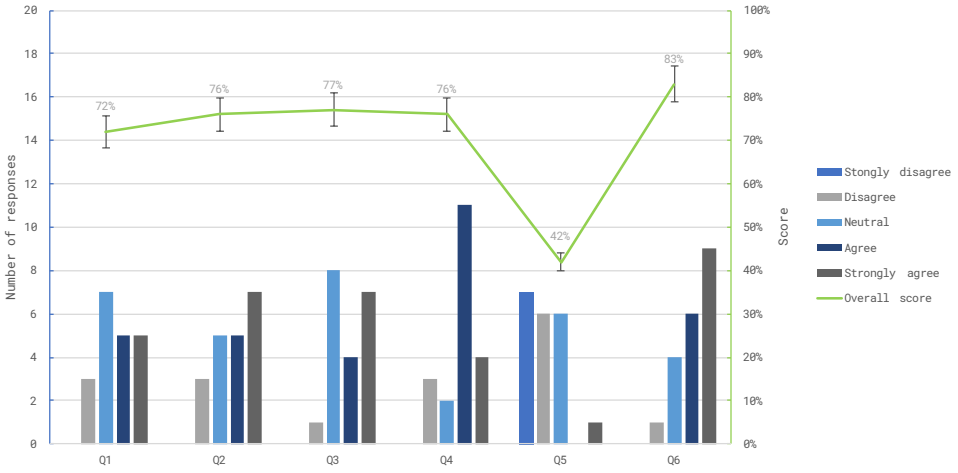


FIGURE 6.40: Bubble chart assessment result

6.4.9 Storyline

The storyline was evaluated against the design, comprehensibility, information visualisation, and interaction. This component achieved adequate ranking scores in all aspects (Figure 6.41). The overall ranking score for this method was 81% (SD=0.07). The interaction has the highest score 92%, followed by the comprehensibility 80%, while the visualisation and the design were scored 79% and 71%, respectively.

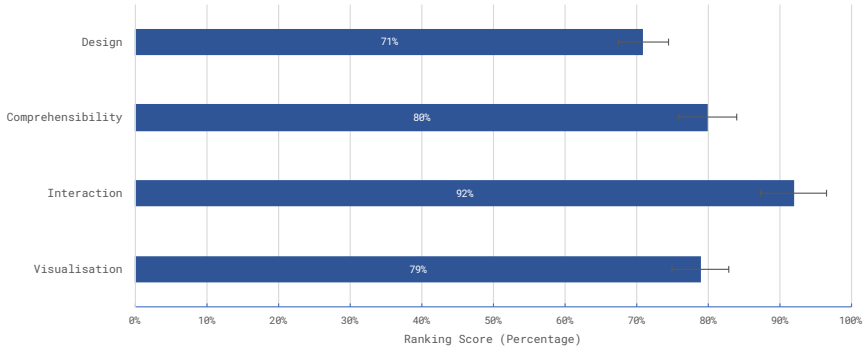


FIGURE 6.41: Storyline evaluation result

6.5 Chapter Summary

This chapter introduced several novel visual components, interactive visualisation methods, and user interfaces – as one of the main contributions of this research (**C5**) – that can be used to envisage the extracted knowledge from personal daily life data. These newly built visual components and interactive visualisation methods (e.g. smart legend, storyline, circular-based visualisation, 24-hour event visualisation, bubble chart, etc.) were designed and implemented based on the preattentive processing and the current base line in the information visualisation domain. Additionally, a number of features were established for an effective user interface to accommodate as well as link the visual components together in an optimum and user-friendly way. All of these can be combined and set out in a new pipeline to picture different dimensions of individual data.

Correspondingly, a set of evaluations was conducted on the design of the novel visual components to assess and to identify weaknesses by analysing the outcome. As a result, this allowed improvement to the designs and the functionality of each component, accordingly.

Now, by establishing the data mining models (in Chapter 4 and 5) and the visual components in this Chapter, a comprehensive pipeline can be illustrated for the visual analytics approach at this stage. This pipeline has the ability to connect to different components via interaction and user involvement to extract and envisage

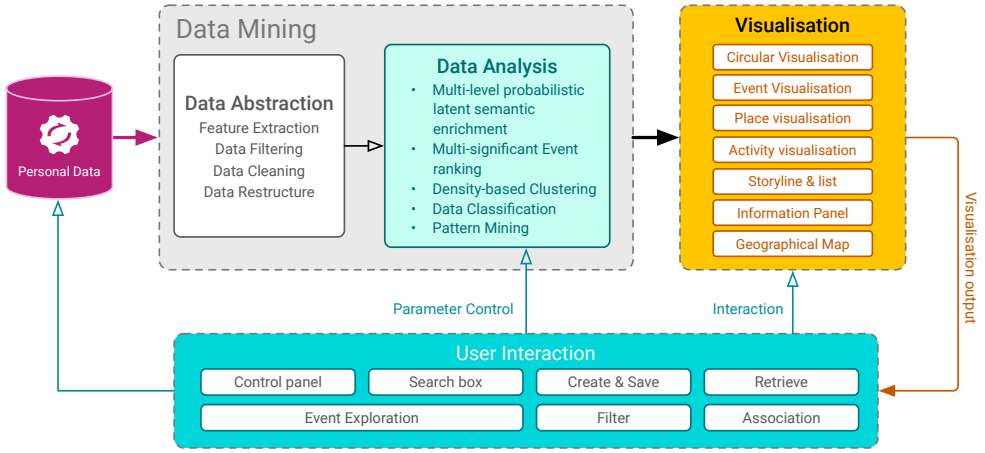


FIGURE 6.42: Main visual analytics pipeline in this research

valuable knowledge rationally to individuals (Figure 6.42). This pipeline including the data mining and visualisation components is deployed to create a unique approach that can cover different aspects of personal life in action. In the next Chapter the use of this pipeline for different purposes is demonstrated in depth.

CHAPTER 7

Integrated Visual Analytics Tools – Platforms

The novel data mining and visualisation established within this research can be used for different purposes or in different domains. The proposed visual analytics approach in this work comprises three main components, namely, data mining, information visualisation, and interaction in order to extract features by analysing the data and finally, visualising the outcome interactively. The data analysis component and interactive visualisation are extensively described in chapters 4, 5 and 6, respectively. In this chapter, the contributions are highlighted and the approach are demonstrated with different criteria that can facilitate the process of knowledge discovery, uncover the life pattern, and support reminiscence, respectively, within the personal life data domain. Three integrated tools are designed and implemented over the course of this research – namely ActivityTimeline, LifeTracker, and MyEvents – to reflect the overall research aim with particular attention to different dimensions of personal daily life. Subsequently, the accuracy, usability, and effectiveness of each tools is evaluated individually.

The general purpose of each platform is as follows, to:

- Support the personal daily life pattern based on the level of activity and movements over time by means of an intuitive temporal visualisation. (ActivityTimeline).
- Support users to explore lifestyle data at different levels of detail and different timescales interactively by allowing them to identify highlighted key places (LifeTracker).
- Support the active involvement of individuals in the process of reminiscence by providing an interactive query and mining the significant events (MyEvents).

7.1 ActivityTimeline

Owing to the widely available sensors in mobile and wearable technologies, different aspects of individuals' life including fitness, movements, health, and lifestyle can be automatically captured. These data, which are highly regarded as personal data, can be utilised to provide a comprehensible understanding and empower an individual's life. However, such data cannot be interpreted by individuals without practical visual representation and user interaction. There are many tools – with common grounds – that typically provide only visualisation of such data with basic encoding and no means of interactive exploration. To fill this gap, a constructive visual analytic platform is designed and developed to support the interactive presentation of large-scale temporal data.

7.1.1 Introduction

This technique is part of MyHealthAvatar, a European Commission funded project aiming at visualising lifestyle data collected by tracking devices or applications

to improve individuals' lifestyle. The amount of collected data is immense as it tends to be collected over a lifetime. Therefore, grasping knowledge from a large collection of continuous data without practical interactive visual representation is beyond the bounds of possibility. To this end, a familiar means of interactive visual analytics, a suite called ActivityTimeline, is introduced to address and facilitate the presentation of such data.

The data used in this platform encompasses three different temporal sources: *Fitbit*, *Withings*, and *Moves* collected by the standalone tracking devices and the smartphone tracking application. The structure of the data is described in Chapter 3. The temporal data from *Moves* contains the automatically detected types of activities followed by step counts and spatiotemporal data, whilst the data from *Fitbit* and *Withings* includes only the step count, calories, and active minutes with no type of recorded activity or GPS information.

In this platform, the data validation together with feature reduction discussed in Chapter 5 are employed to remove any incomplete or noisy parts and filter out unnecessary features, respectively. This leads to a higher performance for encoding the large scale of data visually.

7.1.2 Interactive visualisation

The main focus of this platform is to represent the daily activities of the users. These activities include any movements and places that the users tend to do or stay at in their everyday life. The movements are automatically classified into four standard types, namely, walking, running, cycling, and transport by the tracking applications used in this research. To picture such data including movements and places, the interactive visual components introduced in Chapter 6 can be utilised. Hence, a number of objectives that can be reflected by using the components are set out as follows:

- **Activity Stack**: is aimed at interactively portraying the overall daily activities including walking, running, cycling, and transportation as a stack bar over the timeline.
- **24-hour activity**: to illustrate the user's activities including the movements and places within the 24-hour timeline and provide initial insight into the user's life pattern.
- **Activity Cloud**: to accumulate the activities in different layers over the timeline and show the dominant type of the activity, such as walking or transportation, over time.
- **Activity Bubbles** visualises the daily activities based on their types and durations over the linear timeline and demonstrates the interest in doing each activities (e.g. walking, running)

These components are arranged in an uncomplicated user interface discussed in Chapter 6. The interface is divided into two separate parts, control panel and visualisation canvas (Figure 7.1). The control panel holds a set of useful options such as source selection, time range, and duration, while the visualisation canvas is solely dedicated to picturing the information.

Activity Stack

This component interactively stacks the duration of all user activities throughout the day over the timeline. This allows the user to perceive 1) the overall activities briefly regardless of the occurrence time, and 2) the activity trends over time by providing an interactive and uncomplicated visualisation. Figure 7.2 shows a user's activities between 1st March, 2015 and 7th Feb, 2017. The visualisation demonstrates a change in the user's activities after January 2016. According to the colour-coded activities, the user began using transportation dramatically more than in year 2015. This can be interpreted as a lifestyle change. Another explanation can be that the user reduced the amount of walking during the day, which could lead to a new health condition.

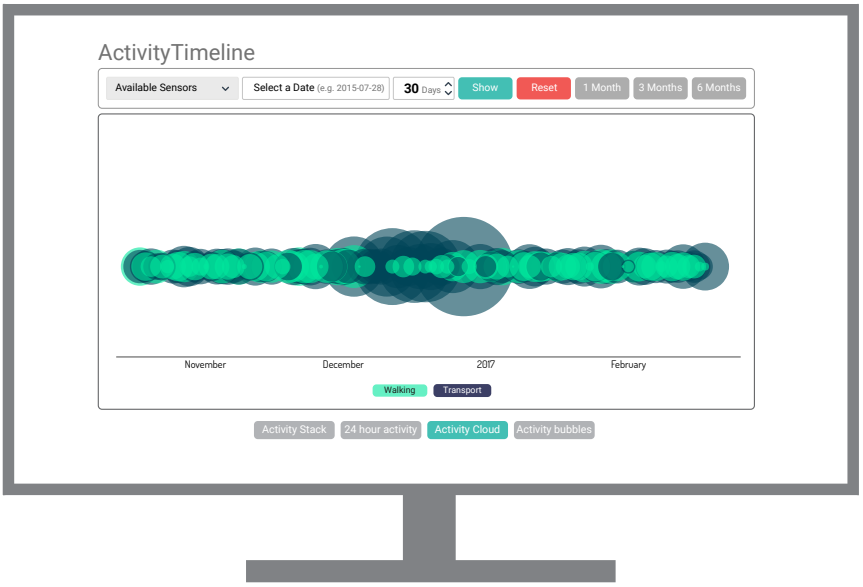


FIGURE 7.1: ActivityTimeline user interface

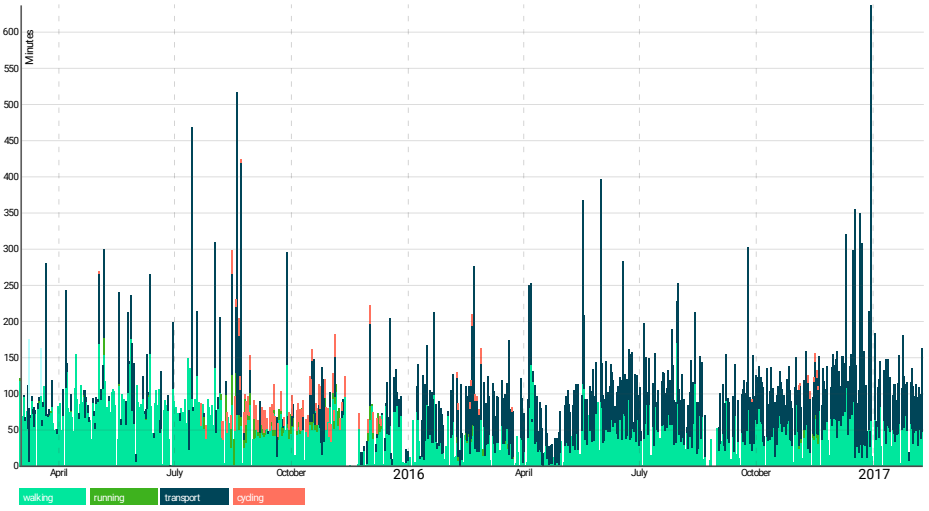


FIGURE 7.2: Activity Stack shows a change in the user lifestyle from 2016 onward. Moreover, according to this visual encoding, there is a short period of cycling in the user data.

24-Hour Activity

This method lines up the daily activities including the movements and places over 24 hours and over the timeline (Figure 7.3). This component provides an interactive colour-coded visual representation to facilitate the process of understanding the daily life and allow for data exploration. This means that the user can benefit

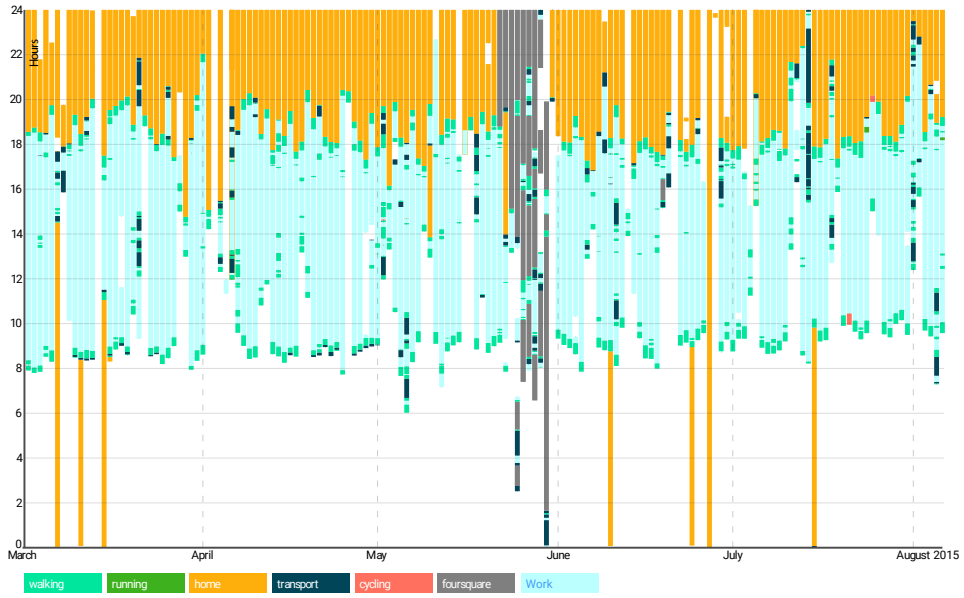


FIGURE 7.3: Daily activities including the movements and places on the 24-hour grid-based timeline.

from the overview and then focus on a particular part of their interest for more information.

Activity Cloud

This component represents the activity with respect to the idea of colour accumulation. The colour accumulation enables the visualisation to indicate the most dominant activity colour. For instance, if there are four activities taking place during the day and three of them are walking, the dominant colour would be the colour of the walking activity. Figure 7.4 is a visual representation of a real activity data from 2015 to 2017. It evidently indicates a sever conversion between walking and transport activities by the user in 2016. This can be interpreted as a lifestyle change.

Activity Cloud does not disregard the other activities by creating a series of plain circles with only the dominant colour. Instead, it displays the activities as they occurred at the time by using the associated colour code and radius.

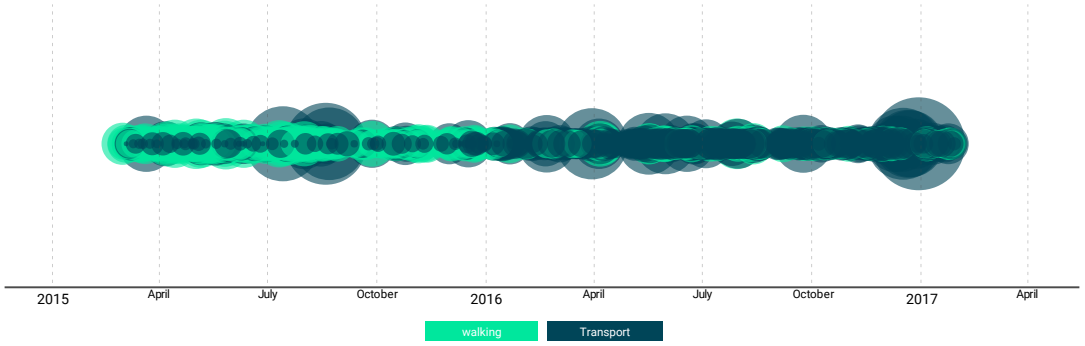
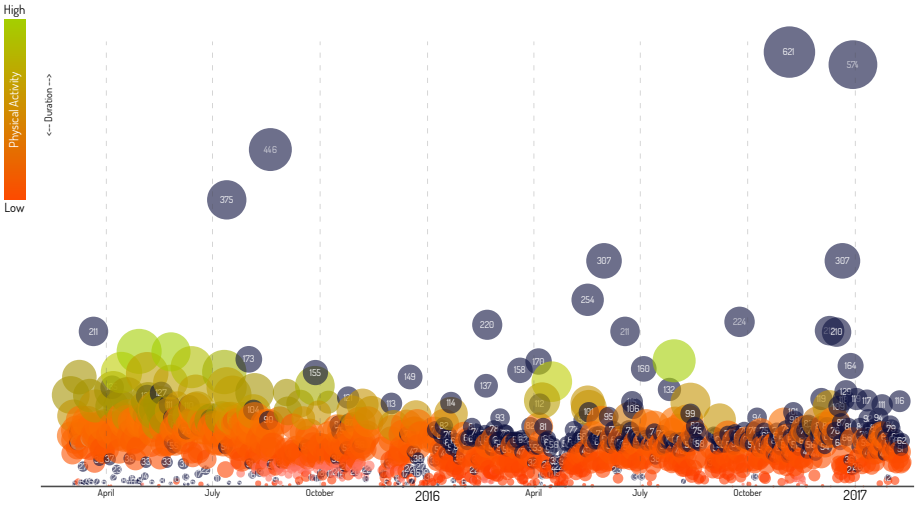


FIGURE 7.4: Activity Cloud indicates a sudden change within the daily life in 2016. The dominant activity change is from walking in 2015 to transport in 2016. It also shows the duration of each activity by the size of the circle. The bigger the circle, the longer the activity.

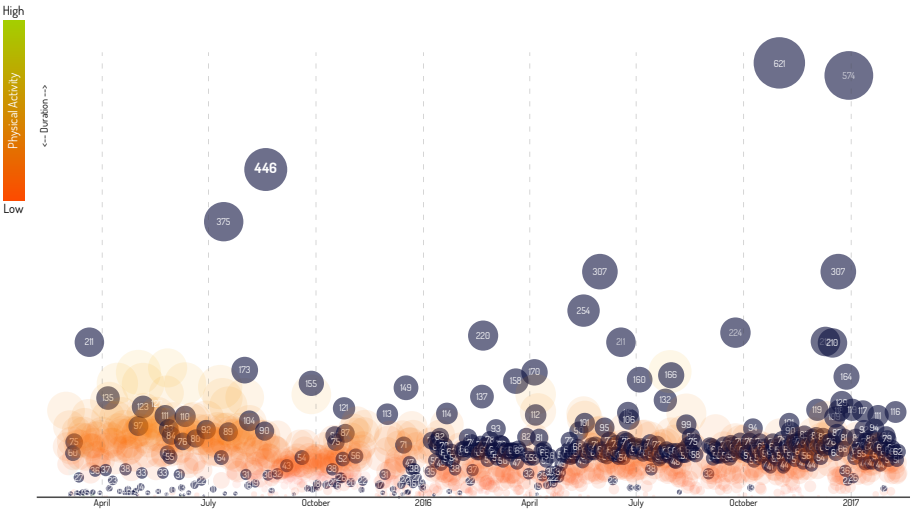
Moreover, this component can be used to identify the long and important activities amongst the years by drawing the user's attention to the bigger circles with the larger radius. The radius of the circle represents the duration of the related activity. Hence, the longer the activity, the larger the radius. From this representation, the user is able to perceive the longer activities. For instance, Figure 7.4 shows a number of circles with considerably large sizes for transportation activity. This can be interpreted as a long-distance travelling destination on or during the particular period of time.

Activity Bubbles

This component is designed to indicate the importance of daily physical activity by dividing the activities into the physical activities (e.g. walking, running, cycling) and the transportation activity. The activities are visualised in a set of circles of different radius and colour. The colour of transport-related activities is dark regardless of transportation kind to emphasise the difference (Figure 7.5). Physical activities are intuitively denoted via a red–green gradient. This gradient is set to show intense physical activity as green and low physical activity as red based on its duration. The setting for classifying the intensity is varied based on the health model.



(a) Overall activities



(b) Interactive filtering

FIGURE 7.5: The position of each circle represents the date of that activity on the x-axis and the duration of the activity on the y-axis. The personal daily life of the users shows that although the level of user's transport activity has increased dramatically, this person had a considerable level of physical activity during 2015 and 2016.

Here, the circles are lined up based on the time of occurrence (x-axis) and the duration (y-axis) in an ascending fashion. Moreover, the interaction on this visualisation addresses the cluttered representation of the activities and also provides more information. The visualisation offers hovering, zooming, and panning. By hovering the pointer on each circle, firstly the same kind as that

activity (e.g. walking) is highlighted and the rest are dimmed out, and secondly additional information regarding the activity such as a precise time, calories burnt, duration, etc. is displayed on a tooltip. The zooming and panning enable the user to drill down to a particular time over the timeline. This results in a less cluttered viewport and improves the focus.

7.1.3 Evaluation

This approach as an uncomplicated platform is assessed to determine the effectiveness of exploring the daily activities and gaining meaningful understanding by asking the participants to complete general tasks followed by the usability questionnaires.

In total, 20 volunteers were recruited – 17 males and 3 females, aged 19–55, in the University, with normal vision without any colour deficiency. The evaluation was conducted via standard university PCs (Intel Core i3 CPU, 8GB Ram, and onboard graphic memory) with standard 21-inch desktop monitors, mice, and keyboards. The platform was run as a web-based page via Chrome browser.

The data for the evaluation was generated from two volunteers who had, combined, actively used a *Fitbit*, a *Withings* and the *Moves* application for approximately 14 months. The personal data were anonymised and concatenated to create a synthetic person.

7.1.3.1 Methodology and procedure

This study began with a short introduction of the platforms to the participants. The participants were given a 15-minute free practice time to familiarise themselves with the platform. The evaluation was designed in two parts: task completion and usability evaluation. The four open-ended tasks in Table 7.1 cover the goal of the platform in each part.

Task	Question
T1	Find out the most active day of the user during 2015 and provide the duration and the type/s of the activities by using Activity Stack.
T2	Provide an approximate number of working hours (start and end times) that the user did during 2015 by using the 24-hour Activity
T3	Identify the period that the user dramatically changed their activity style within all the data by using Activity Cloud.
T4	Provide the approximate period that the user reduced the level of physical activity.

TABLE 7.1: ActivityTimeline evaluation tasks

7.1.4 Result Analysis and Discussion

The evaluation process was analysed and the results divided into two parts, accuracy and usability. Accuracy is determined based on the correct answer to the tasks. Subsequently, the Likert-scale questionnaires were analysed on the scale of 1 (Strongly disagree) to 5 (Strongly agree).

Task 1 and task 3 were answered more than the other two tasks. Task 1 received 85% (SD= 0.06) correct answers owing to the familiar encoding of the information via the stack bar. Additionally, the result of task 3 was 80% (SD= 0.04) correct answers, which shows that Activity Cloud provides an effective encoding of the dominant activities by means of colour accumulation. Task 2 relating to the 24-hour activity reached 70% (SD= 0.04) correct answers. The evaluation revealed that encoding the movements including the physical activities and the places with the same setting results in some degree of confusion. In contrast, task 4 was only answered 65% (SD= 0.8) correctly due to its compact encoding and the colours used to indicate the level of activity as the user could not distinguished the colour-coded physical activities mixed with the level of activity colour.

The usability survey encompasses functionality, efficiency, usability, and reliability, ranging from 1 (Strongly disagree) to 5 (Strongly agree). The scores were transferred into percentages for presentation purposes. Figure 7.7 shows the overall results.

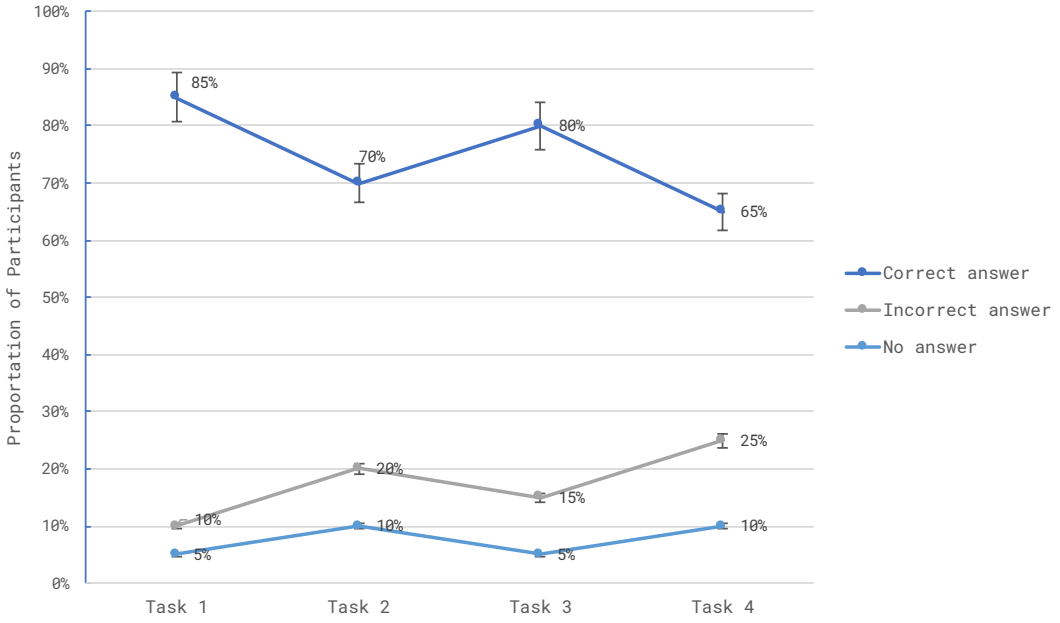


FIGURE 7.6: ActivityTimeline task completion result

The functionality, efficiency, and usability achieved 71% (SD= 0.01), 71% (SD= 0.08), and 70% (SD= 0.05) of overall score, respectively. The reliability received only 63% (SD= 0.07) of the score owing to the fact that the evaluation did not encounter any error and the participants mostly answered the associated question with "Neutral".

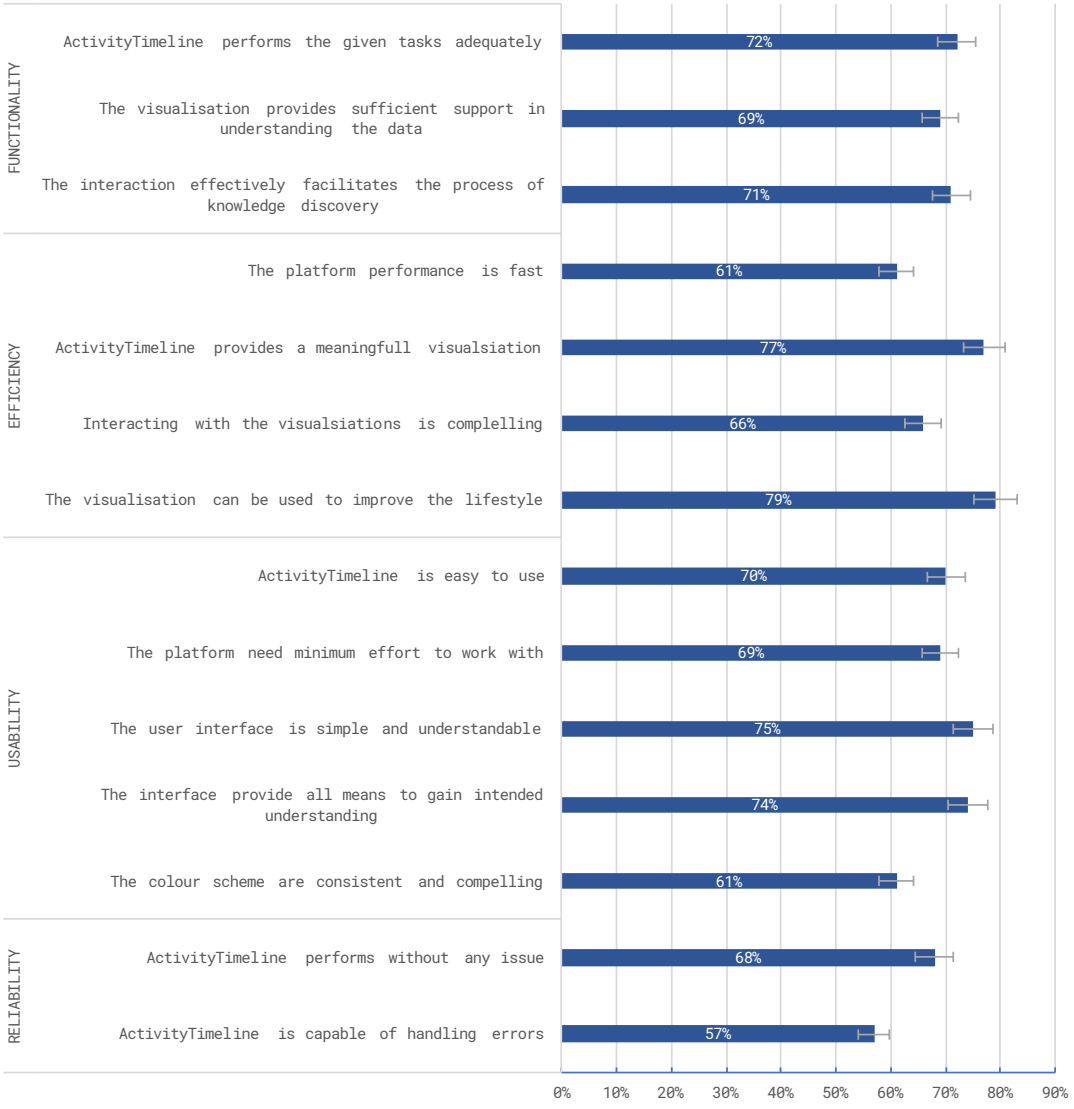


FIGURE 7.7: ActivityTimeline usability analysis result

7.2 Visualisation for Life Pattern (LifeTracker)

There are many potential benefits to be had in areas such as healthcare and personal life management in the recording of a wide variety of personal data related to an individual's day-to-day activities. Data analytics and visualisation can make this data readily accessible to the user, allowing them to view and understand their life patterns. This section depicts on of the use cases, LifeTracker, a suite of techniques that presents lifestyle data and allows users to investigate it interactively. It employs a range of visual analytics techniques to make the outcomes of data summarisation and ranking available to the users, hence allowing them to identify the highlighted key events from the data. This supports the users to explore the data at different levels of detail and at different time scales. LifeTracker follows the principle of "*overview first, zoom and filter, then details on demand*" and offers an integrated environment to present information about individuals' daily activities in one place. LifeTracker has been evaluated by 15 participants who provided positive and critical feedback.

7.2.1 Introduction

LifeTracker addresses the scalability issue by focusing on a particular perspective, which is the data visualisation for presenting information to the users. It follows a renowned "visual information seeking mantra" to offer guidance to visualisation practitioners by describing how data should be presented on screen for the greatest effect. This approach provides a general context – overview – for understanding the dataset by painting a picture of the whole data entity. "Zooming and filtering" involve reducing the complexity of the data representation by removing extraneous information from viewing; and "details-on-demand" provides additional information needed by the users allowing them to focus on data of particular interest, despite any limitations of screen size.

This approach attempts to offer an integrated environment to analyse and present information about an individual daily activities by using a highly interactive medium. This enables the users to explore their own daily life data such as activities, places, and other relevant data by selecting an arbitrary day, month or year.

The interaction within LifeTracker, as it has been discussed in Chapter 6, is designed to offer the functionality of "overview", "zoom and filtering" and "details-on-demand". By overview, the users can gain a perspective about their overall activities and places over a selected period of year or month. By zoom and filtering, the user can select the desired timescales ranging from years to days. By details on demand, the users are able to explore more details in various ways, for example, they can use tooltip to get brief information of a selected data segment; they can use the multi-layered timeline to view occurred historical places down to the scale of days; they can also obtain detailed information about their movements on a selected day through the map view coupled with time information.

LifeTracker features a novel integration of a range of analytics and visualisation techniques to support effective exploration of individual lifestyle and patterns from a sequence of self-logging data over years, including:

- A multi-layer timeline enabling users to easily define a time period of day, month or year by clicking on the timeline to view activities and events that occurred in the selected time period.
- Techniques for the extraction, summarising daily events to highlight important information for a given time period, such as key places and life patterns of the user.
- Interactive visualisation allowing for full exploration of information by the users through a juxtaposed style including a number of integrated views e.g. the map, calendar, life patterns, chart, statistics, etc.

The remainder of this section is organised as follows. A summary of the related work is presented in the next section. Next, an overview of the system is depicted.

The brief data analytics techniques involved in this platform are presented before focusing on the information visualisation techniques used to present the knowledge to the users. The evaluation is described in the last section and, finally, the outcomes are presented together with discussion.

7.2.2 Summary of related work

Spatiotemporal data are used to display time and location information with maps to indicate patterns and changes of individuals behaviour in space and time. Existing techniques mainly use integrated multiple views, colour, and interaction to link the spatial and temporal data [20, 35, 165]. However, such method often produces visual clutter and obstructed data points. Therefore using a timeline become widespread as it can represent data and the temporal ordering more stronger and comprehensible [7, 91, 122] .

Many approaches similar to [10, 21, 163, 179] provide a not scalable and limited interactive timeline to visualise events and offer an event management framework by using text, colour, and streamlines. [10, 57, 232] introduce different method to address such challenges by incorporating different visual properties and interfaces.

7.2.3 System overview

LifeTracker is designed to represent personal life pattern and allow individuals to explore their data and grasp meaningful information. To this end, the novel data mining and a visualisation method are employed. The data mining is utilised to turn the every day life data into an interesting fact. The visualisation in LifeTracker follows the principle of “overview first and details-on-demand”. This is achieved by a multi-layer timeline which allows the users to select a time/date point of interest, at a desired timescale (i.e. year, month, or day). Overviews are provided to allow users to gain an overall understanding of their life patterns and changes over different time periods. The users are able to explore more details at

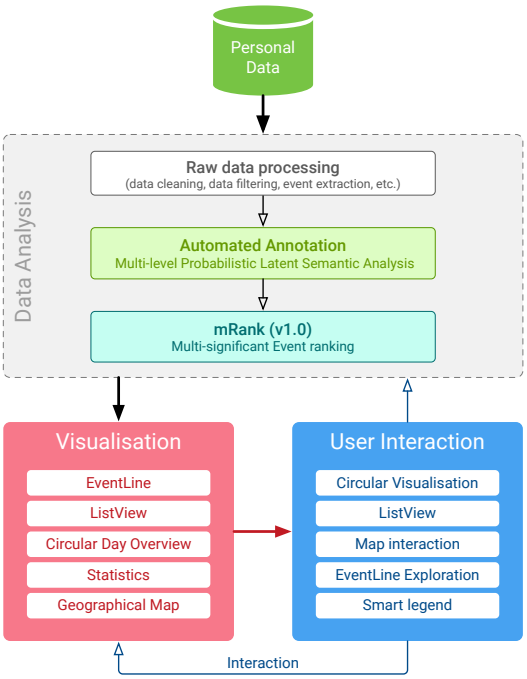


FIGURE 7.8: The pipeline designed for LifeTracker

a finer scale by using mouse hover and mouse clicks (Figure 7.8). The information is presented through a number of juxtaposed panels including:

- Life pattern: shows the daily life patterns of a selected day in detail, or the summarised daily life patterns of a selected month or year.
- Map: shows geographical locations or movements on selected days on a map.
- Physical activity statistics: shows the statistics about lifestyle, such as step count, walking distance, and transportation time.
- Event list view: shows a textual list of events that occurred within a selected timescale and period.
- Statistics view: shows the overall statistics of data within the selected time range.

LifeTracker is empowered by a number of techniques for data analytics, including activity extraction, to allow for the recognition of daily activities according to the

geo-location and movement data, data ranking techniques for the discovery and highlighting of important activities such as the highest step count in a year/month, the longest travelling time, etc., and data summaries at multiple scales, including year and month.

7.2.4 Data Analysis

The data used in LifeTracker is described in Chapter 4. These data consist of sequences of locations and activities over time. As mentioned in Chapter 5, a day event/activity matrix D_{ij}^{d} is used to form a daily matrix on which the activities took place d, where i ranges from 0 to 23 representing the hours of the day, and j is one of the categories. The matrix is:

$$D_{ij}^{d} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,j} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,j} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i,1} & a_{i,2} & \cdots & a_{i,j} \end{pmatrix} \tag{7.1}$$

For simplicity, and without loss of generality, LifeTracker exploits the locations of the activities in nine categories in accordance with Foursquare's hierarchical category guide (v.2015)[1], as follows:

- Residence
- Professional
- Shop & Service
- Food
- Travel & Transport
- Healthcare (GP surgery, dentist, hospital, etc.)
- Outdoor & Recreation
- Art & Entertainment (cinema, theatre, nightclub, museum, etc.)
- School & University

[1] Foursquare continuously updates its service based on new places. This categorisation is based on the information provided in 2015

Based on D_{ij}^{d}, the visible activity at hour i on day d is decided by the following equation:

$$VD_{i}^{d} = \text{argmax}_{j}\, D_{ij}^{d} \tag{7.2}$$

In other words, VD_{i}^{d} is used to calculate and then display the activity at hour (i) on day (d) according to the major activity within the hour. Apart from this primary activity, the other activities are also stored and can be used in the calculation for activity ranking.

The ranking procedure in LifeTracker is explicitly described in Chapter 5.

7.2.5 Interactive Visualisation

The visualisation and interaction of LifeTracker are carefully designed to maximise its usability, and hence to support user exploration of the data requiring only a very few mouse interactions. The visualisation includes the following components: multi-layered timeline, life pattern canvas, map, and statistics (Figure 7.9). The multi-layered timeline has three layers called year, month, and day layers, representing the time at different scales including year, month, and day. By default, LifeTracker is set into the LongMode, which presents a summarised yearly overview of all the available data along the timeline. The user is able to perform the following interaction in order to view the data at different scales:

- *YearMode* via year selection – a year selection can be made by clicking on the year layer on the timeline. By this, the time axis is set to the YearMode and displays all the events and activities across the 12 months of the year selected.
- *MonthMode* via month selection – a month selection can be made by clicking on the month layer of the timeline following the selection of the year. This shows the entire month pattern along the timeline.

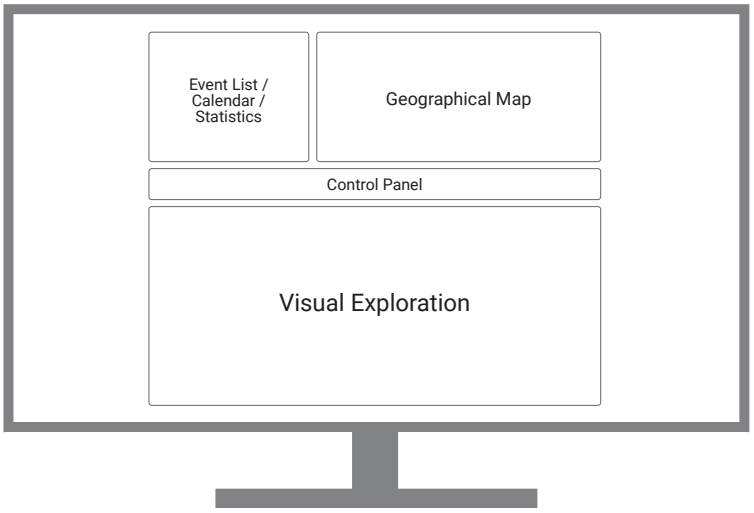


FIGURE 7.9: The interface designed for LifeTracker – a visual analytics approach to represent the life pattern

- *DayMode* via day selection – a specific day can be selected by clicking on the day layer following the selection of the month and year. By doing so, all the activities and related information within the day are displayed by circular-based visualisation (Figure 7.12(b)).

The LifePattern is a canvas to display daily life patterns of the individuals across 24 hours and uses a 2D-matrix view, in which the x-axis shows the time (i.e. day, month, or year) and the y-axis shows the 24 hours of a day. Each box in the matrix represents an hour slot in the day, month, or year. The box is colour-coded according to the analysis of the major category of activities that took place within the hour (Figure 7.10).

The LifePattern is presented at multiple time scales. The time scale is defined by the user selection on the multi-layer timeline, and it can be presented either as the year (LongMode), month (YearMode), or day (MonthMode):

- When the timeline is set in the LongMode, a box with coordinate (i, j) represents a year summary of the ith hour within year j. The year summary calculation is explained in Chapter 5.

- When the timeline is set in the Year mode, a box with coordinate (i, j) represents a month summary of the i hour within the month j of the year. Similarly, the calculation of the month summary is given in Chapter 5.
- When the timeline is set in the Month mode, a box with coordinate (i, j) represents the ith hour of day j of the selected year and month.
- If the user selects a day from the timeline, circular-based views are used to illustrate the activities on the selected day. The activities are colour-coded and placed along the two clocks in order to show the activities within different hours.

The LifeTracker is designed to be highly interactive. The users are firstly presented with the overview (i.e. summary) with key hours highlighted in the LongMode. To explore more details, the users may:

- use mouse hover on the legend for colour coding to see all the corresponding hours in the view (Figure 7.11).
- use mouse hover on the events marked on the clocks for more detailed information of the activities.
- use mouse click on the events marked on the clocks to view the events on the map.

The map view is employed to show daily activities and their locations geographically together. More specifically, it uses:

- Straight lines to show the movement path.
- Heatmap to show the level of activities at different places – red indicates more frequent activities while green and yellow are used to show areas that are less active.

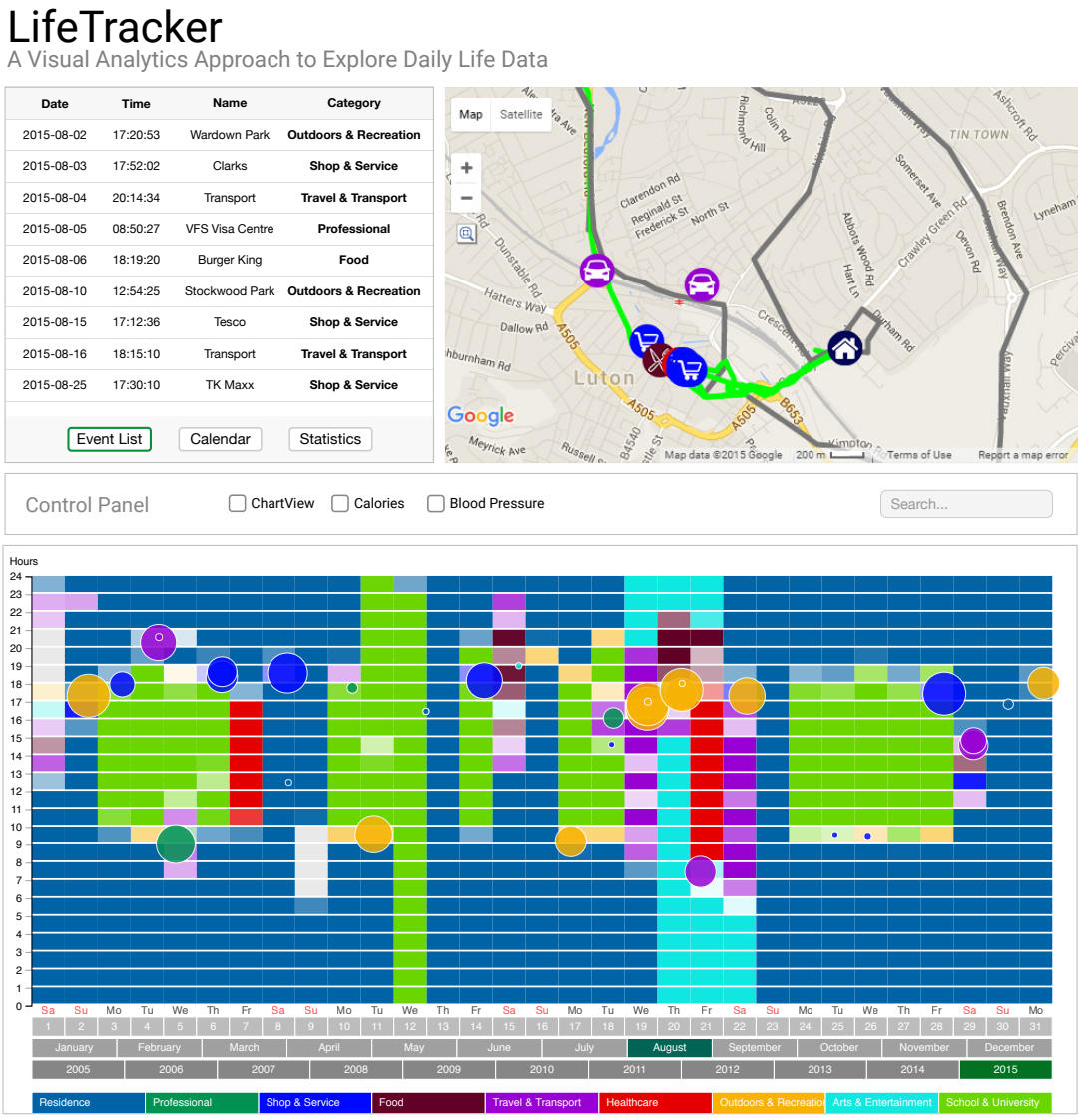
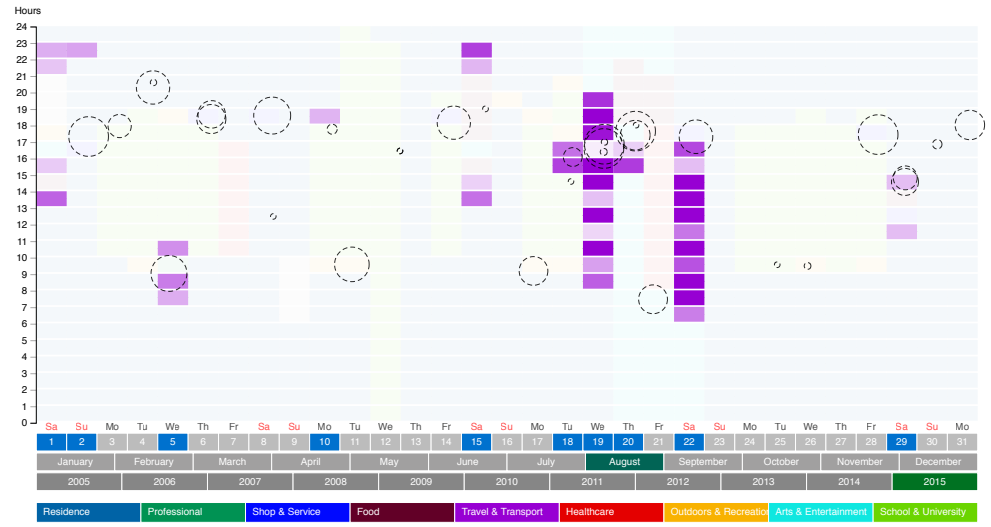


FIGURE 7.10: Overview of LifeTracker implementation to represent the life pattern.

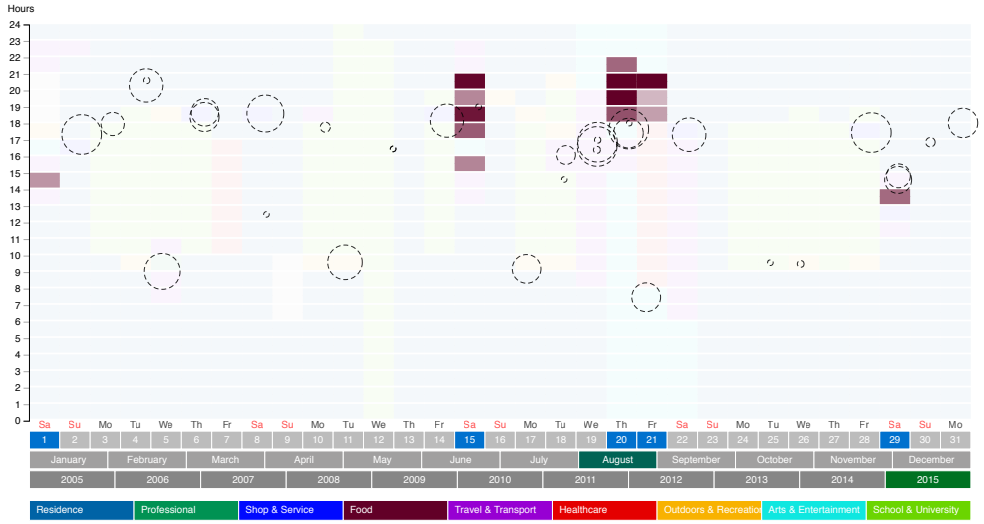
- Glyphs to show the key events in the selected time period (via the multi-layer timeline).

Subsequently, users are able to perform a range of interactions within the Map, including:

- Mouse click to explore more information of events and places on the map.
- Zoom in/out on the map; zooming in can also be achieved by selecting over a region of interest.



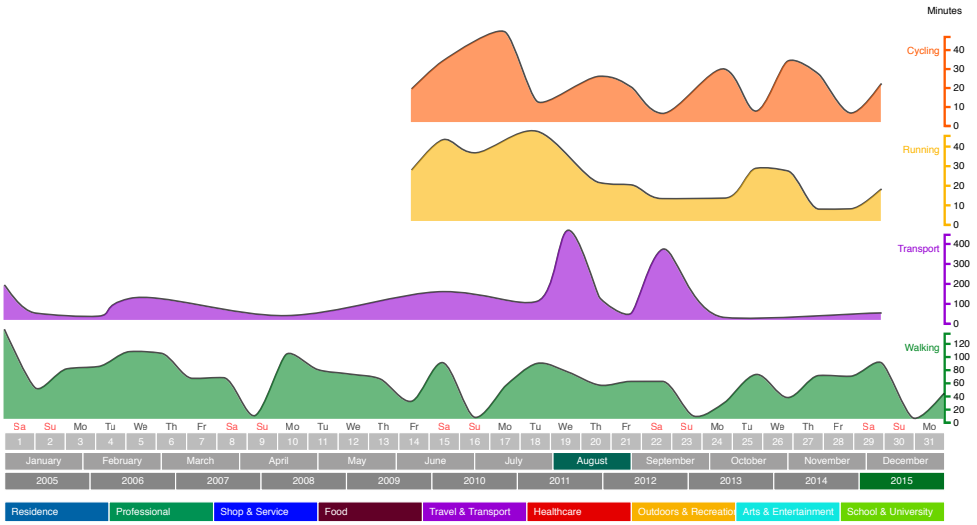
(a) The life pattern shows only the transportation pattern within the selected month (in this case, August) along the timeline. The days with transportation activity are highlighted on the timeline.



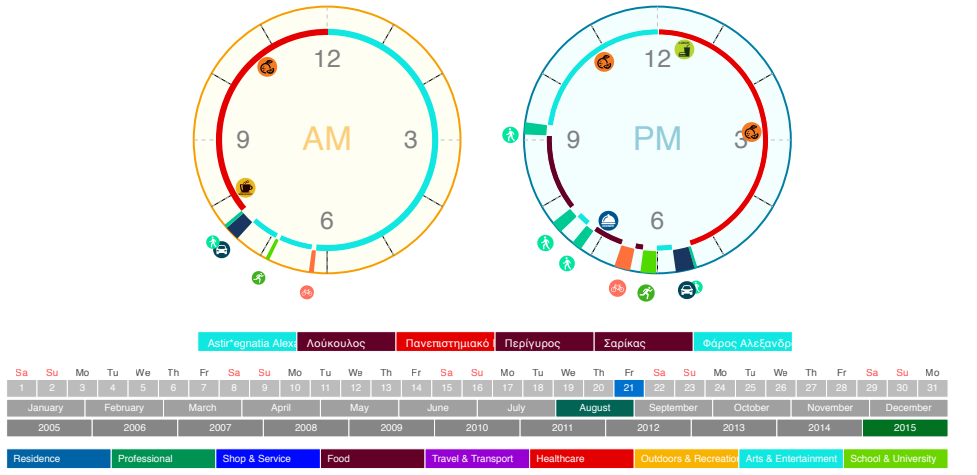
(b) Food-related activities and their occurrences are highlighted in this view along the timeline.

FIGURE 7.11: Interaction with the category legend triggers the data filtering function, which can dramatically elevate the level of focus on a particular category.

The statistics are designed to represent the movements such as step counts, transportation distance, or calorie consumption during the selected period of time by using histograms and line graphs. This encoding is updated according to the different time scale selected by the users (Figure 7.12(a)).



(a) The statistics display the change of the available physical activities as an overlay.



(b) The circular-based overview shows the entire day including physical movements, places, logged food, and sleep information (if available).

FIGURE 7.12: A representation of physical activities and circular-based daily overview.

7.2.6 Evaluation

The evaluation was conducted to investigate the effectiveness of LifeTracker in terms of supporting individuals in understanding their life patterns and exploring key events by asking the participants to complete a number of typical tasks based on their performance on the working prototype of LifeTracker.

In total 14 volunteers were recruited – 12 males and 2 females, aged 18–60, with

different educational backgrounds including business, IT and computer graphics, and English literature. The participants with a computer graphics background had a high level of IT skills, but none of them had used any life tracking devices or exploration in the past. All of the participants had normal vision without any colour deficiency. The evaluation was conducted via PCs (Intel Core i7 CPU, 16GB Ram, and 2GB graphics memory) with standard 24-inch desktop monitors. Standard mice and keyboards were provided for the evaluation. The tests were run using the Chrome browser full-screen without any distraction.

The experimental data were generated from four volunteers who had actively used the *Moves* application for approximately 18 months. The personal data were anonymised and concatenated to create a synthetic avatar that has five years of history. Corrections and validation were performed to make sure that activity patterns for weekdays and weekends were preserved. The geo-locations including track points, were untouched and original, which allowed users' tagging of information to stay valid for the evaluation.

7.2.6.1 Methodology and procedure

The study, in general, began with a brief tutorial to the participants about LifeTracker and its functionalities. The participants were given a free trial time to make themselves familiar with the approach and then asked to complete seven given tasks. After the answers were submitted, the participants were ask to complete usability questionnaires, accordingly. Lastly, five participants with different backgrounds were interviewed to understand their opinions about the approach and the level of knowledge they gained during the evaluation process.

The evaluation was designed in three parts: task implementation, usability evaluation, and interviews. Seven study tasks covered key functionalities offered by LifeTracker including:

- **Searching for places/visits/activities:** These tasks ask a participant to find places visited or activities conducted within a given time period (e.g. day, month, or year).
- **Understanding daily life patterns:** These tasks ask for daily life patterns from the data, such as the daily activities during weekdays or weekends, average time of starting and finishing work, etc.
- **Searching for important events:** These tasks search for important events that are identified by the LifeTracker system.

For each task, the user was expected to provide a straightforward answer to the associated questions, e.g. the number of visits to healthcare centres. The accuracy of the answers and the completion time of the tasks were recorded in order to analyse the performance. Table 7.2 shows the tasks involved in the evaluation process.

Task	Question
T1	Provide a list of all the overseas visits in July 2015
T2	Provide a list of all the visits to health centres in 2015 (inc. date, time, duration and location)
T3	Provide a list of all the activities conducted on 22/08/2015 in chronological order. (inc. places visited and transport taken)
T4	Look into the daily life pattern during the weekdays in May 2015 and provide the following information: the average start and end times of the work; how many days he did NOT stay at home overnight?
T5	Provide a list of yearly important events of 2015 by describing them in term of dates and categories (one item for each category only).
T6	Find out the longest and the shortest walking duration in the last 5 years and write down their dates and time.
T7	Find out the number of days and the total duration of the shopping activities in March 2015.

TABLE 7.2: Tasks designed to study how LifeTracker can help the user gain more insight

Moreover, the usability questionnaires were designed to test a number of usability aspects – functionality, efficiency, usability, reliability, and interface. The participants were asked to provide the answers on the scale of 1–5 for the scale-type questions after their completion of the tasks. Answering these usability questions was not timed.

The interviews were used as a means to further explore the user experience and to obtain their feedback on the system, which are difficult to capture using multi-choice or scale-type questions. The interviews were conducted with five selected interviewees – 3 males aged 33–60 (moderate to high IT skills with computer graphics background) and 2 females aged 27–32 (moderate IT skills with business background). They were asked to explain how LifeTracker would help them identify hidden parts of the data and whether it provides sufficient insight. All the comments and lessons learned materials were used for the further experiments and also to improve the functionality as well as usability of LifeTracker.

7.2.7 Result Analysis and Discussion

The evaluation were analysed and the result is organised in three parts: Accuracy, time performance, and usability. The accuracy of lifeTracker is calculated based on the correct answer to the tasks. The score for the tasks with two individual part where divided by half to deal with the partially incorrect answers by assigning half a score. Subsequently, the related performance to the task completion was analysed individually to determine the complexity of completing each task via using the provided tools. The usability scale-type questionnaires were analysed accordingly by fitting the answer into a scale of 1 (low) to 5 (high).

7.2.7.1 Accuracy

Overall, the participants achieved 76.5% (SD=0.07) accurate answers to the tasks. Figure 7.13 provides more detail about the tasks and the responses. Task 3 and task 6 had the highest accuracy, with participants able to use the combination

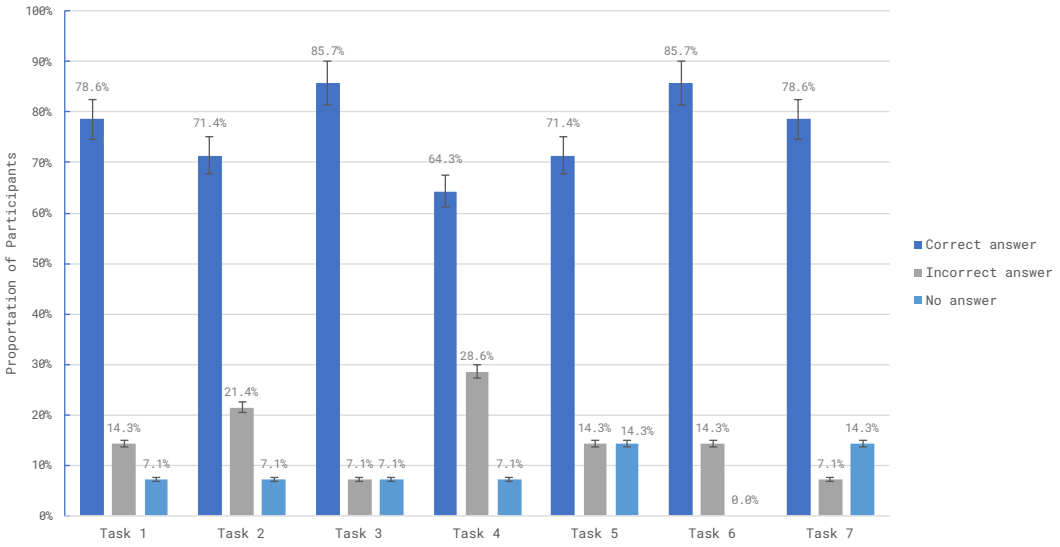


FIGURE 7.13: LifeTracker task accuracy in detail

of different views and interactions to answer successfully. The brief observation revealed that some of the participants double-checked their answers via existing functionalities – which was not anticipated. Correspondingly, task 3 required more focus on the daily activities in which more than 15 items were involved. It has been learned that the participants completed this task in 2 different ways: 67% of the participants (group 1) used the daily circular-based visualisation to get hold of all the activities while 33% (group 2) answered the task using the event list sorted according to time. The result showed that group 2 were considerably faster in providing the summary of the day but with more mouse clicks in order to see additional information regarding specific events, whereas group 1 found the information by interacting with the circular-based visualisation.

On the other hand, task 4 was answered with the lowest accuracy amongst all the tasks. This task required extra attention to two different aspects of the data at the same time, which could be fulfilled by using the legend and its interaction, but it has been learnt that the interaction provided within the smart legend did not sufficiently support the users' needs. The accuracy on task 5 was satisfying. Most of the participants used the life pattern view and the detected rare-event bubbles to complete the task correctly. In task 6, the participants faced an unfamiliar statistical graph with a monotone interpolation and needed more focus to complete

the task. In this task, the participants were required to trigger the physical activity graphs and look for a peak of walking over the period of two years on the timeline. It has been seen that the interpolation was slightly misleading to the non-expert users to find out the exact day of an event, but this did not have any effect on finding the event duration for the second part of the task. Task 7 was performed smoothly and most of the participants were able to make use of the provided functionalities to get the correct answer. However, some wrong answers were found for the day-counting related tasks, which could be a human error. It has been learnt that human errors can affect the accuracy of knowledge and therefore this needs to be addressed within the visual analytics approach.

7.2.7.2 Time performance

The average performance for completing tasks correctly was analysed to determine the minimum and maximum completion times. Figure 7.14 shows the range of the times over which the participants completed each task.

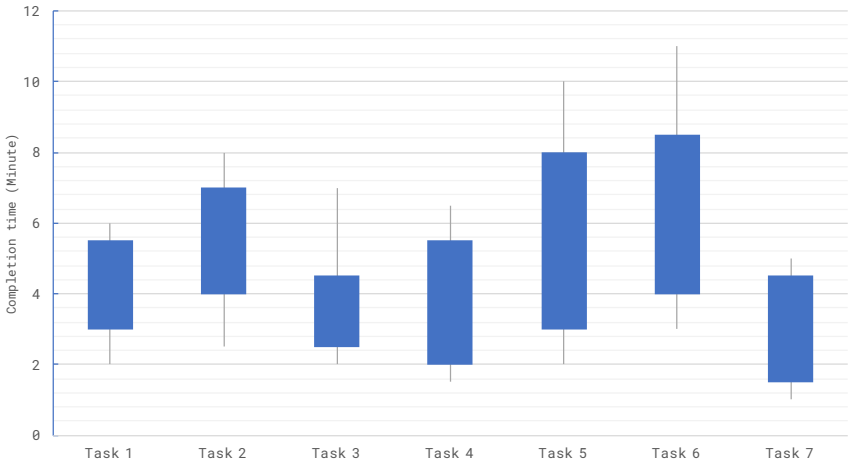


FIGURE 7.14: Completion time range for each task plus the expected minimum and maximum times

On average, each task required 4–5 minutes, and the most time-consuming tasks were tasks 5 and 6 based on their complexity. The analysis showed that the completion time is reasonable. This indicates that LifeTracker as a visual analytic approach does not include a complex method or interface.

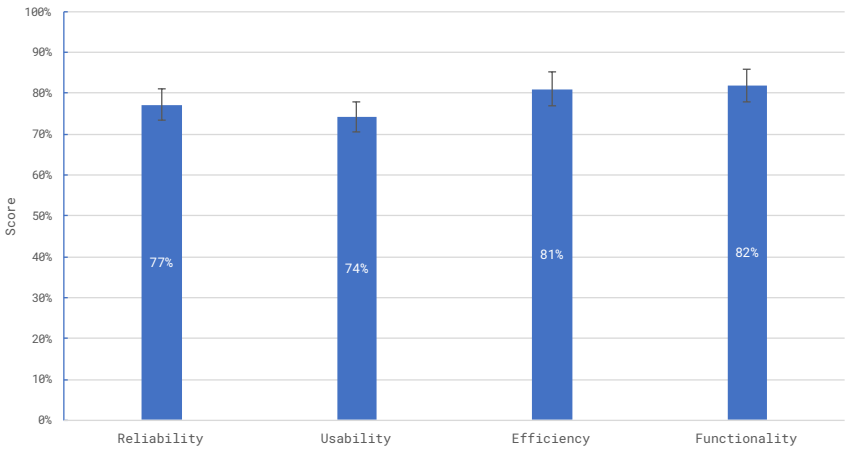


FIGURE 7.15: Overall usability rated by participants

7.2.7.3 Usability

The usability survey consisted of four different parts, namely, functionality, efficiency, usability, and reliability. Each part was rated between 1 (Low) to 5 (High). The usability includes the Computer System Usability Questionnaire by [130] and the user interface satisfaction adapted from [47]. In the following presentation, the scores are changed into percentages. Figure 7.15 shows the overall usability result. More detail regarding each part can be found in Figure 7.16.

Functionality: 82% of the participants were satisfied with the LifeTracker functionalities, in general. The non-expert participants needed more assistance with using the functionalities provided properly.

Efficiency: To some extent efficiency was affected by two factors, age and IT skills. Unsurprisingly, young users with higher IT skills finished the tasks faster than the others. The overall efficiency came to 81%, which reached the platform expectation.

Usability: The average score was 74%. It has been learnt from the comments that some users felt that the a number of alternative options provided as a juxtaposed view were slightly overwhelming or needed more guidance in order to use them.

Reliability: The participants were asked if they encountered errors during the evaluation session. Only the participants with the high IT skills answered this question. On average, LifeTracker achieved 77% for reliability.

Furthermore, some of the collected comments from the participants' feedback form are as follows:

- *It's quite useful, intuitive and reliable for finding out activities pattern.*
- *It is simple and handy to see the life pattern. The shopping pattern was interesting as you could see how often you do the shopping!*
- *Does what it needs to and is easy to understand.*
- *Good idea but a little bit overwhelming as there are too many possible functions.*

7.2.7.4 Limitations and future work

The results of the evaluation show that LifeTracker is able to accurately analyse and visualise the daily life patterns of the individuals and provide effective overviews as well as the detailed information. However, a number of pitfalls within the functionality and usability were identified that required further improvements:

- LifeTracker uses an initial version of significant ranking algorithms in detecting and ranking the events, which may result in detecting insignificant events that are not of interest.
- LifeTracker computes the data in real time; the use of optimised and well-written algorithms would allow handling large-scale personal data with scalability in mind.
- While the category legend and its interaction could significantly help the participants explore the data, there is a need for combined interactions and for allowing the users to filter out results.

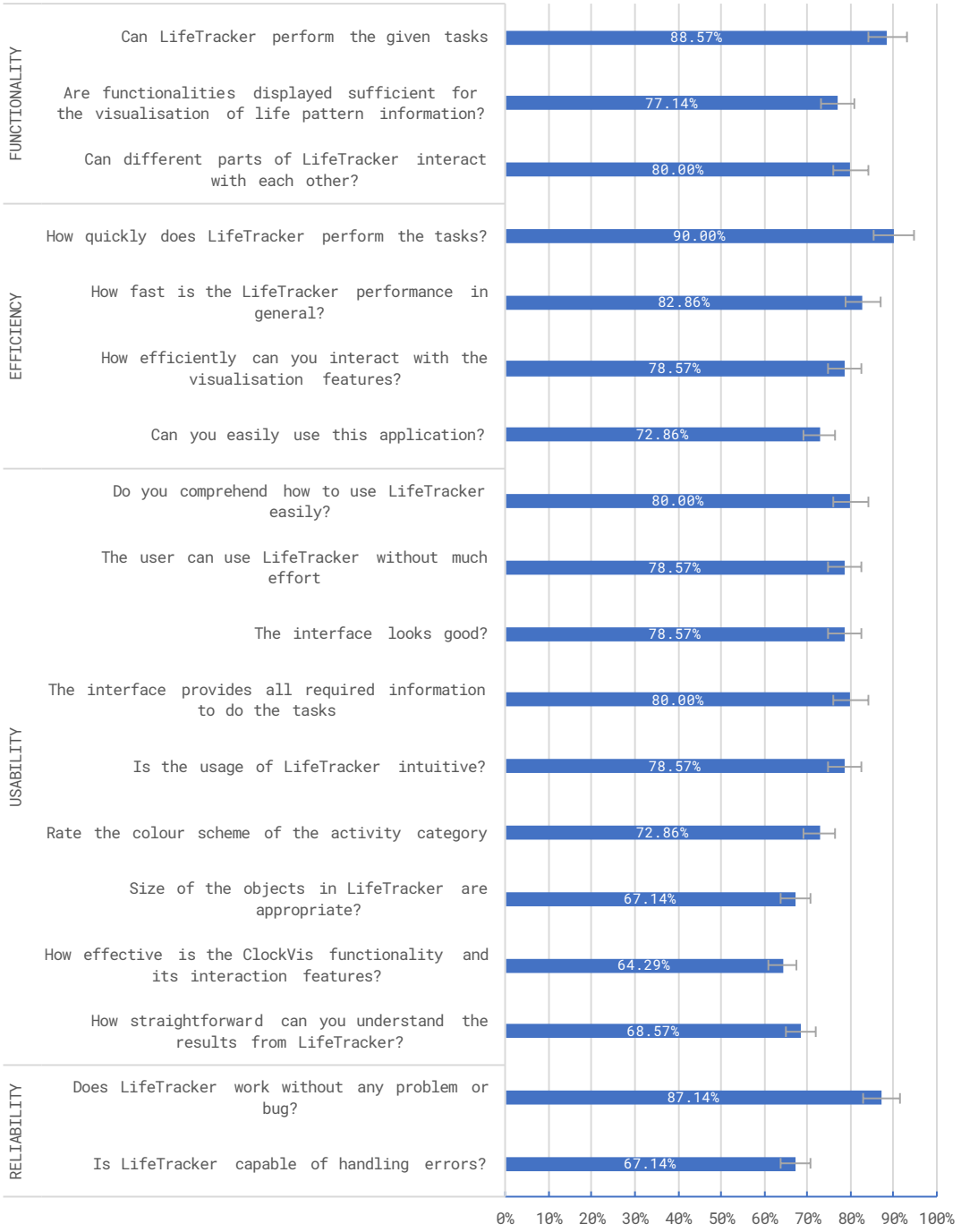


FIGURE 7.16: Usability questionnaire result rated by the participants in detail

- The intuitiveness and engagement of LifeTracker were slightly low.

Furthermore, based on the evaluation results and the observations, further improvements were identified:

- The event ranking model to allow user preferences for identifying significant events and hence provide more effective results;
- The colour scheme for the categories to be optimum and more visible within the visualisation;
- The optimisation of the implemented algorithm to deal with large-volume daily personal data; and
- There is a need for a guidance and a comprehensible user guide to raise user engagement.

Many of these improvements were made available within the next experiment analysing the personal data in support of reminiscence. The next section delves into a particular subject and assesses the visual analytic approach in detail.

7.3 Visualisation for Reminiscence (MyEvents)

In this use case, the aim is to apply and evaluate the proposed approach to support memory recall and reminiscence as an important aspect in personal life. Reminiscence preserves precious memories, allows us to form our own identities and encourages us to accept the past. It is also highly valuable for medical treatment in mental health by offering therapies for conditions such as memory impairment and depression. However, conducting an investigation into the impact of this work for corresponding medical treatments is not within the scope of this research. MyEvents exploits modern sensor technologies to support memory recall through reminiscence, enabling self-monitoring of personal activities on a daily basis and creating data containing massive and valuable hidden information about individual events/movements in space and time. As mentioned in the previous chapters, given the size and complexity of the recorded personal data over time, identifying significant events from daily tracking data for the creation of event mementos still constitutes a key challenge due to the vastly more numerous background of trivial events.

To address this challenge, MyEvents, a web-based personal visual analytics platform based on the proposed approach designed for non-computing experts, is presented. This platform allows for the collection of long-term location and movement data, from which events are extracted and event-mementos can be generated through the personal visual analytics process. The focus of MyEvents is placed on two main goals, namely, selection subjectivity and event familiarity in event reminiscence.

MyEvents features new techniques that have been discussed in chapter 6 to support selection subjectivity from daily events, and novel visual presentations for event mementos to evoke event familiarity. Moreover, the novel automated place annotation with multi-level probabilistic latent semantic analysis (chapter 4) and multi-significance event ranking model (chapter 5) are utilised to enrich the data with systematic annotation and to identify significant and notable events

in personal history according to user preferences for category, frequency, and regularity of events respectively.

This section depicts that how the novel data mining and visualisation methods empower MyEvents in order to support non-expert users in the process of reminiscence. The most research works related to MyEvents such as reminiscence visualisation, personal visual analytics, and time-oriented event visualisation are described in the literature review (chapter 2). In each part, the relevant section or chapter within this thesis in regard to the novel data mining or visualisation is referenced accordingly. The evaluation analysis and result, which show the impact of MyEvents on reminiscence and personal knowledge discovery, are discussed within this use case.

7.3.1 Introduction

Reminiscence is the act of recalling memory of past events and experiences. It constitutes a very important part in our life in terms of preserving precious memories, forming our own identities, and accepting the past [129]. Studies on episodic memory suggest that a constant review of personal life helps improve emotions with families and friends [36], and allows us to seek solutions for present issues by looking into past experiences [222]. Reminiscence is also highly valuable for medical treatment in mental health [95], offering therapies for conditions such as memory impairment and depression.

Mementos are objects that are kept as a reminder of significant experience in the past. Prior work on physical mementos shows the importance of everyday objects for reminiscence [160]. Studying the digital mementos has attracted significant attention in recent years [106, 155, 158, 159, 208, 227]. Compared with many previous works, this use case casts its focus on automatic annotation and the recall of a set of key events that an individual has experienced in their personal history, thus facilitating an important part of reminiscence by bringing back memories at key locations such as home, school, work place, holidays, and other activity

venues. Event mementos also provide important contextual information for object mementos, such as a souvenir bought during a holiday trip or a gift from a friend during dinner in a restaurant and the like. MyEvents takes advantages of modern sensor technologies (e.g. smartphones with built-in GPS sensor), which nowadays enable self-monitoring of personal activities on a daily basis, leading to a spectrum of personal location and movement data along the personal lifespan to support long-term memory. However, given the huge amount of data captured over a lengthy duration, it is difficult for the individuals to find desired information from their own data with a large amount of trivial information. Studies of digital archives also show that collections of large and poorly organised digital objects often become invisible and inaccessible over the course of time [226]. To this end, visual analytics can be of great assistance in terms of supporting effectual organisation, search, utilisation, and encoding of such data for reminiscence. Compared with the previous work on similar topics [25, 204], MyEvents offers an environment that allows for the analysis of long-term location and movement data, from which event-mementos can be readily generated through personal visual analytics. The research focus in these integrated tools is placed on the improvement of the following aspects in event reminiscences:

1. **Selection subjectivity**: the human involvement in the process of annotation, obtaining significant events, exploration, and memento creation; and

2. **Event familiarity**: the presentation of events and information associated with the target events for optimal memory recall.

Human participation plays a key role in the course of reminiscence. Researchers have suggested that autobiographical memory is not a place that simply stores all previous events, rather, it is a subjective interpretation of the past. People like to be actively involved in the process of selecting and organising mementos [79] and providing narrations to support reflection and emotion. In particular, studies have demonstrated that using personal contextual information within the

process of mementos creation is often more relevant in reflecting on memory than impersonal news [177].

In the course of designing MyEvents, supporting the active involvement of individuals was targeted in the process of reminiscence by searching and selecting from personal location and movement data acquired via mobile technologies over a long period of time to expedite generating the digital event memento procedure. Personal visual analytics is involved for interactive exploration of targeted events and their information from the data. This also belongs to the area of "casual information visualisation" [168] in which the targeted end users are normal citizens instead of computing experts.

MyEvents offers an integrated environment of data analytics, visualisation, and human-computer interaction, featuring new techniques to support selection subjectivity from daily events, and novel visual presentations for event mementos to evoke event familiarity. In a nutshell, MyEvents covers the following contributions of this research:

- A novel multi-significance event ranking model called mRank, which identifies significant events in personal history according to user preferences through ranking, allowing the users to efficiently identify key events over a selected period of time based on their personal preference settings to create mementos, including the preference for event category, occurrence frequency and regularity. The frequency preference enables users to make choices between frequently or infrequently visited places. The regularity preference models the occurrence changes. And through the event category preference, the users may set priority for the types of events in their query.

- Interactive visualisation to support heuristic search for significant events. These include the timeline and map-based visualisation to allow users to gain an overview of personal events over the selected time period; a search bar that allows the users to seek events with multi-keywords including name, category, and preferences via multiple and heuristic keyword entry; hints

based on each of the newly entered keywords to guide the search; and a control panel that controls the event search via a graphical interface.

- A novel visual presentation of event mementos by visually encoding a set of heterogeneous information about the event, including time, photos, location, statistics, and contextual information. In addition, associations with other events (e.g. events on the same day) are highlighted to further enhance memory reflection.

7.3.2 Summary of related works

Despite the full review in Chapter 2, a brief overview is provided here to particularly picture the related work to this experiment. Reminiscence visualisation is strongly related to personal visualisation and visual analytics. Unlike the analysis of population data, which is very common in applications such as trajectory mining, personal visual analytics particularly focuses on the visualisation and analysis of personal data.

The work in reminiscence visualisation is tremendously limited. Stream of our Lives (LastHistory) [25] focuses on temporal patterns in personal music listening histories to facilitate reminiscing. An interactive timeline matrix is used to visualise everyday music listening history on a 24-hour basis together with contextual personal information such as photos and calendar events. Daily streamed songs are typically lined up vertically on a timeline in colour-coded circles to represent the genres and songs ranked based on frequency of listening. The interaction of this work shows the links to similar songs and song sequences to facilitate reminiscing and mood analysis. The frequency-based song-ranking algorithm introduced in this method is beneficial for mood analysis. Another work by Dias [60] uses a different timeline-based layout for the visualisation of music-listening history using stacked dots. A filtering feature is provided to view only the selected songs and the ranking is purely frequency-based. AppInsight [23] presents a visualisation tool that helps reminiscing about computer software usage history

with an hourly timeline matrix. A duration-based usage ratio is used to rank the software apps. Details of app usage can be obtained and comparisons of several apps are supported from the visual interface. None of the above work is designed to reveal important events from a large number of trivial events in daily life logging data. In contrast, this research addresses the problem by proposing a novel importance-based ranking algorithm in Chapter 5.

In some ways, this platform is close to VisualMementos [204], which integrates a timeline with a map followed by semantic clustering of GPS logs to analyse and visualise personal movements at different temporal and spatial scales, hence supporting reminiscing for self-reflection and memento sharing. They use different sizes of circular map segments along a time axis in chronological order to illustrate visits or repeated visits and associated duration within a geographical area. In comparison, the approach proposed in this thesis casts more focus on selection subjectivity: the multi-layered significance event ranking together with the interactive visualisation are designed to support users' involvement in the creation process of the digital mementos by facilitating their selection of key events in their personal history for reminiscence. Meanwhile, MyEvents involves additional contextual information in the visualisation of the event mementos to evoke event familiarity, such as other events which occurred on the same day. Such contextual information is proved to be important in terms of helping memory recall. Finally, the data mining and visualisation in MyEvents offer more semantically meaningful outcomes through the provision of event category information.

As trajectory data often do not come with semantic information, most of the existing works focus on pattern mining from the trajectories of a large population or the automatic generation of semantic information. Krueger et al. [119] introduce context data into trajectory data analysis and their main work is to find potential places from a large group of trajectory data. Andrienko, Andrienko, Hurter, Rinzivillo and Wrobel [13] extract significant places from population trajectory data. They, recently, began to study privacy-respectful discovery of place semantics [18]. The most relevant work to this approach is the semantic enrichment of movement by Krueger et al. [120]. This work attempts to semantically enrich the

trajectories by undertaking a preprocessing to identify the destinations (places) and using a POI decision model to interpret the identified points only in a categorical way by calculating the overall category score of the nearby places, which can lead to two main issues. First, the enrichment only contains the most likely category of the identified places but the actual names are still not discovered. And second, the calculated categories are less compelling within the places with dense POIs as the calculation only incorporates the distance, number of check-ins and users. This drawback is fully addressed in Chapter 4 by introducing the multi-level probabilistic latent semantic analysis.

7.3.3 Definition and data

MyEvents provides an automated semantic enrichment as well as interactive query and mining for significant events via a user-centred interface consisting of: a search box, MyMoment (an interactive presentation of event mementos), a control panel, EventLine, which provides an overview of all events and the event category legend – see Figure 7.17.

The data collected by *Moves* and *SmarTrack* constitute low-level semantic information for the places names and the movement types. This data can either undergo an automated place annotation with latent semantic enrichment (Chapter 4) or be manually labelled by the user with high-level semantic information through the tracking application. The automated place annotation extracts high-level contextual information by using the Foursquare POI service for the process of automatic annotation. Most of this high-level information is used to address the presence of unknown facts regarding events or places. Foursquare uses a hierarchical structure for its place categories and classifies available places (venues) into 10 top-level categories. Each of these categories has more detailed sub-categories to which a place belongs. For instance, the Food category comes with variety of restaurants or coffee shops in a different group – see Figure 4.3. The top-level categories used to classify events in this work are as follows: Residence, Professional, Shop &

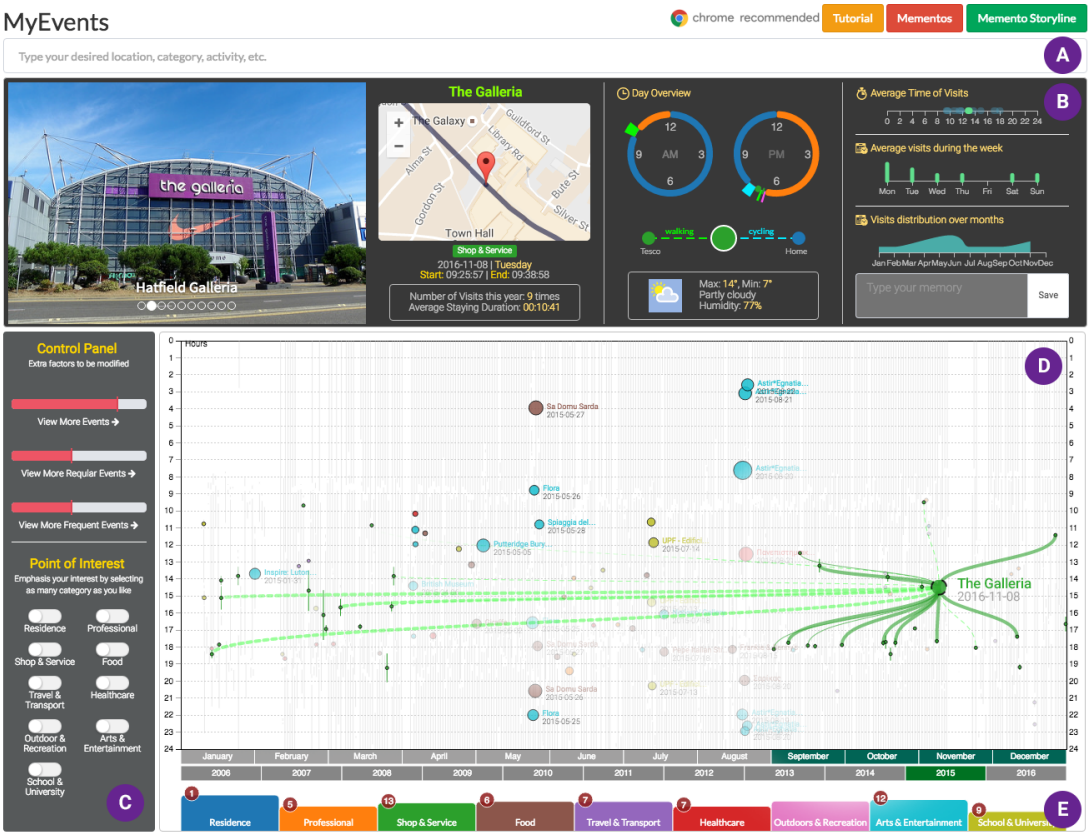


FIGURE 7.17: MyEvents main interface: A) Search box for event query; B) MyMoment: an interactive presentation of event mementos; C) control panel; D) EventLine: visualisation of events along a timeline with indications of their significance ranking; E) event category legend

Service, Food, Travel & Transport, Outdoor & Recreation, Art & Entertainment, School & University, Event, Nightlife, and Healthcare[2].

7.3.4 Design goals and requirements

The design goal within MyEvents is to advocate the process of personal memory recall for reminiscence by identifying significant events, visualising the events and their possible correlations, and creating event mementos via interactive visual analytics for personal data, with particular focus on two goals: event selection subjectivity and event familiarity. Event selection subjectivity is related to

[2] Healthcare is part of the professional category but based on the MyHealthAvatar requirement, healthcare is deliberately considered as an individual category to emphasise in this instance

personal involvement and user experience in the reminiscence process, in which users specify targeted events, grasp knowledge about them or any association between the other events, reminisce about past memories, and attempt to create digital event mementos or retrieve previously created mementos. Event familiarity refers to the presentation of the targeted events. A well-defined presentation can decisively support evoking memory familiarity by providing related information about the event, such as statistical facts, photos, and weather information to users.

Research on autobiographical memory has highlighted the importance of subjectivity in reminiscence, as human memory is not an objective storage of all previous events; instead, it is a personal self-reflection that requires significant involvement from human subjects to decide the events that reflect their memories. Whilst the location and movement data are available through self-monitoring sensors, users need support in order to retrieve and revisit their personal history from the big data and make decisions about their memorable places. MyEvents is designed for this purpose.

In this use case, the targeted end-users are considered as average users who are not familiar with information visualisation by any means. To reflect this, a number of strategies are employed to ensure the effective design of a simple and intuitive user interface. In fact, the interactive review of personal history itself is also part of a reminiscence process, and one of the missions for MyEvents is to offer a sound interactive tool to help users to enjoy this personal memory recall process through self-knowledge discovery and visualisation. The MyEvents design endeavours to strike a sound balance between the level of automation and user control, allowing users to seek desired information based on their own preferences with the support by the data mining techniques to achieve efficient information retrieval.

To fulfil the design goals, the following key requirements are identified through user requirement analysis using a set of questionnaires (Appendix A), and through literature reviews as described in Chapter 2.

Selection subjectivity (G1)

MyEvents needs to fulfil a set of identified requirements in selection subjectivity. Each of these requirements is briefly portrayed and also numerically referenced.

Gain overviews R1 : There is a need to provide an overview of all the personal events within a selected period of time to users. The overview should be obtainable before a user drills down for more details in the data. Hence, MyEvents needs to follow the "overview-first", "details-on-demand" methodology in the information visualisation.

Enquiry-based retrieval R2 : In a similar fashion to many online search services (e.g. Google, Bing) or well known websites, users are often keen to use natural languages that may involve a number of keywords to customise enquiries about their personal event history, such as names and categories of event location, date, time of the day, and weekday or weekend. Furthermore, they are interested in discovering events whether occurring frequently or infrequently, and regularly or irregularly. For instance, a frequently visited place could be either the user's home, school, or workplace, and an infrequently visited place could be associated with a holiday trip or the like. MyEvents requires providing suggestions and initial hints to users in order to support adjusting the search results adaptively.

Control panels R3 : This option should also be made available to users who prefer to set their search targets and preferences via graphical user interface components. Users should be able to define points of interest, frequency, and regularity together with their level of significance within the control panel.

Events visualisation and highlights R4 : Location and time information should be made available in the event visualisation. Important events need to be highlighted to facilitate user selections. Users need to define the importance of an event according to its name, category, date and time, together with preference settings, e.g. frequent or infrequent, regular or irregular. The system needs to highlight highly ranked events according to calculations based on user input. Furthermore, events with low rankings should be faded out in the process of

representation to avoid a cluttered view and overplotting. Users may also wish to view the associations between the events, for instance, repeated and sequential events within a certain time range.

Data upload R5 : MyEvents needs to allow users to upload their personal data. The platform needs to be equipped with a database that enables the storage of long-term personal data as well as mementos created by multiple users.

Create, save, and retrieve mementos R6 : The platform needs to allow users to save their digital mementos after exploration for re-access in the future. Users should be able to view and interact with the created mementos in one place. MyEvents should offer downloading of created mementos from the system to users who want to share their memories with their family or friends.

Event familiarity (G2)

Contextual details R7 : There is a need to allow users to view event information according to their location (e.g. geographical map), date, and time. In addition, contextual information about the events, including indications of other events that occurred on the same day, are also desirable to users.

Photos R8 : Photos taken by users during events or public photos from search engines to show the event venue are tremendously useful in terms of supporting memory recall.

Statistics R9 : Statistical information about an event is beneficial for supporting users to reminisce how often similar events have occurred. The statistical information may include average duration, average duration during the week, and distribution of the occurrence over weeks or months.

7.3.5 System overview

MyEvents is designed in harmony with the design goals and requirements. The system is composed of four modules, namely data repository, data analysis, visualisation, and interaction. The data repository accommodates the personal life tracking data, the automated annotation result, and the created mementos. The security of the data repository is built into the core of the system to protect users' sensitive data. The server is located in a secured cloud service in the UK, which is in compliance with regional legislation and data protection laws. The HTTPS protocol is enforced for the whole platform to protect users' privacy from misbehaviour such as eavesdropping. Authentication and authorisation is built on the Spring Security framework, which is the most widely used Java web security solution and tested by millions of users around the world.

> **Note**: The data repository and related procedure is out of scope of this research. More information regarding the database can be found on the official MyHealthAvatar website.

The data analysis module includes three sub-modules: the preprocessing of the raw data, automated annotation, and the significant event ranking computation, each of which are described in Chapter 4 and Chapter 5. The visualisation module represents the output from the data analysis module. It includes a number of visual components that have been defined in Chapter 6. The visual components are customised in MyEvents based on the requirements and users' needs. These components, namely, the event visualisation (EventLine), geographical map, My-Moment, memento storyline, and memento list are described within this section. Moreover, user interaction has a direct impact on the visualisation and the event ranking. It can allow users to explore events, get additional information, customise the event ranking (by using the control panel and search box), and save or retrieve the mementos – see Figure 7.18.

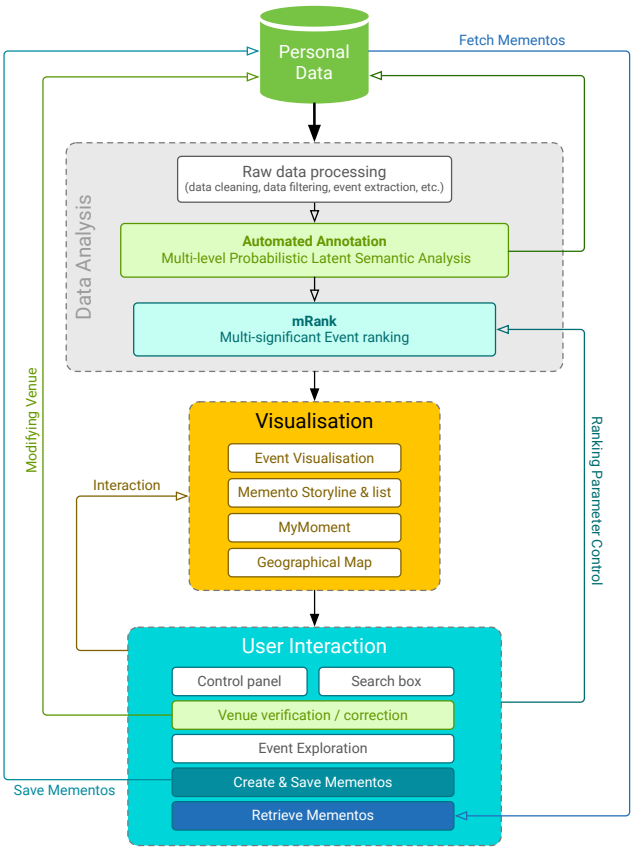


FIGURE 7.18: MyEvents pipeline overview

7.3.6 User interface

The user interface of MyEvents is composed of a search bar, a control panel, and a number of components for event visualisation. Figure 7.19 illustrates the MyEvents layout. While the search bar and control panel are designed for user-controlled event retrieval, the components for event visualisation constitute the key to the event mementos. These components, which described in Chapter 6, aim at effective assistance in exploring the events and gaining preliminary knowledge. In addition, the event memento is designed to present details of a memento in terms of its time, location, duration, and the like in order to provide the best support within the process of memory recall and reminiscence. The specific components of the event visualisation embrace the EventLine, geographical map, MyMoment, and event category legend. This section portrays the user interface and its components.

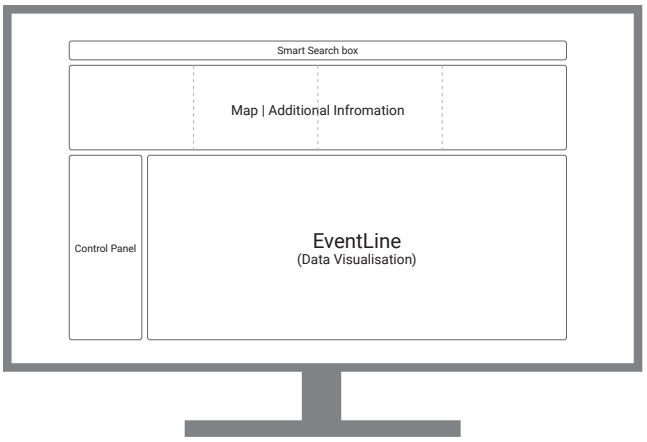
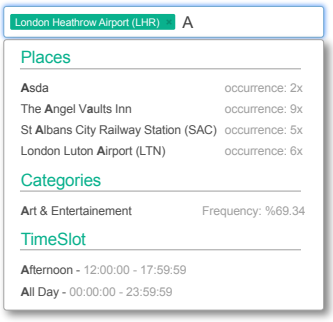


FIGURE 7.19: MyEvents layout wireframe

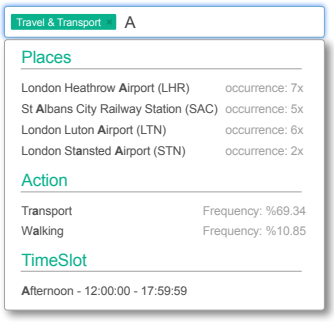
Search box: The search box is designed to allow event queries using keywords for named entities in natural language. This component fulfils the requirements from non-expert users who are keen to make queries about events according to their place names, time, category, and other preference settings such as frequency and regularity – see R2.

The search box offers smart assistance within the search. It provides categorical suggestions as well as initial facts within the process of search. The search can be incremental, allowing for entry of multiple named or condition entities. Auto-suggestion is provided to assist the search. It employs a fuzzy matching mechanism to instantaneously respond to the entered words by providing possible matches of places, categories, or time slots. The named entities (i.e. place names, categories, time, and day) plus the search preference can be entered in an arbitrary order to facilitate the user search. Figure 7.20 shows the search box functionality by an example of typing a keyword. Furthermore, the search box displays more hints in regard to the occurrence or frequency of the places and categories, respectively.

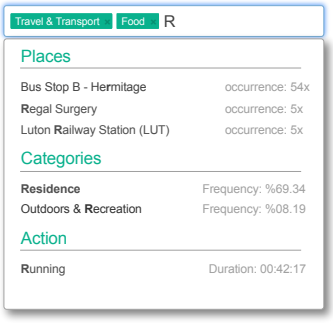
The incremental search can narrow down the process of search based on the keywords entries. This means that time slots, categories, and places can reflect the search and limit the suggestions in conjunction with the multiple entry keywords. As an exemplar, the user looks for a particular category by entering a keyword name, then if the user attempts to add another keyword (e.g. place name, time),



(a) The search box suggests the possible queries in a categorical way together with introductory hints regarding places, categories, etc.



(b) The search box reflects the rest of the search based on the provided category – Travel & Transport – keyword and suggests only the available places, actions, and times within this category.



(c) The search shows the suggestions based on the combination of two provided category keywords. The search still suggests other categories that match the entry.

FIGURE 7.20: The search box suggests the possible events, categories, or time slots based on the user input. The occurrence, frequency, and duration of the events, categories, and actions are also displayed as a hint to facilitate the search process.

the search box suggests places or times that are included in the provided category and limits the search result. This functionality can effectively increase the focus on the particular part of the data.

Control panel: The control panel is also designed as an alternative means to search, in which users can set the search parameters via a graphical user interface – see R3. More specifically, as shown in Figure 7.21, three range sliders are made available to allow users to control how many events to visualise, view more or fewer regularly occurring events, and see more or less frequently occurring events in the

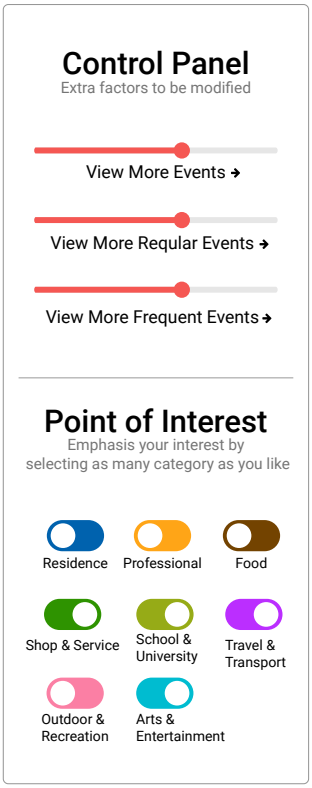


FIGURE 7.21: The control panel with two sections: a) the range sliders to control the number of events on EventLine, the regularity, and the uniqueness of events; b) the toggle button to indicate the categories of interest.

event visualisation. This adjustment is supported by the underlying significance ranking model that computes ranking of the events, which subsequently decides the size of the event icons in the event visualisation – EventLine.

The second part of the control panel allows users to modify the ranking and visualisation by indicating their category of interest.

The control panel is designed to provide prompt feedback during the interaction process by sending the modified parameters to the ranking module to recalculate and update the visualisation in EventLine, so that with every change in the control panel users see the change and can compare the visualisations.

Geographical map: The map shows the geographical locations of the significant events extracted by the ranking module by using simple markers. This provides a simple overview of the location of all the calculated significant events – see R4.

Moreover, additional open-source libraries are employed to cluster the points in a dense area and display them as a whole with a number of included points.

EventLine: The EventLine is a major visualisation component to present personal events along a modified multiscale timeline. A multiscale timeline shows time along with data information in only years and months through two individual layers. The modification of the timeline allows the users to select a year to display the events over 12 months. The second, month layer, unlike the LifeTracker, is not selectable but similarly can be highlighted to show in which months the selected event occurred.

EventLine offers a gainful overview of all the events in a selected period by displaying them along the timeline in a form of light grey solid lines – R1. It visualises the events according to their significance as calculated by the mRank model – R4. More specifically, the events are presented in a grid structure according to their time and duration, in which the horizontal axis denotes day, month, and year, while the vertical axis indicates the 24 hours of a day. As mentioned in Chapter 6 (section 6.3.1), many previous works [23, 85, 115, 164, 192] use a straight line as a natural choice to represent an event in which the length demonstrates the duration and the line width the significance. However, such a representation would visually persuade users to perceive a very significant event with only a short duration as less significant than a normal event with long duration. Increasing the link thickness cannot mitigate this problem due to the notable overlapping caused. Some previous works such as [25, 60] use a circle to address the problem but this can also lead to overlaps if there are many events with lengthy duration. A solution to this, which is extensively described in Chapter 6, is to presented an event as a glyph with a combination of a circle and a line by:

- using a circle to particularly indicate the significance of events according to its size;
- displaying the duration of events via using a solid line; and
- employing colour scheme to indicate the event category.

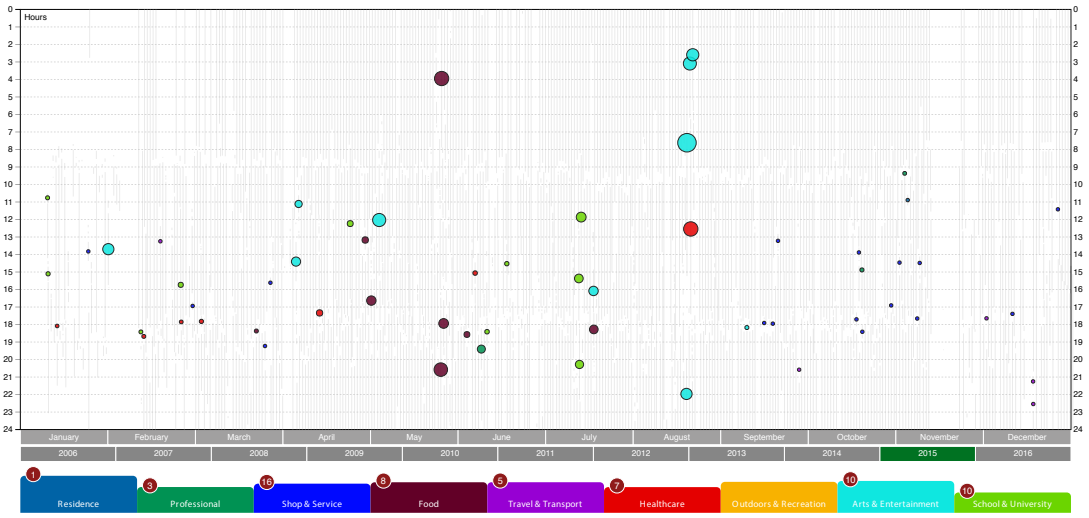


FIGURE 7.22: This approach in visualising the significant events together with all the available events within the data can assist the visualisation to avoid visual clutter.

Figure 7.22 shows all the recorded events as well as significant events by means of only grey solid lines and circles with different sizes and colours, respectively.

MyMoment: To evoke event familiarity (G2), selected events are presented in a panel called MyMoment, showing the contextual information R7, photos R8, and statistical information R9 about the selected event.

The contextual details include map, time, duration, weather, and day overview. Sequential events that occurred before and after the selected event are also visualised in lines and circles. A compact form of circular-based visualisation (ClockView) is used to show when the event occurred in the context of the entire day (day overview) using two interactive radial clocks for morning and afternoon, respectively. The events and movements that happened on the day are placed on the AM/PM clock with alignment to the time and consistent colour coding according to their categories.

Photos from the event are also displayed within the MyMoment frame. These include private photos from users themselves or public photos that show the event venue and surroundings.

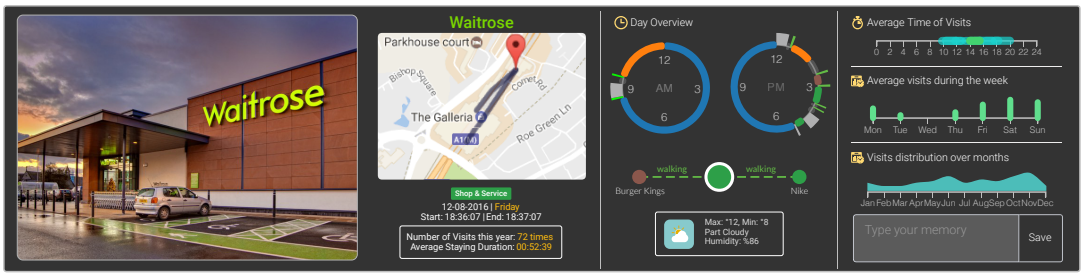


FIGURE 7.23: A complete overview of a selected place (a supermarket) in the MyMoment panel.

In addition, statistical information is also made available, including the number of occurrences and average duration of the event, the histogram of the occurring time (i.e. hour of the day), the weekly occurrence, and the annual occurrence.

There is also a text box for users to enter narration about the event in free text. The notes can be stored together with the event information by the save button during the process of creating digital mementos. This functionality supports users in the process of memory recall and reminiscence by reading their own text related to the event.

Memento storyline and memento list: This is a modal window that displays all the saved mementos in two different forms, namely, either a grid-based or storyline-based style – see R6. Within the grid-based style, the saved event mementos are displayed similarly to card information and arranged in a grid. Each memento carries essential information such as photos, narration, time, and location but not statistics (Figure 7.24). The grid-based style allows users to download these card-style mementos and share them with their friends and family. In the storyline-based style, these event mementos are placed along an interactive timeline. The storyline is designed in a compact form that supports zooming, panning, and mouse hovering for more contextual information (Figure 7.25).

Memento are displayed linearly in the interactive storyline. A combination of circles and texts with different sizes and colours is employed to represent the event mementos' importance scores and categories. In addition, hovering a mouse on each memento discloses complete information such as photos, narration, geographical

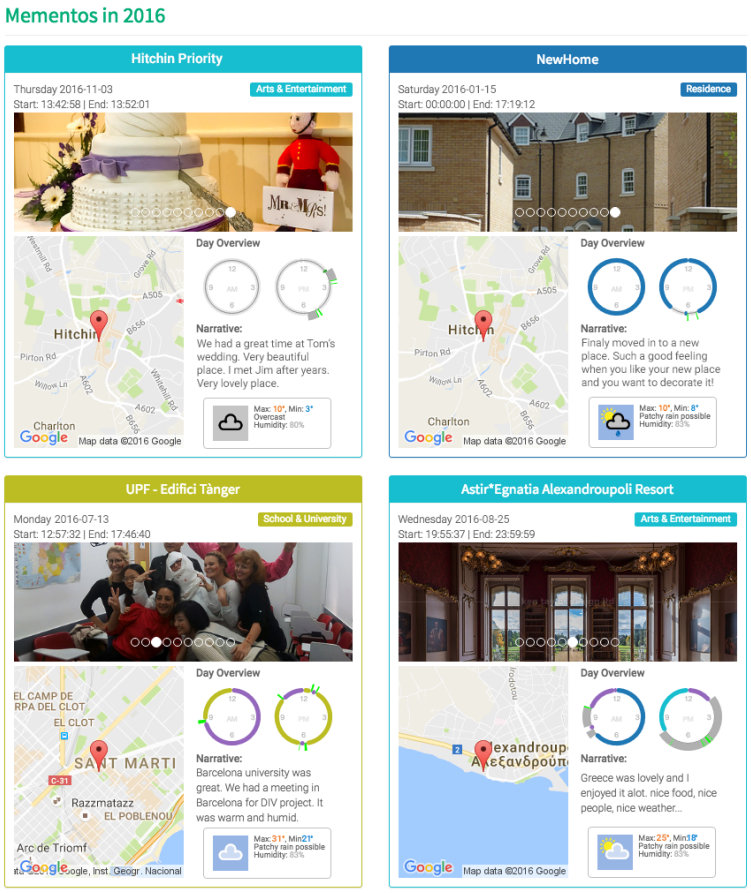


FIGURE 7.24: All the saved mementos are shown in a grid-based layout. The user can download them to share with friends and family.

map, daily view, and statistics. See Chapter 6 (Section 6.3.1) for more details about the storyline structure.

Memento storyline and memento list are only shown on the top layer of the viewport (visualisation page), followed by fading the rest of the visualisation components upon user selection so that the display of the saved mementos normally does not overlap with the main user interface. This also boosts the user focus, particularly on the created memento during the process of reminiscence.

Event category legend: The event category legend located at the bottom of the EventLine indicates

- the categories of the events in different colours,

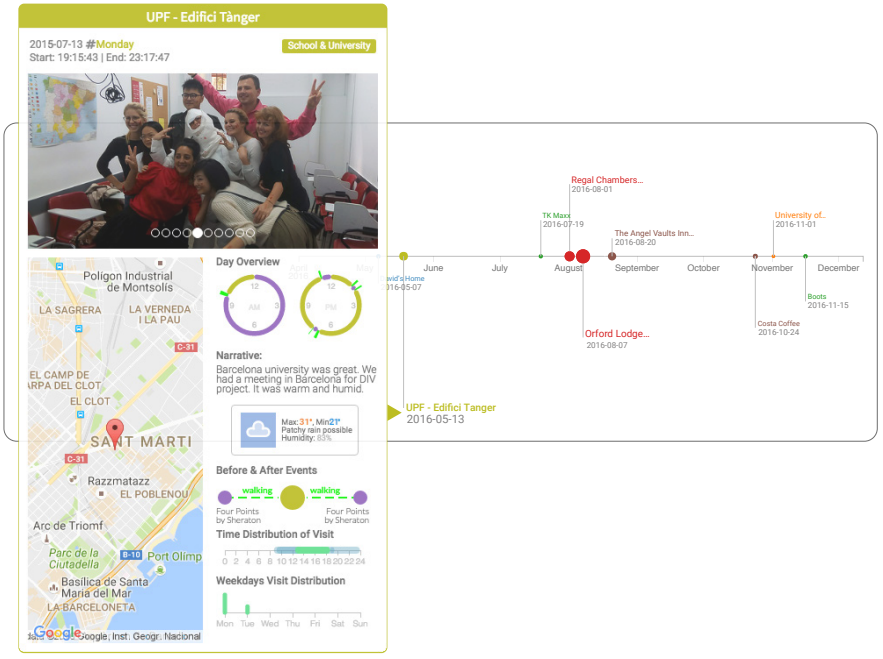


FIGURE 7.25: Interactive Memento Storyline shows all the saved mementos along the timeline, allowing for interaction to present extra details in a compact form.

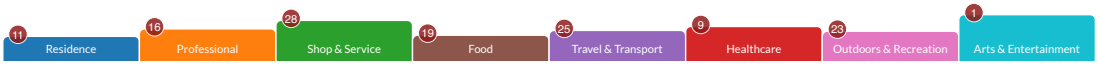


FIGURE 7.26: Event category legend shows the classification of the categories within the data, the influence factor of each category, and the events included.

- the importance of each category using different heights, and
- a list of top-ranked places within each category.

As discussed in Chapter 6 (section 6.3.1), each category has a distinct colour, and also different heights to help users differentiate the classification of events and provide information based on the calculated influence factors within the selected time period, respectively. Figure 7.26 shows a real-world example of how the legend displays the information.

7.3.7 Interaction

With the user interface and its components described in the previous Section 7.3.6, MyEvents offers a range of user interactions for data exploration such as overview,

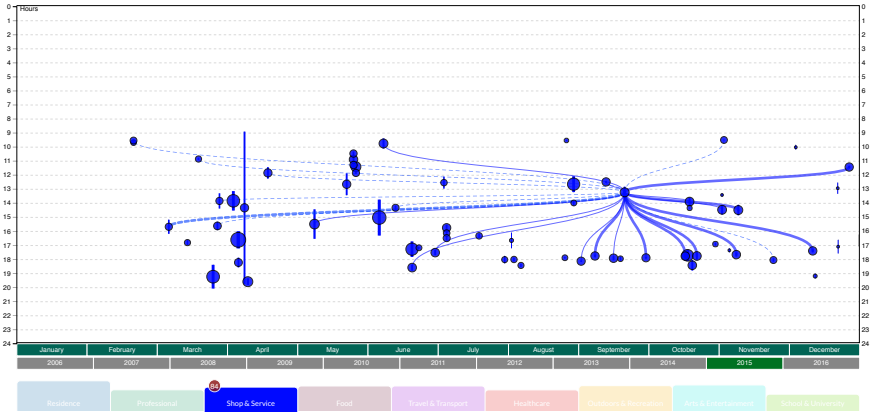
query, data upload, create and save mementos. Each of these interactions yields assorted functionalities within MyEvents to assist users grasp better understanding of their own personal data and facilitate the process of reminiscence for which MyEvents is designed and implemented. Although the interaction of each visual component is described in Chapter 6, it is necessary to illustrate how MyEvents incorporates interaction in its design to deliver an effective method and achieve its goals.

Gain overview of personal events: By default, MyEvents visualises the personal events in the current year with indications of their significance ranking in EventLine, allowing the user to gain an overview of all the activities in the year – R1. Through the multiscale timeline, users are able to select different years and explore the data.

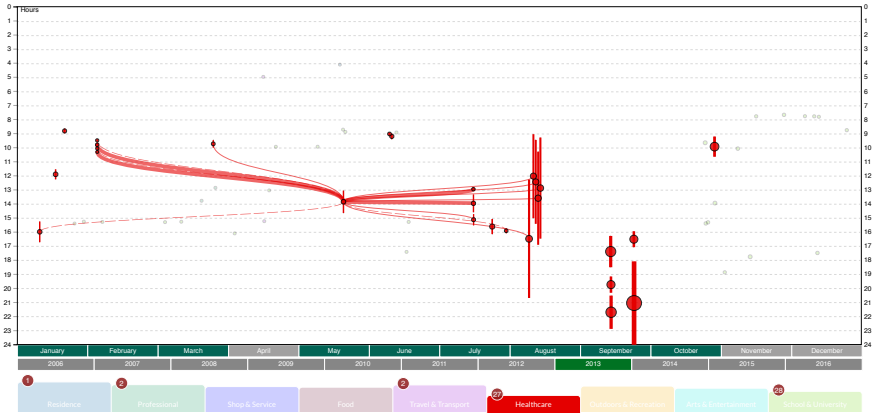
Event query and exploration: The event query allows users to search for events in EventLine and in the map based on the customisable multi-significance event ranking. This functionality can be evoked by typing a set of keywords for named entities (e.g. event name, category, time, regular or irregular, frequent or infrequent) in the search box R2, or by the control panel R3. MyEvents then finds the relevant events and the mRank ranking model computes the significance before they are visualised in EventLine and maps R4. The users can also see and select top-ranked places within each category in the event category legend.

Interacting with event circles in EventLine can filter some of the information or provide additional details. The correlation between the events in the same category is one of the features that can be illustrated by hovering the cursor over an event to view the connections – these can be, for example, repeated events, same place at different location (e.g. shop chains), and nearby places within the same category. The width of the connection indicates the geographical distance between the places (the thicker, the closer), while the same places at different locations or the nearby places are joined with a dashed line – see Figure 7.27.

Data upload: Data upload is straightforward. Users are required to link their *Moves* or *SmarTrack* mobile app to MyEvents – R5. This process needs to be



(a) The correlation between the events within the Shop & Services category.



(b) The correlation amongst the health-related events.

FIGURE 7.27: The correlation between events within the same category. Similar events are connected by a solid line whereas nearby events or similar events in different locations are shown via a dashed line. The widths of the lines are related to the geographical distance between two events, of which the closer has thick linkage, whilst the distant ones have thin linkage

done only once, and the data upload becomes automatic when the device is linked to the system.

Memento viewing, saving, download, and retrieval: Users can explore events from EventLine and obtain further knowledge by clicking on the event circles to see detailed information in MyMoment – R4. This can initiate and boost the process of memory recall. The provided input text box in MyMoment is designed to be used for free-text note writing about the event that can support memory recall. The event mementos can also be saved from MyMoment – see R6

– and be retrieved and shown by using the memento storyline or the memento list – again see R6.

7.3.8 The mRank Event Ranking Model

MyEvents uses the multi-level significance ranking model – hereinafter mRank – to calculate the significance of personal events. A brief overview of this model is portrayed in this section and more details can be found in Chapter 5. The data first undergoes preprocessing including data cleaning, data filtering, event extraction, and event validation before passing to the mRank model. The data cleaning removes errors and noise, whilst data filtering removes unnecessary entities (e.g. logs) from the raw data. In addition, the event extraction extracts the candidate events based on the predefined requirements for identifying an event and then verifies them.

The model considers three factors in event ranking, the event category, frequency, and regularity. Through the user interface and interaction described in the previous section 7.3.7, users can customise these three factors based on their personal preferences. The importance of these factors is expressed in three weighting coefficients in the model:

- w_3: for event category
- w_2: for event frequency
- w_1: for event regularity.

Correspondingly, the ranking score consists of the following three components:

- Event $(tf - idf)$
- Regularity score $(\hat{R}_e)$
- Category score (influence_{ctg}).

All of these components are included in the final equation to calculate the multi-significance event score, which can be used to support the process of memory recall in MyEvents.

$$S(e|d) = (1 - w_1) \times \Big(\text{tf-idf}_e \times \text{influence}_{ctg}\Big) + w_1 \times \hat{R}_e \tag{7.3}$$

where $S(e|d)$ is the final score of the event, tf-idf_e is the term frequency of the event, influence_{ctg} is the influence factor of the category that the event belongs to, and $\hat{R}_e$ is the regularity of the event according to the calculated variance σ_e^2.

7.3.9 Evaluation

In this section, the evaluation of MyEvents, in all aspects, is depicted. The evaluation was organised in two phases, iterative design evaluation and conclusive evaluation. The iterative evaluation was conducted by using simulated personal data to assess the interface, the effectiveness of visualisation techniques, and corroborate the user requirements, respectively, while the conclusive evaluation was aimed at usability and verifying the effectiveness of the approach in supporting memory recall and reminiscence by involving participants with real personal data.

The effectiveness of visualisation techniques is related to how the visualisation can enable users to read, explore, understand, and interpret the visual encoding easily and accurately. This definition can be extended to that the effective visualisation can be perceived to a higher standard by users, consequently interpreted faster, and leads to less error during interpretation, and as a result to more sensible conclusions.

Iterative evaluation was exercised in different stages of the design process. The process included an evaluation of interface design and general effectiveness of the visualisation techniques (visual efficacy) introduced in Chapter 6 to support the process of reminiscence. At the outset, the initial prototype was designed to work with mock-up data on a web browser with limited access in order to

assess the approach in terms of visualisation together with the provided interface and interactions. The analytic modules – automated annotation and significant ranking – were not the subject of this evaluation as the data were simulated and participants would be able neither to judge the accuracy of the annotation nor rate the significance of the events.

The conclusive evaluation was conducted to ratify how MyEvents supports individuals in the process of memory recall and reminiscence by fulfilling the design goals and requirements. The conclusive evaluation involved the participants with the personal daily life data collected by *Moves* or *SmarTracker* applications.

7.3.9.1 Methodology and procedure

The evaluation of MyEvents is based on the participants responses to a set of designed questionnaires regarding different aspects of this approach. The evaluation focuses on four different aspects within the design process to acquire users' needs and preferences. These aspects – interface, usability, visualisation, and interaction – assisted the design process to improve the initial prototype and thus the effectiveness of the approach. For the final evaluation of MyEvents, a boundless task-based assessment followed by a usability questionnaire were used to examine the approach in depth.

In the iterative design evaluation, participants were asked to use the online prototype while some observations took place to form an opinion of MyEvents and complete the online questionnaire. The questionnaire was designed as part of the user centred design approach to enhance the interface and optimise MyEvents functionality.

The conclusive (final) evaluation was carried out by a task-based usability survey and a set of individual interviews. The participants were asked to explore, gain insight, and complete a number of typical reminiscence tasks on the platform. All the tasks reflected the design goals and requirements discussed in Section 7.3.4. The participants' progress during the evaluation was not recorded by any analytical

tools such as Google Analytics to assure participants about the confidentiality of their own data. Instead, participants were asked to think aloud while a simple observation was conducted. This helped in understanding how participants deal with the approach, its interface, and provided visual components to complete the tasks.

A short online tutorial about MyEvents was provided prior to each and every survey based on the version of the prototype to familiarise participants with the approach by introducing each part, indicating the improvements compared to the previous stage, how to use each component, and the expected results. The users were provided with the online consent form at the beginning of each round of evaluation and they were free to withdraw at any time.

In each stage, questionnaires together with observations were carefully studied and a number of interviews took place to obtain more details in regard to certain parts, such as interface, visual components, or interaction within the bound of the MyEvents approach. Subsequently, all the inputs from observations, surveys, and interviews were analysed and further improvements were carried out.

Capturing data In the first run of the iterative evaluation, the questionnaires were divided into three parts, namely, user demand, preferences, and effectiveness to obtain complementary information about the MyEvents elements. In addition, extra attention was paid to SmartSerach box and its functionality to learn how participants prefer to query their own personal data by using the search box. The participants were asked to try both search box and graphical interface – control panel – and answer the associated questions. The questions can be found in Appendix A, Table A.1.

In the second round of the iterative evaluation, 5-point Likert-type questions ranging from (1- Strongly disagree) to (5- Strongly agree) were employed to assess and get the measure of design and functionality improvement according to the first round of evaluation. Table A.4 in Appendix A shows the questions of the second MyEvents evaluation.

No.	Questions
Tasks	
T1	Look into the user's healthcare events during 2011 and tell us how many times the user has had healthcare events, for how long, and was it repeated or not?
T2	Identify how many times the user has been to a restaurant or any related activity during 2016, what was the average time of visit, and on which days of the week (Mon-Sun). Can you see any correlation between the events?
T3	How many times has the user been to the hospital or any related healthcare place during 2013? Which days of the week, and during which hours was the hospital visited by the user?
T4	Query the shopping events that occurred in the morning during 2012 and explain which super market was visited most frequently by the user.
T5	Find out what the user normally does before and after going to the school during 2013.
T6	List all the unusual (rare) events in 2015 by using the search box or the control panel and check if they happened regularly or not.
Interface and functionality likert question	
Q1	MyEvent is easy to use and user-friendly.
Q2	Interactive exploration provides additional information.
Q3	Interaction facilitates gaining better understanding of the daily life.
Q4	The visualisation is not cluttered.
Q5	This approach can help reminiscing.
Q6	MyEvent can improving the user lifestyle.
Q7	I enjoyed seeing the events and their relations on MyEvents.
Q8	MyEvents allowed to learn new information about the user.

Table 7.3: The third round of evaluation tasks and questions

The third round of the evaluation comprised a number of qualitative tasks and Likert-type questions. In total, six tasks were designed in relation to the simulated personal data followed by the Likert-type questions to capture participant progress and to assess usability and effectiveness of MyEvents. - see Table 7.3 for the tasks and Likert questions.

The conclusive evaluation was the final round of the evaluation procedure based on the stable version of the enhanced prototype. In the conclusive evaluation, nine tasks were designed to reflect the design goals and requirements in MyEvents

by involving participants with recorded personal daily data for more than two years. The first six tasks reflected the first goal **(G1)**, whilst the rest covered the second goal **(G2)**. In addition, a Likert-type scale questionnaire ranging from 1 (Strongly disagree) to 5 (Strongly agree) was used to collect user responses in connection with usability.

The purpose of the tasks was to gather the participants' subjective opinions on different parts of MyEvents derived from the design goals and requirements, including: gaining an overview; event query and retrieval via search box and control panel; event exploration and visualisation; data upload, creating, saving, downloading and retrieving digital mementos; and the effectiveness of contextual information, photos and statistical information within the reminiscence context. Participants were required to query, explore, and interact with the visual interface to get meaningful insight (for reminiscence purposes), and at the end explain their experience. They were, subsequently, asked to complete a number of tasks, each of which was designed to measure the usability, performance, and quality of the computed events and visualised information. The tasks can be found in Table 7.4.

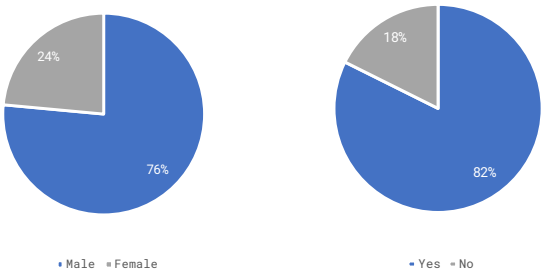
7.3.9.2 Results

Participants Two groups of participants including academic, research fellows, and students were recruited for the evaluation of MyEvents. These groups were solely created to distinguish the participants with and without recorded personal data for the evaluation purpose. The first group incorporated all the participants without any limitation in terms of owning personal life logging data, whilst the second group included only the participants from group 1 with personal life logging data. Group 1 took part in the iterative design evaluations and group 2 participated in the conclusive evaluation. A profile of the two groups and their involvement is depicted in the following.

In total, 5 students and 12 research fellows with British and international backgrounds were recruited (group 1) but only 3 students and 7 research fellows own a personal life logging device (group 2). Students who participated in the evaluation

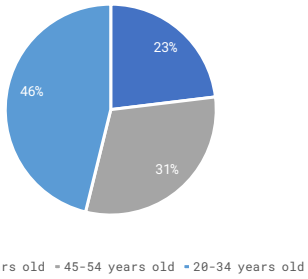
Task	Questions	Expectations
Selection Subjectivity		
T1	Select different period of time and make observations to the overviews of all the events.	Comment on whether you can gain a good overview of your activities within a selected period.
T2	Use the search box to discover an event according to location and category. You may also try the keywords of "frequent, rare, regular or irregular".	Share your experience in retrieving a target event using the search box which features incremental search assisted by the clues provided by the system.
T3	Use the control panel to discover an event according to the event category in conjunction with the three parameters ("view more event", "view more regular events", "view more frequent events") which define the importance of an event according to its name, category, date and time together with preference settings.	Describe your experience in retrieving target events via the control panel.
T4	Interactively view the events which are visualised in the EventLine along a timeline with indications of their locations and time.	Share your experience in the event visualisation in terms of the visibility of important events (namely the performance of the system in terms of highlighting highly ranked events and fading away events with low rankings to avoid over plotting), and the associations between the events. (namely the events at the same place, and sequential events occurred one after another).
T5	Link your Moves app to enable data upload.	How easily you could link your Moves app and upload your data.
T6	Save and retrieve a memento (you may use the memento list or interactive MementoLine).	Describe how easily you could explore the events, create, save and retrieve a created memento.
Place Familiarity		
T7	Select an event from the EventLine and view the map, date, time, clock views, weather and sequence view in the event panel.	Discuss your experience in recalling the events information according to their locations, date and time and context.
T8	View the pictures of the selected event.	How helpful are these pictures in terms of bringing back the memory of the event.
T9	View the statistical information of the selected event.	Talk about how helpful the information such as average time of visits, average visits during the week, distribution of the visits over months for reminiscence.

TABLE 7.4: The final evaluation tasks and expectations



(a) The participants' gender.

(b) The majority of the participants owned some sort of tracking device.

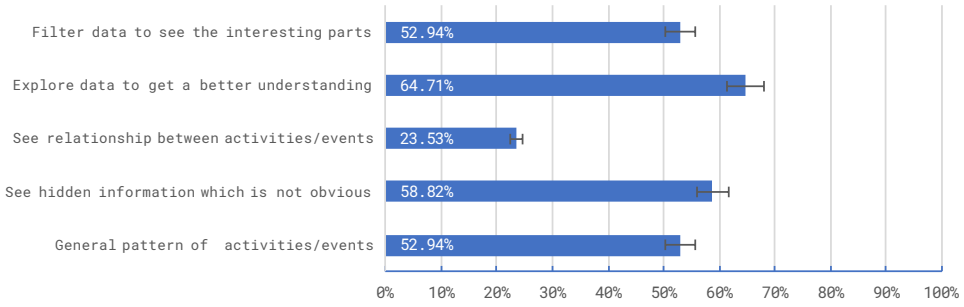


(c) Three age ranges were involved in the evaluation. The majority of the group (46%) were 20–34 years old.

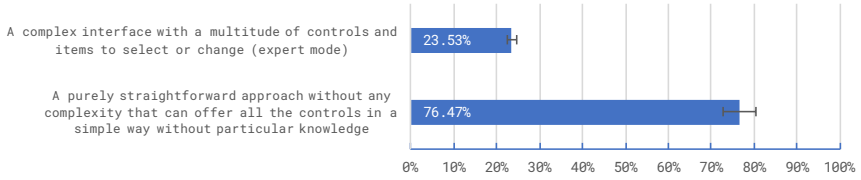
FIGURE 7.28: Demographics of participants.

held a mixture of BSc and MSc degrees, and came from from the Computer Science and Business departments, whilst research fellows were members of the Computer Graphics institute at the University of Bedfordshire. The students and research fellows were not considered as experts in the visual analytics field and, therefore, could participate in the MyEvents evaluation with the aim of supporting reminiscence for non-expert individuals. The demographics of the participants are depicted in Figure 7.28.

Iterative design evaluation The results of the iterative evaluation are shown in this section. Questions are put into three sets in order to examine the user demands, usability, and functionality of the initial MyEvents prototype.



(a) Participants' expectation of what they would like to see about their daily data within MyEvents.



(b) Non-expert users' preferences in a visual analytics approach – MyEvents.

FIGURE 7.29: User requirements analysis in MyEvents.

The results of analysing the user requirements are illustrated in Figure 7.29. According to the analysis, the majority of the participants with 76.47% were in need of a non-complex approach that could perform an intelligible exploration and provide a comprehensive visualisation of the personal life data that can support reminiscence. Moreover, the analysis of user demands shows that participants, to a great extent, were interested in exploring the data to get better understanding and discover hidden information by learning a general pattern in their activities along with filtering trivial information.

Subsequently, extra attention was paid to the filtering and information retrieval in MyEvents. The SmartSearch box is introduced as an intuitive and non-complex means of filtering the data in contrast to the traditional control panel. The SmartSearch box and the control panel were separately compared in two different user interfaces, A and B, to determine users' outlook on both tools (an A/B test). The results of this comparison are presented in Figure 7.30. Notwithstanding that the SmartSearch box received a highly positive response, the need for graphical interface control – with a clear form – was mentioned by more than 88.24% of the participants.

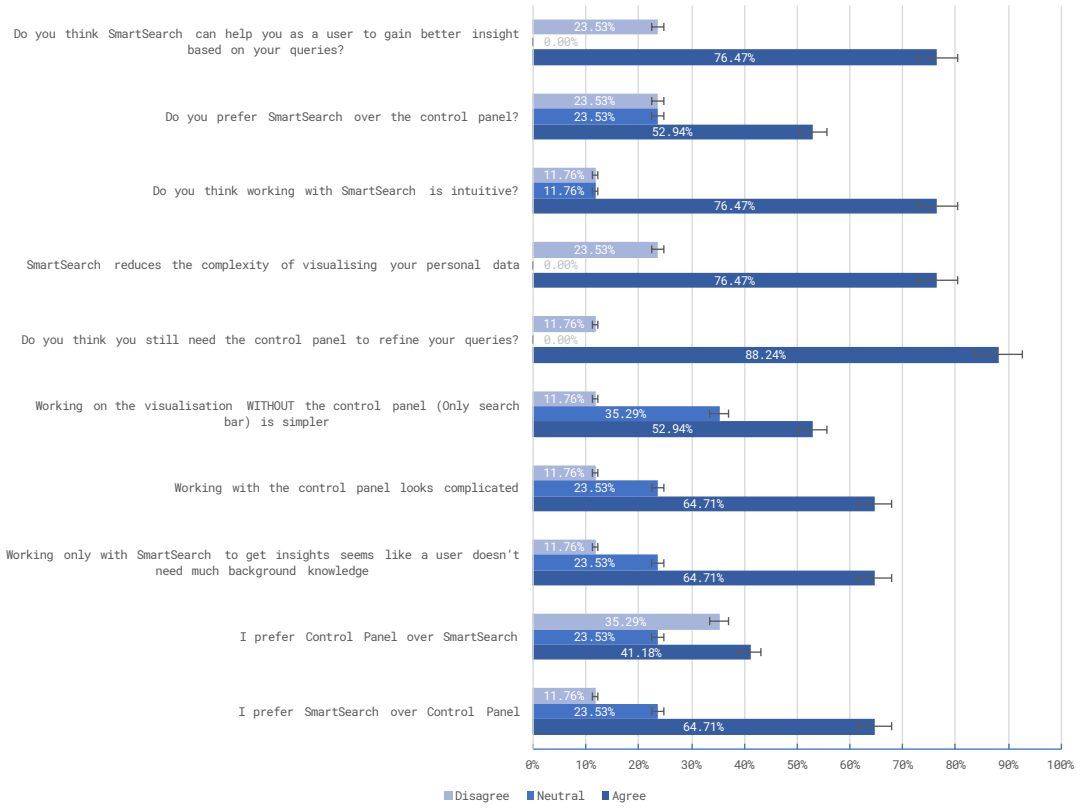


FIGURE 7.30: An individual analysis of the search box and control panel in MyEvents.

As part of the iterative design, the visual components were examined in terms of soundness and functionality as a whole within the research procedure. The result of the first round analysis of the designed visualisation and interaction is shown in Figure 7.31.

A number of improvements were made to MyEvents in connection with the analysis results from the initial evaluation. The second round of evaluation was conducted – by considering 14 participants with valid response to all parts of the evaluation (scale from Strongly disagree to Strongly agree) – to assess the approach and its improvements compared to the previous version of MyEvents. The result (Figure 7.32) showed that MyEvents provided a high level of usability with 71.43% positive responses by the participants. However, 66.67% of the participants gave a positive rate to the clarity of the event representation due to the known issues (which was uncovered in the previous evaluation) in presenting the overall events as well as informative tooltips. To address the issue, the suggested encoding

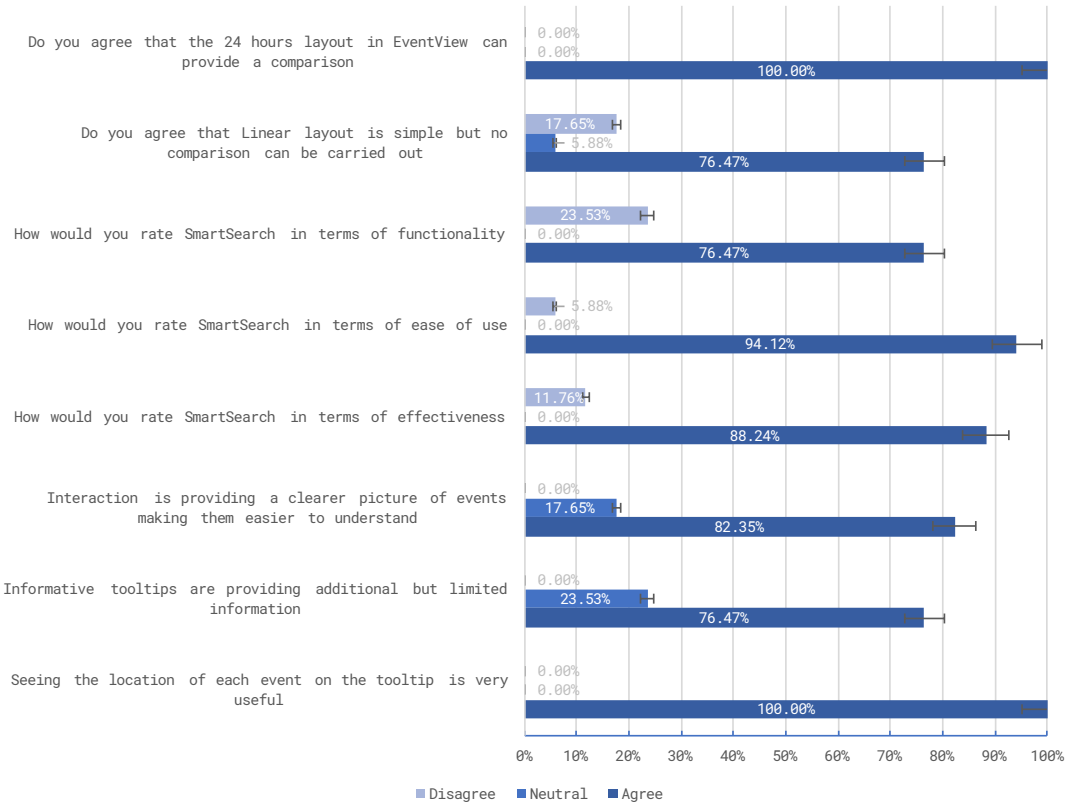


FIGURE 7.31: Overall result of visualisation analysis including the functionality and effectiveness within the iterative design evaluation.

and functionalities in the second evaluation were tested by providing the possible enhancement in the second prototype. Figure 7.32 shows the result for the provided prototype with enhanced or added features. The result indicated that the suggested terms can improve MyEvents' functionality and usability. The rate of positive response to presentation of events and tooltip functionality – based on the suggestions in the second prototype – expanded to 76.19% and 78.36% respectively. While the use of informative tooltips to provide additional information was valued by the participants, a "two-level hover and click" and "a side panel" were introduced to enhance the use of this feature and to accommodate the extra information in a side panel to prevent impeding the EventLine during the exploration. As shown in Figure 7.32, the second approach received a greater rate of acceptance amongst the participants. More detail about the result can be found in Appendix B.

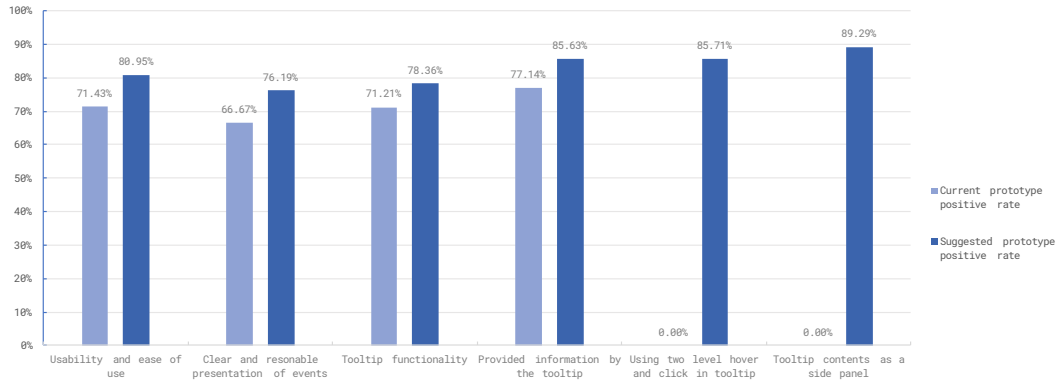


FIGURE 7.32: Positive (agree) results of the MyEvents analysis including the event visualisation and informative tooltip based on the current and suggested functionalities.

The result of the task completion in the iterative design evaluation with the synthetic personal data is illustrated in Figure 7.33. The task completion was conducted by using 10 participants but 9 participants with the valid response were considered for the evaluation. The related tasks[3] can be found in Table 7.3. According to the result, the percentage of correct responses to tasks 1, 2, and 3 is above 70%, whilst tasks 4, 5, and 6 received 55.56%, 66.67%, and 44.44% correct responses, respectively. It has been learned that the tasks related to the use of the search box (**T4**) and control panel (**T6**) were answered with a high level of error due to a lack of sufficient knowledge of operating these components. It has been observed that the users were required to know a set of keywords to start with the search box, and also some initial knowledge to modify the attributes provided within the control panel. The percentage of correct answers to task 5 is acceptable as the participants were able to discover more insight than expected. This led to having a wider range of relevant but not entirely correct answers – See Figure 7.34 for the task completion rate.

The overall analysis of the iterative evaluation is used to enhance the approach during the design process. Moreover, a number of functionalities are added and series of improvements are made to the visual encoding. This process were assisted using the user preferences towards implementing an effective approach.

[3] Some of the tasks included two or more sub-tasks which were assessed equally and the overall mean average were calculated.

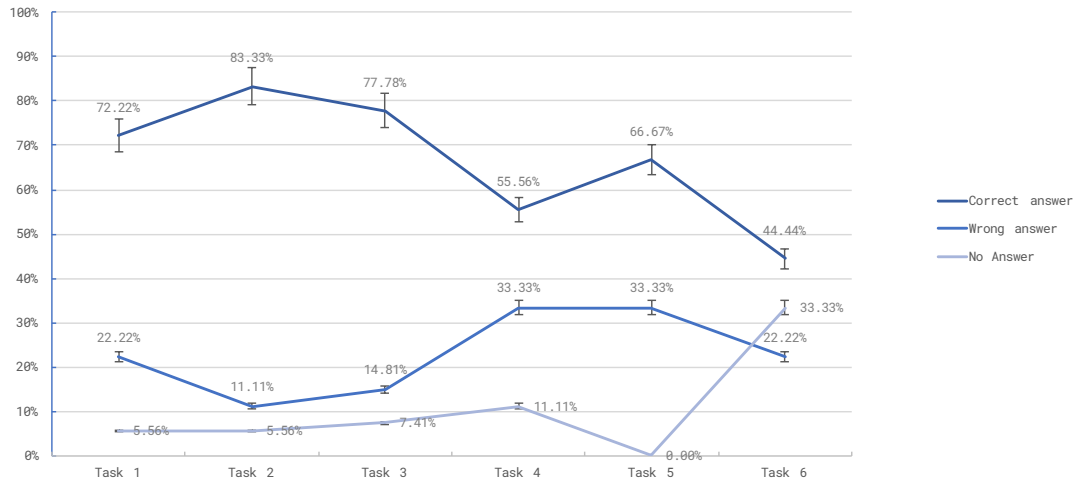


FIGURE 7.33: Iterative design task completion result in detail

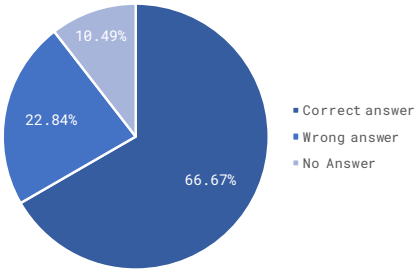


FIGURE 7.34: Overall task completion success rate based on the simulated personal data

Conclusive evaluation At first, the participants were given a brief tutorial of how MyEvents generally supports personal recall and individual reminiscence, and subsequently how it addresses the specific goals of selection subjectivity **(G1)** and event familiarity **(G2)**.

In total, 10 participants aged 28–50 were recruited in this process, with different educational backgrounds. All the participants have used a tracker application (*Moves or SmarTracker*) on their smartphones for a few years to record their daily life activities. The data used in this stage of the evaluation was owned by the participants and divided into individual years to make two factors datasets and increase the number of instances for the evaluation. This allowed this research to

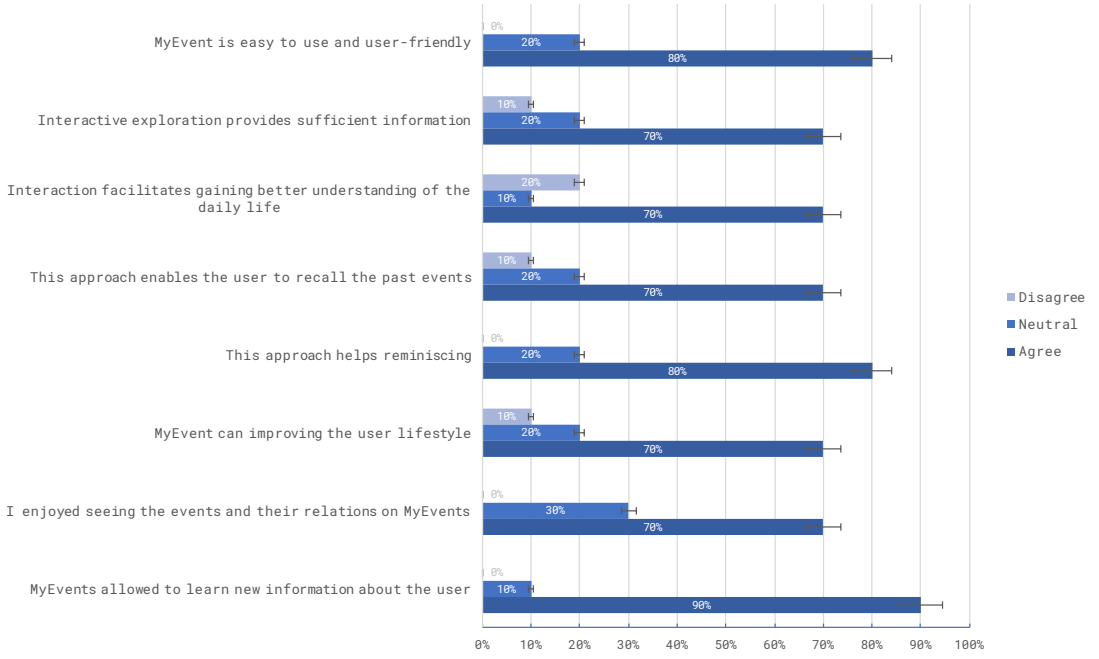


FIGURE 7.35: Result of the usability and functionality questionnaire

address the limited number of participants with real personal data in the final evaluation. For instance, if the user has two years of recorded personal data, the data is divided into two subsets of one year to be examined in two different sessions, accordingly. By doing so, the number of evaluation was expanded to 20 cases with the same number of participants.

The conclusive evaluation consisted of task completion, usability, user interface satisfaction, computer system usability, and effectiveness profile questionnaire. The result of the usability and functionality questionnaire is portrayed in Figure 7.35 based on the tasks defined in Table 7.4.

The result showed that this approach is beneficial in supporting reminiscence by employing novel data mining models and visualisation approach. Eighty percent of the participants stated that MyEvents is easy to use and has a user-friendly interface; 70% of the participants mentioned that the provided information was sufficient for the purpose of reminiscence; whilst 70% mentioned that the interaction facilitated gaining better understanding. Seventy percent indicated that this approach enabled them to recall the past by means of interactive visualisation

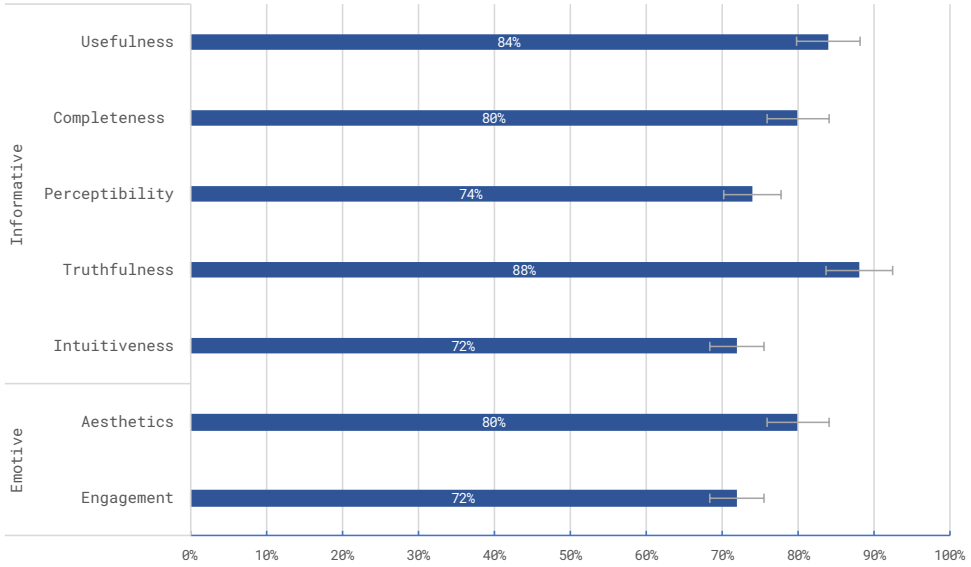


FIGURE 7.36: MyEvents effectiveness profile

and the provided visual components. Overall feedback showed that 80% of the participants mentioned that MyEvents helped the process of reminiscence. In addition, 70% of participants mentioned that MyEvents can also improve one's lifestyle. Seventy percent of the participants stated that working with MyEvents was interesting and they enjoyed exploring the data and creating event mementos. And finally, ninety percent of the participants agreed that MyEvents allowed the discovery of new knowledge about their own daily lifestyles during the exploration.

Furthermore, the result of assessing the effectiveness in Figure 7.36 shows that MyEvents is a useful approach to support memory recall from personal daily life data by providing perceptible and relevant information via a user-friendly interface. The evaluation of the effectiveness is inspired by [75] in accordance with the informative and emotive terms. The participants were asked to rank each terms between (1-5) and the result is transformed to overall score percentage – See the questionnaire in Figure A.1. According to the result, this approach gained overall 80% (SD=0.06) in informative terms, with the highest score in truthfulness (88%) and the lowest score in intuitiveness (72%), and overall 76% (SD=0.05) in emotive terms.

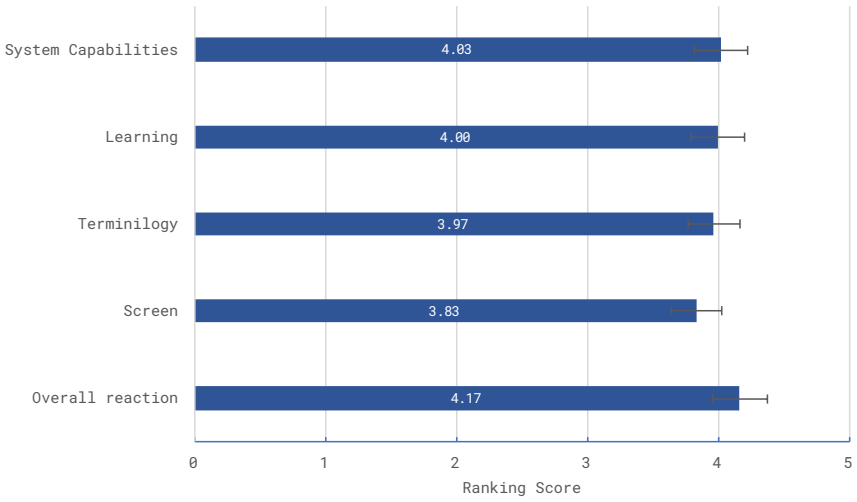


FIGURE 7.37: MyEvents user interface satisfaction

The user interface of MyEvents received a well-grounded feedback (Figure 7.37). The result (consists of the ranking score between 1-5) showed that the overall reaction to the interface, learning, and system capabilities gained averages of 4.17/5, 4.00/5, and 4.03/5 scores, respectively. In addition, the screen and terminology achieved scores of 3.83/5 and 3.97/5, respectively. The most positive aspects of this approach mentioned by the participants were clarity, simplicity, overall design, and interaction between different events, whereas the most negative aspects were the lack of known keywords for the search box and confusion between two types of memento retrieval.

7.3.10 Result Analysis and Discussion

MyEvents offers a distinctive way to explore daily personal location and movement data via visual analytics and supports memory recall by creating event mementos for reminiscence. To identify key events from large personal data, the novel significance event ranking model, mRank, plays a key role in terms of providing significance ranking based on three factors of category, frequency, and regularity, and hence allows for discovery of events according to user preference to facilitate user viewing and creation of mementos. The user interface allows users to set their

preferences, either through the control panel or by using the keywords "regular, irregular, frequent, or rare" in the search box. Such user involvement constitutes the key in selection subjectivity. The evaluation shows that both the search box and control panel have a direct impact on selection subjectivity.

The subjective selectivity is also well supported by the timeline-based visualisation (i.e. EventLine) and other visual components such as the event category legend. The evaluation shows that the visualisation in MyEvents plays a very positive role in terms of facilitating the users in the selection of personal events for reminiscence. The visual components in MyEvents allow users to gain a clear overview of personal history, to easily discover and hence quickly access a set of candidate events of interest. The interactive visualisation enables users to explore more details of the events, including obtaining more information, viewing links between events, and visualising event mementos, hence allowing users to find more pertinent and sensible results. Studying the results reveals that the participants were very interested in using the platform to explicitly explore their own data with the help of the ranking model and the visual components (e.g. EventLine, search box) and to create mementos for reminiscence. They also used MyEvents to gain knowledge about their own daily life (e.g. how often they went to a hospital, how many times they visited a specific place).

The presentation of the mementos was effective in terms of supporting memory recall. Collectively, the contextual information, photos, and statistics provided valuable information and helped the users re-experience the past. The interactive clock view was very helpful to the users, since recalling other relevant events contributed positively to the memory. The memento saving and retrieval functions were also very useful from the user perspective: firstly, they did not need to search for the same events repeatedly; and secondly, the stored events were made available to the users according to their preferred layout; for example, they could access their recently stored mementos immediately through the grid style, or alternatively, they could view all the stored the mementos in an interactive storyline.

7.3.10.1 Recall and individual reminiscence

According to the participants' responses, MyEvents effectively involves the users in the reminiscence process and presents events in a favourable way to evoke event familiarity: "*MyEvents gave me a freedom to select an event based on my interest and showed me a gripping overview. I could select each event, view the related photos, then write my memory about the selected event and save it. It provides a beautiful interactive storyline of my saved memento. I was able to show them to my wife and talk about how pretty were those places. Fascinating!*" Another participant stated that the system helped him see a very rare and special event that he had entirely forgotten which results in recalling the good memory of that: "*This tool showed me an event that I totally forgot amongst all my day to day events. I used provided interaction to get more information about it by looking at the photos, map, clock view, etc. It put smile on my face when I remembered it.*" The positive comments showed that the system, particularly, the ranking method, highlights significant events in personal history very well. On the other hand, it has been identified that there is a need for a complete form of tutorial and a short training session before exploring the personal events via MyEvents. In addition, some visual improvements in the events were mentioned by the participants in order to make the visualisation more engaging: "*I like the layout. Overall this gives me a very good assistance in recalling the event in terms of location and date but I would like to see the time and date of the events more obvious (Bigger). Also, I prefer to see better overview of my events on the map.*"

7.3.10.2 Selection subjectivity (G1)

The participants were asked to assess how well MyEvents helped them gain an overview of their own daily data as one of the design requirements. The result of the selection subjectivity and its requirements can be found in Figure 7.38-(G1)[4]. According to the overall subjective feedback, MyEvents provides an appealing

[4] The result contained the users' subjective responses ranging from (Strongly disagree) to (Strongly agree) and were transferred to the score of maximum 5

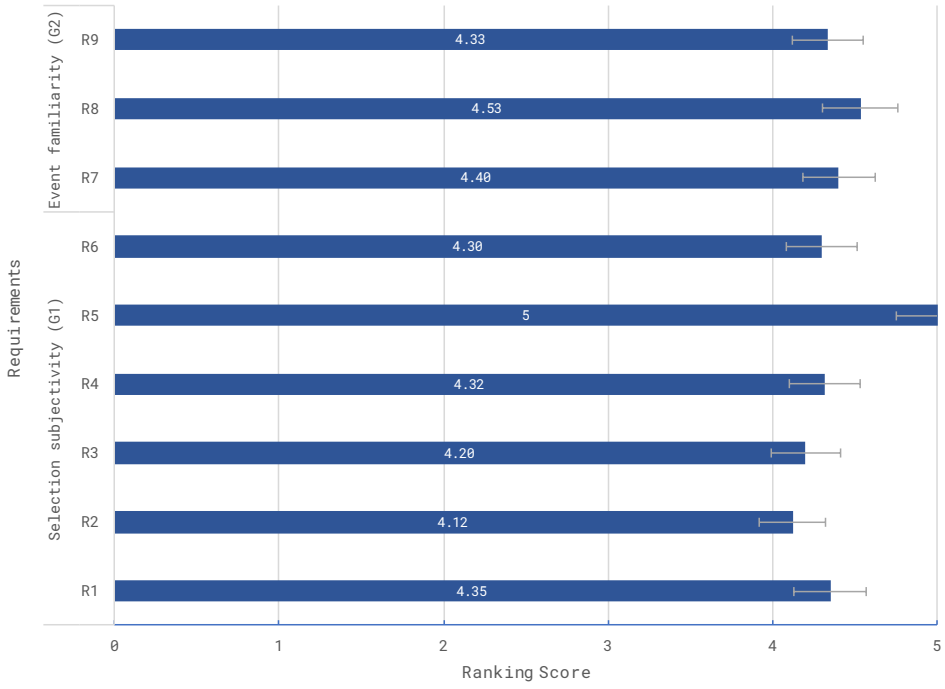


FIGURE 7.38: MyEvents goals and requirement results

overview of daily life events **(R1)** with mean 4.35/5. The participants were satisfied with the overview and in some cases they discovered interesting facts about their life: "*Overall, MyEvents gives me a good overview at a selected time period and support further interaction to retrieve the events*", "*Very interesting, I notice that I went to the hospital a lot more than I thought in 2015.*" Also, participants made some general suggestions: "*I would like to see some abstract information next to the event circles to be able to recall my memory without any further interaction*".

To assess how the search box and control panel help retrieve and discover events **(R2, R3)**, the participants were asked to retrieve their own data via the search box and control panel and rate the functionalities accordingly. The feedback from the participants shows that the search box was very useful (mean 4.12/5) in the search, whilst the control panel was regarded as a more intuitive way of starting the search (mean 4.20/5): "*The search box helped me to find a lot of events that I did not remember by myself. The suggestion are amazingly works for me*", "*I particularly like the auto completion and the incremental search features*", "*Event*

search through control panel is very intuitive, it especially helped me when I just came to use the platform." Meanwhile, the feedback showed that there is a need to provide more hints about available keywords within the search box to help new users to perform their task correctly: "*a bit more guide information could be very helpful, for example, I did not know these keywords such as regular or irregular until somebody told me*". Moreover, it was mentioned that the search box and control panel need to show a feedback message while they are delivering the required process on the fly.

The event visualisation in EventLine aims for a direct influence on the reminiscence process by representing the events along a timeline together with their associations. The visualisation of events (**R4**) achieved an mean of 4.32/5. The comments were mostly positive, such as "*The events are displayed in different size and colours along 24 hours, which helped me easily find what I was looking for. I also like the category legend, in which I could see the significant places I have been to in each category*", and "*I like this design. It shows the event, which can be my potential mementos, along the timeline. It shows me the association between some of my events by hovering the mouse that I never thought about!*". However, in some cases, the participants had a problem with the EventLine interaction that was classified as a bug in the system: "*The visualisation Highlight the event does not work properly, I need to click multiple times to trigger an action*".

MyEvents is also designed to provide an opportunity to save mementos for later retrieval. One of the main focuses in the evaluation was to assess the support of the system towards saving, downloading, and retrieving of digital mementos **(R6)**. The provided functions attained a mean of 4.30/5, as the participants were able to create a number of mementos and later on show them to their family members: "*The layout of the memento is impressive! I think, it is very important to me to come back and see the mementos that I created before*", "*I love the note taking very much in which I can write some memorable notes about my events which really helps me to recall most of my past events!*", and "*I used the interactive memento storyline to see all my mementos along the timeline. I showed our trip to Germany and France to my wife and children. It was nice to see our photos and*

what we did during that time. I will certainly come back to see my mementos over and over again as it is giving me a good feeling!" However, providing two different ways of retrieving the mementos caused some confusion although the layouts were consistent. This issue was addressed in the layout by assigning the grid-based memento list for sharing purposes and the interactive storyline for demonstration purposes.

7.3.10.3 Event familiarity (G2)

Providing additional information to invoke event familiarity is a vital feature in the system – see Figure 7.38–G2. The participants' comments and responses to the Likert-scale questions were analysed to assess how effectively the system can fulfil **(R7)**. The result revealed that MyEvents satisfies the requirements by achieving a mean of of 4.40/5: "*I like the layout and the way the information like map or time are represented. This gives me very good assistance for recalling the event in terms of location and date*", "*I enjoyed the context information and found recalling the event through other activities that happened on the same day (looking at the clock view or sequence view) is very useful*", "*I recalled my London trip in January 2015 with the help of map and clock view. Amazing!*" Although the feedback was positive, a few suggestions regarding the use of a bigger map were pointed out. It appeared that the participants preferred to interact with a larger map in order to see more information about the event location.

To further support invoking event familiarity, the users were allowed to view photos of and statistical information about events **(R8, R9)** as this information, according to [23, 79, 92, 106, 155, 158, 204, 208, 226], can significantly contribute to memory recall. According to the evaluation analysis, using photos broadly supports the reminiscence process (mean 4.53/5): "*I like the pictures very much. I think they are the most important items that help me to recall the old memories.*" Moreover, using the statistical information is regarded as a useful feature for discovering knowledge about events **(R9)** (with mean 4.33/5), which can create a more enjoyable experience: "*The statistics features are quite interesting although*

they are not essential. They helped me see how often I have been to a specific place and on which days I normally went", and "*The statistical information of my selected events are fascinating! Specially for regular and frequent events. Something that I haven't seen in any tracking devices dashboard.*"

7.3.10.4 Limitations and future work

The evaluation and analysis also helped with comprehending a number of limitations of the current approaches in MyEvents and hence, future work can further explore areas such as:

- Keywords hints – It has been learned that users needed further assistance in entering named entities using the correct keywords in the search box. More hints need to be provided to show available keywords in the search box.
- Online sharing of MyMoment – The current version of the platform allows users to share their information by downloading. Further work may link the platform to other social media platforms like Facebook and Twitter, hence allowing users to share their event mementos online.
- Events vs. activities – The current work mixes the concept of event and activity by assuming the same activity at one place. This limitation can be addressed upon the availability of more activity data within one place.
- Other data – It has been learned that people are also interested in visualising more personal data captured in their daily life alongside the location and movement data: for example, health-related data. Correspondingly, the ranking model may need to consider more factors to calculate the significance ranking.
- There is also a potential that, upon the collection and availability of longer and more extensive personal history data, MyEvents can be used to present the lifespan of an individual by showing all the key events in the person's life.

7.4 Chapter Summary

This chapter presented three integrated visual analytics tools, ActivityTimeline, LifeTracker, and MyEvents designed for non-expert users. Each of these tools is designed and implemented as part of the research objectives to show how and how effectively the proposed visual analytics approach can proceed towards different purposes for supporting personal daily life and gaining meaningful knowledge.

ActivityTimeline demonstrated the simple but powerful means of interactive visualisation to support the visualisation of lifestyle based on the level of activity and movements over time via a basic form of control panel and four interactive visual components – Activity Stack, 24-hour activity, Activity Cloud, and Activity Bubbles. This platform was evaluated in term of accuracy, functionality, efficiency, usability, and reliability. The results show that ActivityTimeline is capable of visualising the lifestyle and its activity levels. A number of drawbacks were also identified during the evaluation and these were used as lessons learned in LifeTracker and MyEvents.

LifeTracker presented a visual analytics approach to explore daily life data and extract life patterns from individuals' life logging data. LifeTracker benefits from a simple user interface with an integrated timeline, life pattern view, event list, and map. LifeTracker provides an overview, and detailed life patterns on the 24-hour grid-based timeline on the dynamic categories. It used a time matrix to analyse the data and fit them to the associated hours and categories. LifeTracker utilised a number of visual components from Chapter 7 to encode the information into comprehensible knowledge by displaying the dominant activity during the specific hour of a day while storing all the other activities in matrices. Although the evaluation of LifeTracker shows that it allows users to explore the life logging data and gain valuable knowledge with acceptable accuracy, there are many ways in which LifeTracker can be extended.

MyEvents adopted a personal visual analytics approach to query, rank, and visualise personal events to support personal reminiscence. A novel multi-level

automated place annotation and a novel multi-significance event ranking model were employed to automatically enrich the raw trajectory data and take into account event category, frequency, and regularity, allowing for user preferences through either the graphical user interface (control panel) or using keywords (search box). The evaluation result showed that MyEvents is highly beneficial in supporting reminiscence and helping individuals gain meaningful knowledge about themselves. MyEvents fulfiled two main goals in reminiscence. The first goal addresses selection subjectivity, which enables users to identify the events of interest by modifying the ranking preferences, making event queries, gaining overviews of personal history, and exploring further details about the events through interactive event visualisation along a multi-scale timeline. The second goal addresses event familiarity, which evokes memory familiarity by presenting the user's selected event with key contextual information, photos, and statistical information. Users can save and subsequently retrieve and download the mementos. MyEvents were evaluated in terms of effectiveness, usability, and user interface by involving a number of participants who have self-tracking location and movement data for a few years. The evaluation result shows that MyEvents is beneficial to support reminiscence, enable individuals to gain knowledge about themselves from location and movement data, and present a sound approach to address an interesting research issue in personal visual analytics.

CHAPTER 8

Conclusions and Future Work

This thesis has set out a novel visual analytics approach to gain meaningful knowledge from accumulative personal life logging data. This approach has used two inventive data mining models to enrich and rank the included data points, in particular, events amongst large-scale day-to-day personal data. Furthermore, this approach uses a carefully designed visualisation to visualise the extracted knowledge and facilitate the process of knowledge discovery through the set of visual components. Each part of the approach was assessed individually during the design and implementation process to ensure the proposed methods are accurate, effective, and viable. The proposed approach was exploited by three integrated visual analytics tools for different purposes to demonstrate its effectiveness and versatility. The evaluation confirmed that this approach facilitates the process of knowledge discovery effectively and can be adapted for different purposes in the personal domain.

This chapter concludes this work by reviewing the objectives, highlighting the contributions, revisiting the research questions, and outlining future research within this context.

8.1 Thesis Summary and Contributions

This work focused on systematic knowledge discovery by utilising advanced data mining models, data visualisation, and interaction, rather than the traditional data visualisation with its lack of scalability and expansion towards gaining meaningful insight from personal data.

The literature review in Chapter 2 reviewed the related visual analytics work in three main areas – data and knowledge mining, data visualisation, and evaluation – with particular focus on personal data, time-oriented data, and spatiotemporal data (**OBJ1**). Although existing techniques are important for mining and encoding data, certain gaps lie in semantic enrichment, significance ranking, and data visualisation. This work took advantage of good practice to initiate its new data mining models and visualisation design accordingly.

Proposing a new approach requires efficient and rigorous evaluation of its components and use cases to ensure that the approach meets corresponding goals and requirements. Chapter 2 investigated the practical evaluation scope within the visual analytics context that allows the evaluation to be conducted at different phases (e.g. precondition, design, prototyping, and the like) within the process (**OBJ2**). In addition, the number of methods that are actively used for evaluating the visual analytics approach were identified, such as evaluating visual data analysis and reasoning, evaluating user experience, and evaluating visualisation algorithms.

Modelling Automated Place Annotation with multi-level probabilistic latent semantic analysis was extensively presented in Chapter 4 . This model fills the gap of effective and accurate multi-level semantic enrichment discussed in Chapter 2 with scalability in mind (**OBJ3**). The automated place annotation is explicitly modelled to attach multi-level semantic information into the unidentified trajectory data by determining pertinent POIs and computing the probability value using incremental probabilistic latent semantic analysis with contributing historical location data and prior knowledge. This model is evaluated by using

six datasets with known ground truth at different levels and achieved overall of 72% accuracy in the first level, 86% in the second level, and 100% within the 7th level of the annotation. Furthermore, the benchmark results for the efficiency and scalability of the implemented algorithms show that this model runs with $O(n \log n)$ complexity time (**OBJ4**).

Chapter 5 presented the multi-significance event ranking mode that comprises a set of fixed and customisable factors as coefficients to extract significance events. The model is formed of three different components, namely $tf - idf_e$, event regularity, and category influence factor with consideration of three external factors w_1, w_2, w_3 to reflect user preferences in selecting the importance of regularity, frequency, and category. This model is assessed to determine the accuracy of the outcome and the performance of implemented algorithms accordingly. The evaluation showed that the model can identify the significant events with 72.43% (SD=0.07) accuracy for the real-world data based upon user preferences. Moreover, the benchmark result showed the logarithmic $O(\log n)$ time complexity for the significance ranking algorithms and $O(n)$ for the event extraction in which it scales well as the number of data points expands (**OBJ5**).

The data visualisation design including the visual encodings and visual components was demonstrated in Chapter 6. A comprehensive list of key requirements that indicates in what way the visualisation needs to accommodate users' needs were established in this chapter (**OBJ6**). This ensured that the visualisation technique effectively visualises the extracted knowledge towards gaining meaningful understanding of different aspects of daily life with minimal level of learning through its visual components and uncomplicated interface.

Furthermore, Chapter 6 presented the process of design, implementation, and evaluation the advanced technique that endeavours to strike a sound balance between the level of automation and user control, allowing users to effectively seek for desired information based on their own preferences via interaction and the support of the underlying data mining techniques (**OBJ7**). The visualisation technique comprises perceptual visual encodings (e.g. shape, size, colour, glyphs),

high quality visual components, and an optimised user interface with respect to daily life. The number of visual components such as multi-layer timeline, storyline, circular-based layout, 24-hour event visualisation, and smart legend are introduced to form innovative visualisation solutions within this context. These components are developed in the light of preattentive processing and formulated to be versatile, which allows for different types of knowledge discovery and visualisation. The visual components that form the data visualisation were iteratively examined to ensure the quality and perceptibility of the encoded information (**OBJ8**).

This thesis has presented a feasibility study based on the three integrated visual analytics tools which were demonstrated in Chapter 7 and show the versatility evaluation of the proposed visual analytics approach including its established novel data mining models and visualisation components within the personal life domain (**OBJ9**). The integrated visual analytics tools are designed with particular attention as follows:

- **ActivityTimeline**: to visualise the level of physical activities – walking, running, cycling, and transport – and support personal lifestyle patterns by means of an interactive time-oriented visualisation. This tool uses a control panel and four interactive visual components namely, Activity stack, 24-hour activity, Activity cloud, and Activity bubbles. The evaluation indicated that ActivityTimeline is capable of visualising the lifestyle despite a number of drawbacks, which were taken as lessons learned.

- **LifeTracker**: to interactively explore the personal life logging data at different levels of detail and different time scales by allowing identification of highlighted summaries and key places. LifeTracker consists of timeline, life pattern view, event list, and map that can provide a summary overview, and detailed life patterns over 24 hours based on dynamic categories. The underlying data mining models were employed to address the low or no semantic information and to extract significant events. The evaluation result showed that this tool allows users to interactively explore the life logging data

and gain valuable knowledge regarding the general or specific life patterns with adequate accuracy and efficiency.

- **MyEvents**: adopts a personal visual analytics approach to query, rank, and visualise personal events to support personal reminiscence. This tool employed an automated multi-level place annotation and a novel multi-significance event ranking model to involve user preferences during the data mining process. It fulfils two main goals in reminiscence – selection subjectivity and event familiarity. The former allows for identifying the events of interest by incorporating user preferences and the latter evokes memory familiarity by envisaging the user's selected event with essential contextual information, photos and statistical information. This visual analytics tool was evaluated in an iterative way by involving two groups of participants with and without the self-tracking movement data. The evaluation result showed that MyEvents is highly capable of supporting reminiscence by assisting users to gain meaningful knowledge about themselves from the location and movement data, and presents a sound approach to address the identified gaps in personal visual analytics.

The contributions of this work, in a nutshell, are outlined as follows. An extensive literature review of data and knowledge discovery, data visualisation, and evaluation was conducted with particular focus on personal data to identify the gaps within this context and the best practices for evaluating such an approach (**C1**). Moreover, the novel automated place annotation was designed and developed to enrich the trajectory data with prominent level of semantic information (**C2**). The proposed multi-significance ranking model was also implemented, which allows for user involvements and personal preferences within its settings (**C3**). Furthermore, the interactive information visualisation including inventive visual components was designed, developed, and evaluated accordingly by involving the users in the design process to effectively visualise the extracted knowledge (**C4**). A robust visual analytics pipeline was developed to support the process of knowledge discovery and exploration within the personal life logging data (**C5**). Three integrated visual

analytics were developed and evaluated, namely, ActivityTimeline, LifeTracker, and MyEvents, based on the proposed visual analytics approach including data mining and interactive information visualisation, all of which showed accurate and high quality results according to their purposes (**C6, C7, C8, C9**).

8.2 Reviewing Research Questions

Q1. What is the impact of semantic enrichment in personal spatiotemporal data and how does it contribute to mining meaningful knowledge?

- In general, personal spatiotemporal data lacks adequate semantic information that can be used in extracting knowledge from logged events. Without meaningful contextual information about such data, extracting any knowledge regarding the behaviour or interest of users is beyond the bounds of possibility. However, such data can be tremendously improved by employing an automated semantic enrichment method and can be used by analytical reasoning to extract invaluable information (Chapter 4). Using the semantic enrichment model enables the data mining process to obtain the users' behaviour not only based on the geographical position but also based on the classification of places, rate of occurrence, and regularity.

Q2. How important is significant event ranking to the visual analytics of personal data and what is the impact of effective event ranking in knowledge discovery? Also, what are the most influential factors on the result of event ranking in persuasive knowledge discovery?

- As described in Chapter 5, spatiotemporal data in the personal daily life domain contains a great number of trivial information that are mainly not of interest. Visualising such data without considering the significance yields an ineffective outcome with limited

features and quality. Hence, discovering what information is more of interest is indispensable as it can bring the momentousness to the fore of the approach and amplify the quality of provided knowledge.

Subsequently, the event ranking models should consider a number of factors that contribute in acquiring the significance. These factors should reflect individuals' preferences within the process to provide a sensible result. This thesis introduced a multi-significance event ranking model with three main factors as follows: 1) frequency and uniqueness, 2) regularity, and 3) category influence factor of events. Additionally, three external factors were identified which reflect user preferences as follows: 1) the weight of importance for event regularity, 2) the weight of importance for event frequency, and 3) the weight for event category as a point of acquiring the significance.

Q3. How important is the visual representation of extracted knowledge in personal visual analytics? And how effective does eloquent visualisation support the process of knowledge discovery?

- The visual representation plays a vital role in the human cognitive system to convey the extracted knowledge to users. The visualisation is the connection between the human visual system and computer strengths that can facilitate the process of forming hypotheses and knowledge exploration. An effective visualisation including the well-designed visual components (Chapter 6) strengthens the ability to perceive the data with minimum effort and allows for further exploration by providing sufficient interaction, hence influencing the process of knowledge discovery. This thesis provided a concrete interactive visualisation and user interface in Chapter 7 based on the fundamental aspects of human cognition, which benefits from computer strengths in

analysing data and providing effective visual components in order to reinforce the process of knowledge discovery in this context.

8.3 Future Work

Personal visual analytics has become a focus of much research due to availability of related personal data. However, research on individual daily life still faces fundamental problems such as lack of adequate semantic information and effectiveness. Despite the fact that the proposed visual analytics approach, in this thesis, manifested constructive knowledge discovery via its underlying data mining and visualisation, further research is required to address the limitations and add more merit to the current approach. Potential future work directions are briefly described as follows:

- **Semantic enrichment for episodes in daily life**. The automated place annotation in Chapter 4 can go further to calculate plausible grounds for combination of movements and places within daily life. This requires extensive research on how to form a number of possible hypotheses based on latent probabilistic grounds and accept or reject them in conjunction with user historical behaviour in space and time.

- **Automated place annotation enhancement**. The model demonstrated in Chapter 4 can be improved to yield higher accuracy within the initial suggestions by engaging powerful and scalable machine learning algorithms such as linear regression. Subsequently, this model can be deployed as a standalone enrichment module and plug into different analytical or data mining approaches to address the absence of the contextual information.

- **Additional external factors in the multi-significance event ranking model**. Although the model implemented in Chapter 5 demonstrated a positive result, it can be expanded by considering a number of external factors that might affect individual life in its significance score calculation,

such as emotional and social impact. This requires extensive research on feasible factors that have direct as well as indirect influence on the significance ranking model.

- **Automated sequence ranking within the daily life data**. Research on identifying the significant sequences within the daily life can be investigated by adopting the automated place annotation and multi-significant ranking models, and pertinent algorithms such as Hidden Markov Model (HMM). The model requires identifying the sequences, calculating the rank according to the similar factors in Chapter 5, and creating a textual purpose to them by adding meaningful semantic information. This results in determining a purpose for each sequence and can highly contribute to the process of knowledge discovery in this domain.

- **MyEvents to support memory recall by providing multi-level information**. In Chapter 7, MyEvents as an integrated visual analytics tool was introduced in order to investigate the effectiveness of the proposed visual analytics approach in supporting reminiscence. However, due to a limitation of this research, it did not attempt to measure the level of memory recall. This can be a future research direction to uncover the correlation between different levels of information and memory recall towards supporting memory impairment. The current information is limited to photos, narrations, geographical location, and finite textual content. Future research can extend this information to video, music, sound, social media, and news feed along with the current facts.

APPENDIX A

Supplementary Materials – Questionnaires

No.	Questions	Type
User Demand		
Q1	What would you like to see about your daily activities/events in general?	Multiple answer
Q2	What kind of approach would you prefer to work with and explore your data?	Multiple choice
Q3	How important is it to have the ability to customise visual encoding by yourself in order to understand the visualisation better?	Likert-based
Q4	Do you think SmartSearch can help you as a user to gain better insight based on your queries?	Likert-based
Q5	Do you prefer SmartSearch over the control panel?	Likert-based
Q6	Do you think working with SmartSearch is intuitive?	Likert-based
Q7	SmartSearch reduces the complexity of visualising your personal data	Likert-based
Q8	Do you think you still need the control panel to refine your queries?	Likert-based
Q9	Would you like to use a text-based search box or a control panel with several facets to make the query?	Multiple choice
Q10	How important is the auto completion and suggestion feature in the search box?	Scale

Continued on next page...

TABLE A.1 – continued from previous page

No.	Questions	Type
Q11	How would you rate seeing a list of suggestions based on the text that you already entered into the system?	Scale
Q12	How would you rate the ability to search for multiple locations that took place over multiple time periods?	Scale
Q13	How would you like to enter the text in search box (SmartSearch) in an arbitrary order or a fixed order?	Multiple choice
Q14	How would you rate selecting the date and time information? By pre-defined time range e.g. morning, afternoon, evening or start and end?	Likert-based
Q15	SmartSearch can narrow down your interests based on place, time, category and frequency. How would you rate SmartSearch in the following terms (functionality, ease of use, effectiveness)	Scale
Q16	What information would you like to add to a query in SmartSearch? (e.g. time, location, category, etc.)	Multiple answer
Human-computer Interaction		
Q17	How appealing is the interaction of a whole platform including the interface and visualisations to you?	Scale
Q18	Interaction is providing a clearer picture of events making them easier to understand	Likert-based
Q19	Informative tooltips are providing additional but limited information	Likert-based
Q20	Seeing the location of each event on the tooltip is very useful	Likert-based
Q21	It is not necessary to grey out the rest of the events	Likert-based
Usability and functionality		
Q22	How do you rate the ranking approach that determines the importance of events based on their frequency, degree of interest, and the user preferences?	Scale/Rank
Q23	Rate how easy you can see the pattern, understand the visualisation, and finding the start, end, and overall pattern of the given example.	Scale/Rank
Q24	The categories colours are distinguishable in the DARK background than the WHITE background	Scale/Rank
Q25	The DARK visualisation background makes the visualisation easier to read than the WHITE visualisation background	Scale/Rank
Q26	Which colour scheme for categories is the most distinct based on what you see on the above figures?	multiple choice

Continued on next page...

TABLE A.1 – continued from previous page

No.	Questions	Type
Q27	We provide SmartSearch to fully eliminate the traditional control panel and enhance usability. Based on the description, and the two given layouts for MyEvent (One of which contains SmartSearch, map, and visualisation canvas, and the other comes with map, control panel, and visualisation canvas), which one is more appealing to you?	Multiple choice
Q28	Working on the visualisation WITHOUT the control panel (Only search bar) is simpler?	Likert-based
Q29	Working with the control panel looks complicated?	Likert-based
Q30	Working only with SmartSearch to get insights seems like a user does not need much background knowledge	Likert-based
Q31	I prefer Control Panel over SmartSearch	Likert-based
Q32	I prefer SmartSearch over Control Panel	Likert-based
Q33	How would you prefer to see where your events/activities took place, on the provided map or on the tooltip?	multiple answers
Q34	According to the description regarding the Degree of Interest term and the given methods (Method 1 with important events and dimmed out the rest of events, and method 2 with only important events), which method do you think can be more informative without losing a track of visualisation?	Multiple choice
Q35	Amongst the two provided layouts for visualising events, one of which with a 24 hour grid style (the Y-Axis as a 24 hour scale and the X-Axis as days) and the other a linear timeline, which type of timeline layout is more appealing to you?	Multiple choice
Q36	Do you agree that the 24 hours grid layout can provide a comparison?	Likert-based
Q37	Do you agree that Linear layout is simple but no comparison can be carried out?	Likert-based
Q38	Based on the provided timelines in MyEvents, multi layered and traditional linear timeline, which timeline do you prefer?	Multiple answer

TABLE A.1: The user demand questions in the iterative evaluation first run

No.	Questions
About Shopping	
Q1	How often do you do the shopping during the week? (e.g. 2 times, 3 times, etc.)
Q2	Do you do the shopping in the morning / afternoon / evening?
Q3	Which days do you normally do the shopping? (e.g. Mon, Tue, etc.)
About Food related actions	
Q1	How often do you go to any food related place such as restaurant, coffee shop, etc.?
Q2	When do you go to the food related place? (Morning/afternoon, evening)
Q3	Which days do you go to any related places? (e.g. Mon, Tue, etc.)
About Outdoor & Recreation activities	
Q1	How often are you exercising or going to park or similar places?
Q2	When are you going? (Morning/afternoon, evening)
Q3	Which days do you go for any outdoor and recreation activity? (e.g. Mon, Tue, etc.)
About Arts & Entertainment	
Q1	How often are you going for any arts or entertainment related activity?
Q2	When are you going for these kind of activities? (Morning/afternoon, evening)
Q3	Which days do you normally select for entertainment? (e.g. Mon, Tue, etc.)
About your Profession	
Q1	What time during the day are you working? (Morning/afternoon, evening)
Q2	Are you working day shifts or night shifts?
Q3	Are you working within the weekdays or weekend?

TABLE A.2: The user profile questionnaire

ID	Questions
Q1	The event with 24hr grid layout can provide a comparison
Q2	Linear event visualisation is simple but no comparison can be carried out
Q3	The design of the event visualisation with 24hr layout is compelling
Q4	Using degree of interest and dimming out trivial events is appealing
Q5	The pattern can be perceived with not much effort
Q6	The visualisation is comprehensible
Q7	The interaction is sufficient and effective
Q8	The importance of the event can be perceived easily by using the circles with different size of radius
Q9	Visualising the event by using a combination of line and circle (line as a duration and circle with different radius as an importance) is appealing
Q10	The correlation between the events are beneficial and appealing
Q11	Highlighting the event by dimming out the rest of the events, using the transitional animation, and adding a addition to the event is effective
Q12	The interaction is sufficient and effective

TABLE A.3: The evaluation questionnaire for 24-hour event visualisation

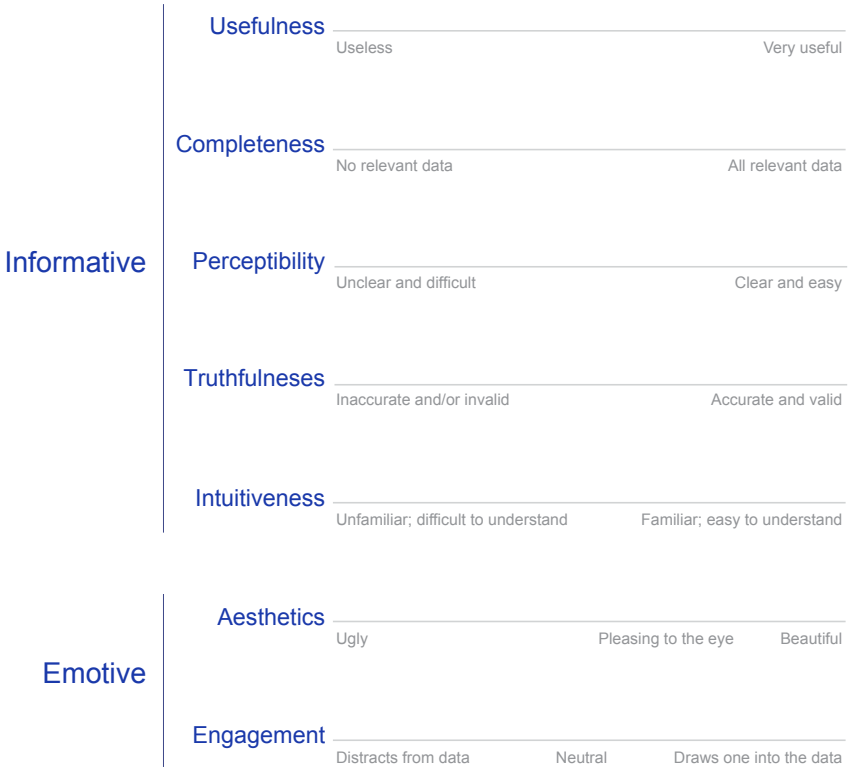


FIGURE A.1: Effectiveness profile by Few [75]

No.	Questions	Type
Interface and Functionality		
Q1	MyEvent needs the control panel to adjust some of parameters that cannot be done by the search box	Likert-type
Q2	Interactive exploration provides more details	
Q3	Interaction facilitates gaining knowledge	
Q4	The visualisation is comprehensible and user-friendly	
Q5	The interface is easy to understand	
Q6	The visualisation is not cluttered	
Q7	This approach can help the user to reminisce about events	
Q8	This technique can help to improve the user's lifestyle	
Event View		
Q9	The events appear reasonable and clear on Event View	Likert-type
Q10	Interaction provides sufficient support for exploring the events	
Q11	Correlation between the events can be triggered without any complexity	
Q12	Seeing the duration of the events is important	
Q13	The visualisation does not require a high level of computing skill to use	
Q14	Category colours are sufficiently distinct from each other	
Q15	The correlation between events (if any) stands out in the visualisation	
Tooltip interface		
Q16	Two-level hover and click allows the user to control the amount of displayed information	Likert-type
Q17	Tooltip provides sufficient facts about the selected event	
Q18	By using the tooltip, the user can discover interesting points about the selected and similar events	
Q19	Seeing the weather history information is compelling	
Q20	Seeing what led up to the event and what happened after is helpful	
Q21	Average time and distribution offers engaging information	
Q22	Seeing the frequency of the event during the week is compelling	
Q23	The tooltip is providing excessive amount of information which is not necessary	
Q24	How do you rank the evolution of the tooltip?	
Geographical information		
Q25	It would be helpful to display the tooltip information as a side panel at all times	Likert-type
Q26	A side panel could show additional information in order to compare the current and previous events	
Q27	The side panel prevents the Event view being impeded by the tooltip	
Q28	Compact form of the provided information is still effective on the side panel	
Presentations of events on a radial layout		
Q29	Which one the following presentations of events on a radial layout is more compelling to you?	Multiple-choice
Q30	The glyphs make the visualisation easier to understand	Likert-type
Q31	There is a need for a glyphs legend	
Q32	The glyphs make the visualisation cluttered if there are many activities and places	
Q33	Glyphs do not have any influence on the radial layout	
Q34	Text info is preferable to glyphs	

Table A.4: The second round of iterative evaluation questions

OVERALL REACTION TO THE SOFTWARE			
Q1		terrible	wonderful
Q2		difficult	easy
Q3		frustrating	satisfying
Q4		inadequate power	adequate power
Q5		dull	stimulating
Q6		rigid	flexible
SCREEN			
Q7	Reading characters on the screen	hard	easy
Q8	Highlighting simplifies task	not at all	very much
Q9	Organization of information	confusing	very clear
Q10	Sequence of screens	confusing	very clear
TERMINOLOGY AND SYSTEM INFORMATION			
Q11	Use of terms throughout system	inconsistent	consistent
Q12	Terminology related to task	never	always
Q13	Position of messages on screen	inconsistent	consistent
Q14	Prompts for input	confusing	clear
Q15	Computer informs about its progress	never	always
Q16	Error messages	unhelpful	helpful
LEARNING			
Q17	Learning to operate the system	difficult	easy
Q18	Exploring new features by trial and error	difficult	easy
Q19	Remembering names and use of commands	difficult	easy
Q20	Performing tasks is straightforward	never	always
Q21	Help messages on the screen	unhelpful	helpful
Q22	Supplemental reference materials	confusing	clear
SYSTEM CAPABILITIES			
Q23	System speed	too slow	fast enough
Q24	System reliability	unreliable	reliable
Q25	System tends to be	noisy	quiet
Q26	Correcting your mistakes	difficult	easy
Q27	Designed for all levels of users	never	always

Table A.5: Questionnaire for User Interface Satisfaction by Chin et al. [47]

APPENDIX B

Supplementary Materials – Evaluation Results

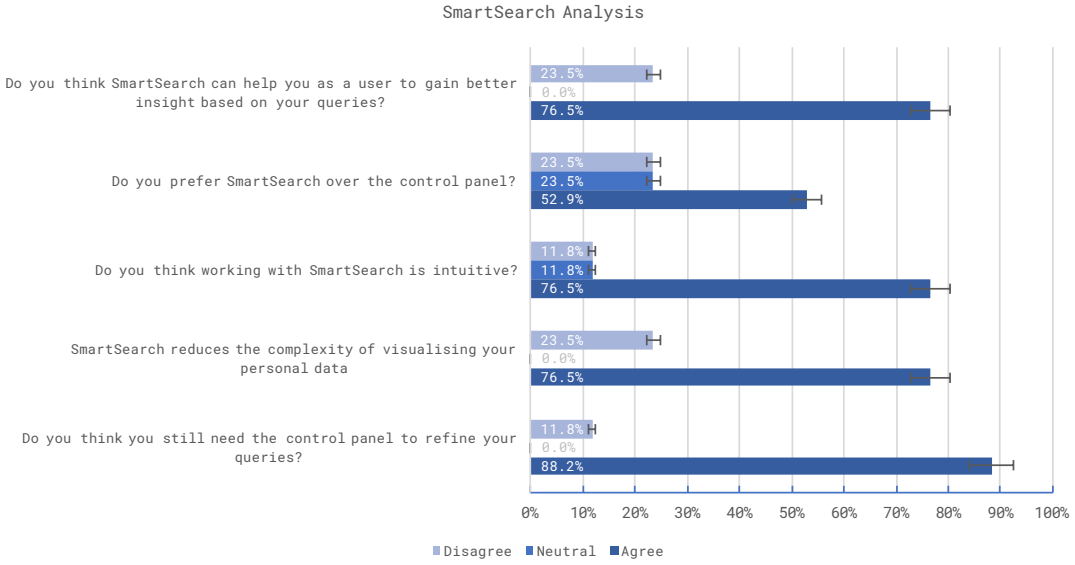


FIGURE B.1: SmartSearch design evaluation

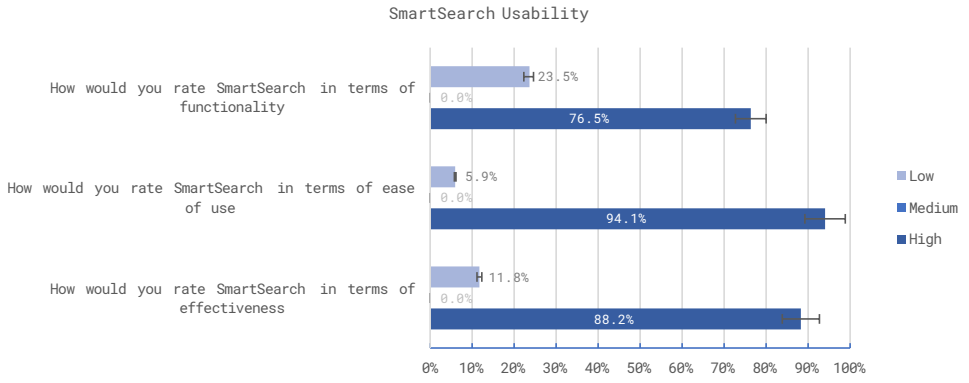


FIGURE B.2: SmartSearch rating result

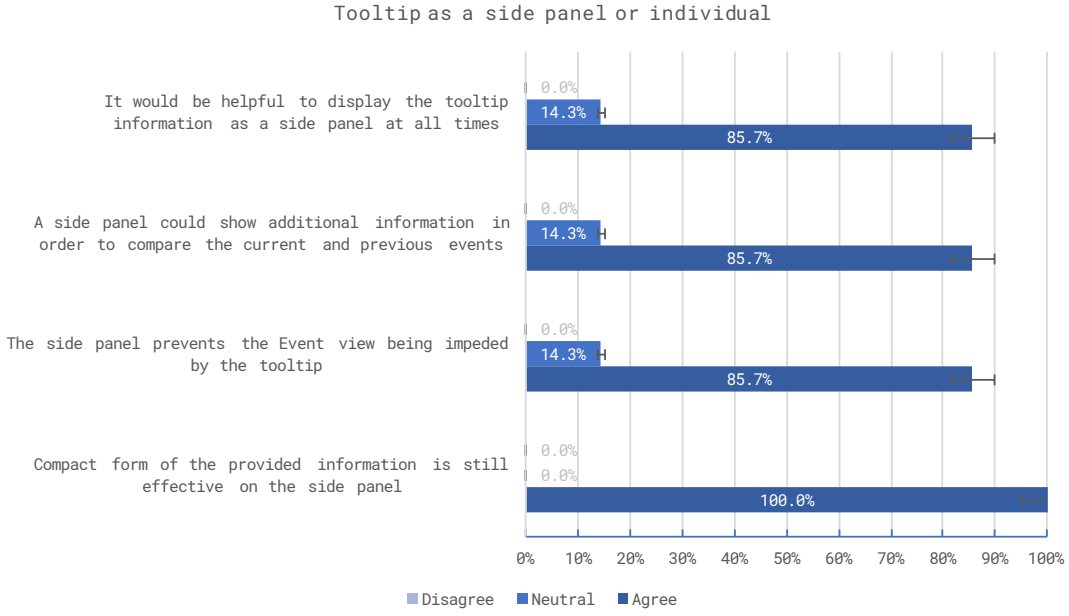


FIGURE B.3: Result of the tooltip as a side panel evaluation

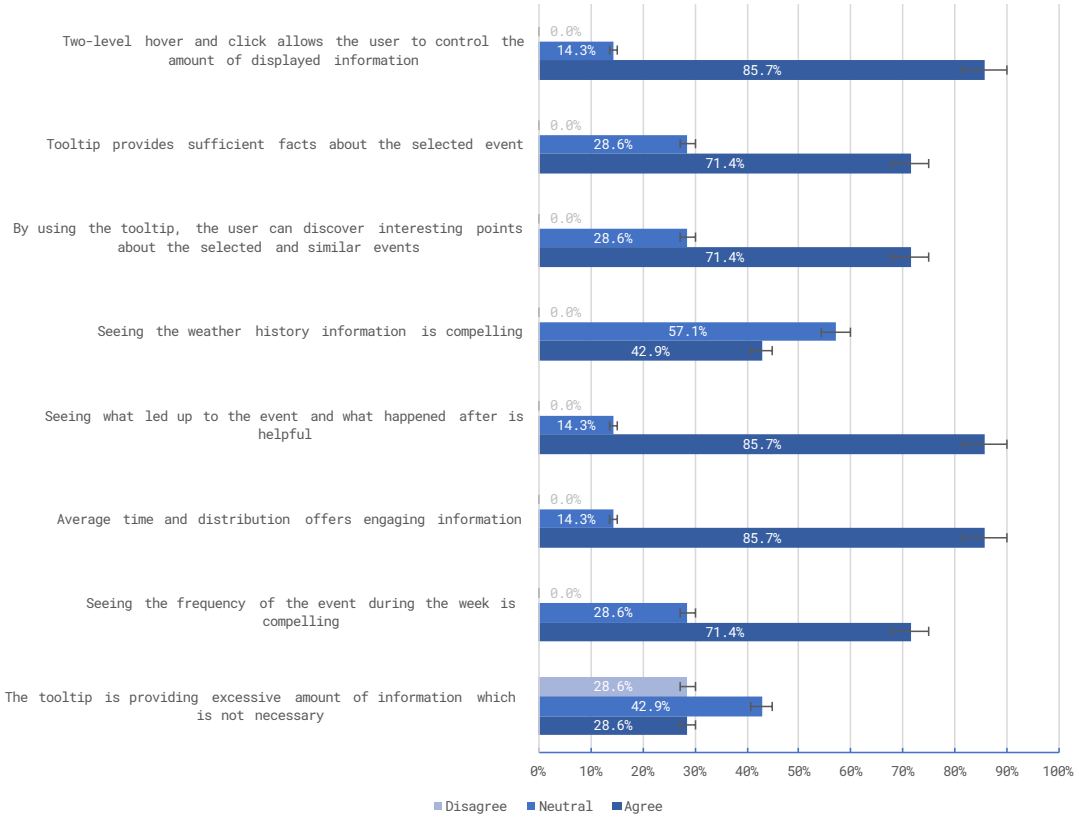


FIGURE B.4: Tooltip and its provided information evaluation

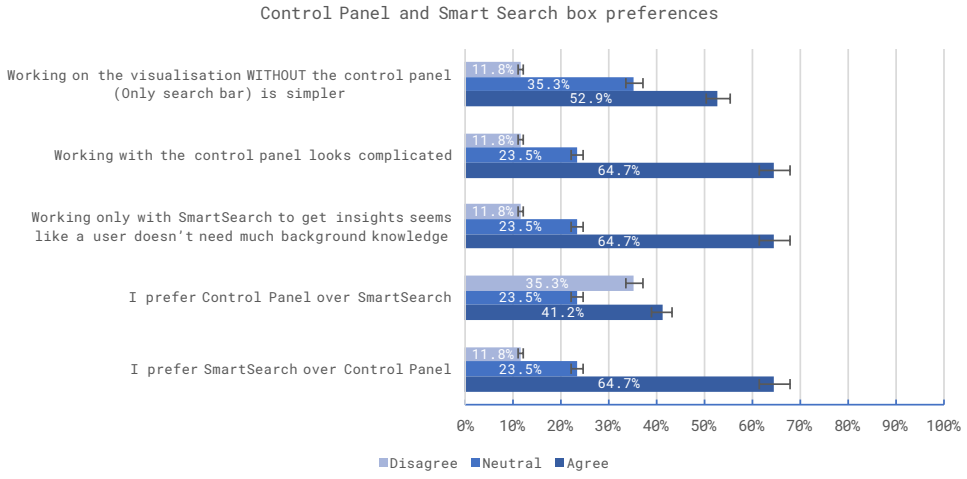


FIGURE B.5: Control panel and SmartSearch comparison

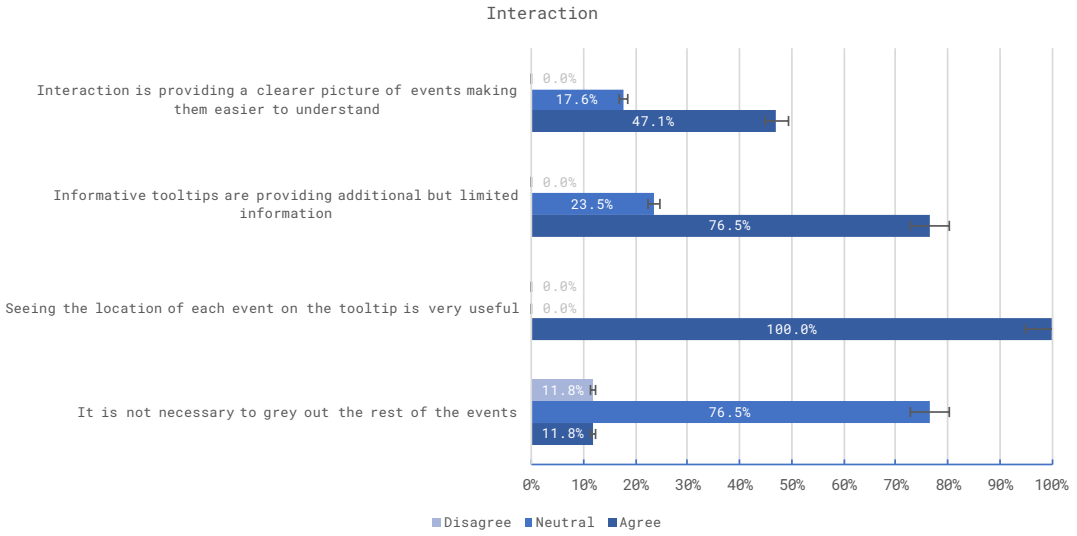


FIGURE B.6: Interaction evaluation for visualising events

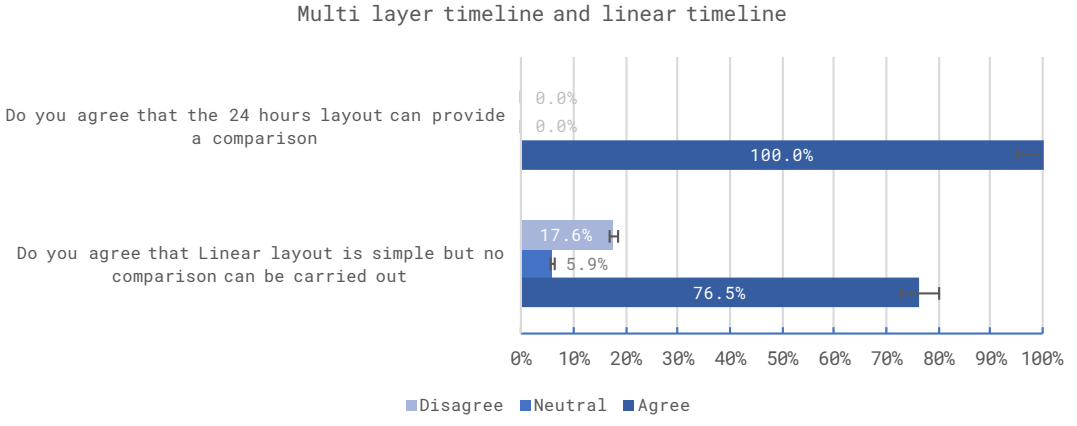


FIGURE B.7: Multi-layer timeline and linear timeline comparison

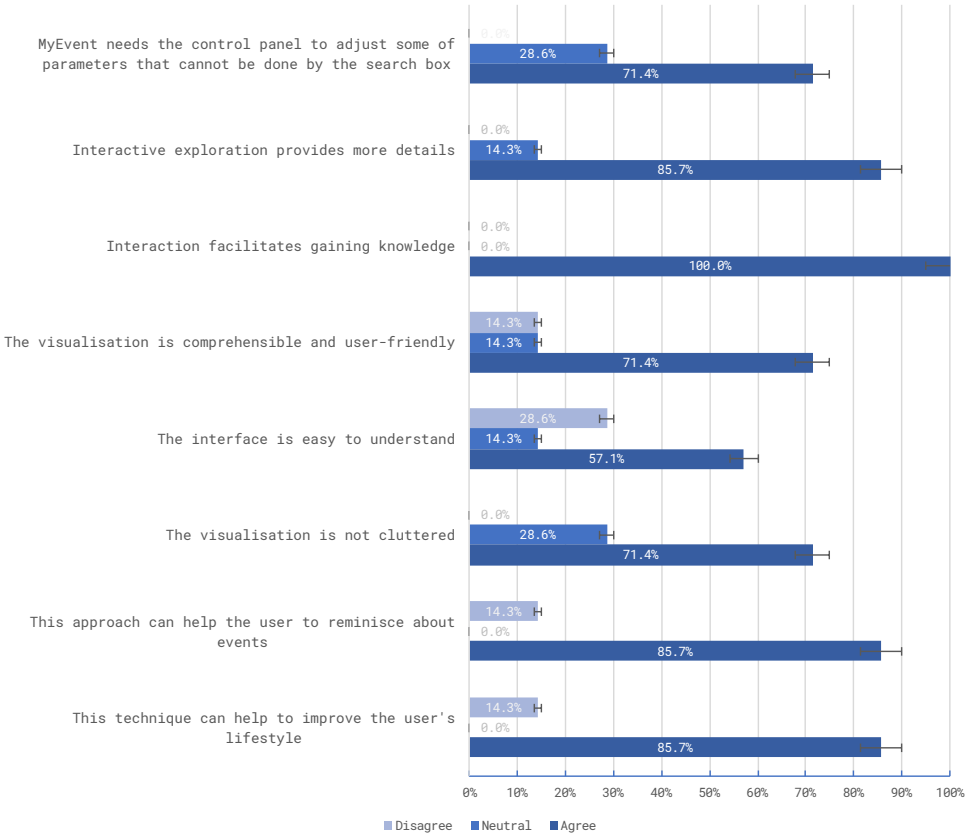


FIGURE B.8: MyEvents second round of interface and functionality evaluation result

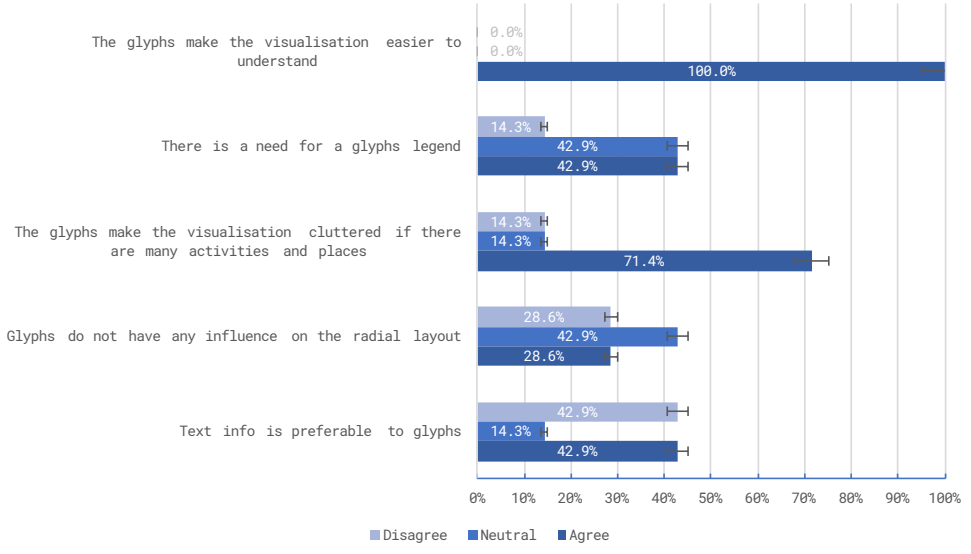


FIGURE B.9: Evaluation result of using glyphs within the circular layout in MyEvents

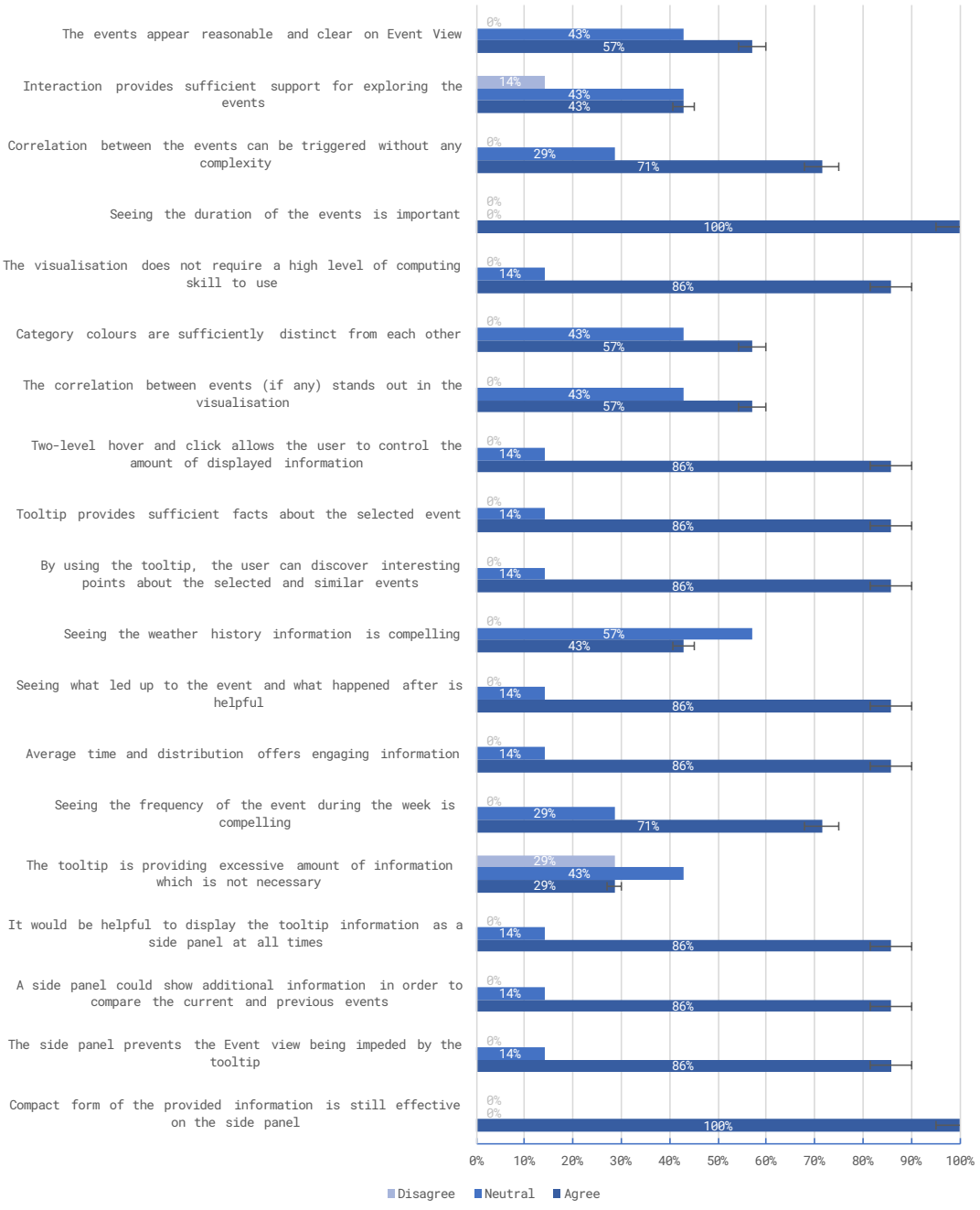


FIGURE B.10: Evaluation result of MyEvents components and functionalities in second round

APPENDIX C

Pseudo Codes

C.1 Automated place annotation pseudo code

```
var isNodeJS = typeof window === 'undefined';
if (isNodeJS) {
    var fs = require('fs');
    var async = require('async');
    var moment = require('moment');
    // var d3 = require('d3-array');
    var d3 = Object.assign({}, require("d3-array"), require("d3-scale"));
    var math = require('mathjs');
    var haversine = require('./js/haversine.js');
    // console.log(haversine);
    var request = require('request');

    var top = {};
    var nf = function () {};
    // mute jQuery operation in nodeJS environment
    var $ = function () {
        return {
            progressbar: nf,
            append: nf,
            hide: nf,
            attr: nf, // -> progressbar
            show: nf,
            ajax: nf
        }
    }
}

function access_foursquare(api_uri, callback) {

    if (isNodeJS) {
        request({
            method: 'GET',
            uri: api_uri
        }, function (error, response, body) {
            if (!error && response.statusCode === 200) {
                // console.log(JSON.parse(body));
                callback(JSON.parse(body))
            } else {
                console.log(response);
            }
        });
    } else if (typeof $ !== 'undefined') {
        $.ajax({
            crossDomain: true, // added in jQuery 1.5
```

```
            headers: {
                'Access-Control-Allow-Origin': '*'
            },
            async: false,
            type: 'GET',
            dataType: 'jsonp',
            // async: false,
            url: api_uri
            // success: crosSucessHandler(result, status)
        }).done(callback);
    }
}

// DBSCAN CLUSTERING ---->
function handleGPS(data, eps, minpoint) {
    var gps_process = []
    var gps_geo = []
    var downloadData = []
    var bias = 1.5; // multiply stdev with this factor, the smaller the more clusters
    var dbscanner = jDBSCAN().eps(eps).minPts(minpoint).distance('HAVERSINE').data(data)
    var dbscan_res = dbscanner()
    var cluster_cent = dbscanner.getClusters()

    top.clusterOfdata = cluster_cent
    cluster_cent.sort(function (a, b) {
        return d3.descending(a.dimension, b.dimension)
    });
    var tmp_zeros = []
    for (j in tmp_zeros) {
        cluster_cent.push(tmp_zeros[j])
    }
    return cluster_cent
}

// Create categories list according to Foursquare
data_handling('cats.json', function (json) {
    top.ctg = json
    var tmpCat = []

    for (i in json) {
        tmpCat.push(json[i].name)
    }
    top.catList = tmpCat
})

// Convert time to a single value
function get_time(time) {
    time = new Date(time)
    var hours = time.getHours(),
        min = time.getMinutes(),
        sec = time.getSeconds(),
        t = hours + (min / 60) + (sec / 3600)
    return t
}

// Create the XHR object.
function createCORSRequest(method, url) {
    var xhr = new XMLHttpRequest()
    if ('withCredentials' in xhr) {
        // XHR for Chrome/Firefox/Opera/Safari.
        xhr.open(method, url, true)
    } else if (typeof XDomainRequest != 'undefined') {
        // XDomainRequest for IE.
        xhr = new XDomainRequest()
        xhr.open(method, url)
    } else {
        // CORS not supported.
        xhr = null
    }
    console.log('xhr is returning')
    return xhr
}

// Weekdays - Categories - Questionnaire
var weekdays = ['monday', 'tuesday', 'wednesday', 'thursday', 'friday', 'saturday', 'sunday']
var categories = ['Residence', 'Professional & Other Places', 'Shop & Service', 'Food', 'Travel & Transport', 'Outdoors &
 ↪    Recreation', 'Arts & Entertainment', 'College & University', 'Event', 'Nightlife Spot']

// LOCAL DATA ---->

data_handling_multiple([question, sampleData, venueData], function (questionnaire, trackPoints, venueData) {
    var geo = trackPoints[0],
        venueData = venueData[0];

    var tracks = [],
        gps_data = [],
        output = [];

    // SmartTracker App Data formatting  ------->
```

```
function mobileData(mobi) {
    console.time("Benchmark - Handel SmarTracker data")
    for (i in mobi) {
        for (j in mobi[i].Place) {
            if (mobi[i].Place[j].duration > 700) {
                var tmp = {
                    name: mobi[i].Place[j].name,
                    lat: mobi[i].Place[j].CenterPoint_lat,
                    lng: mobi[i].Place[j].CenterPoint_lon,
                    identifier: mobi[i].Place[j].CenterPoint_lat + ',' + mobi[i].Place[j].CenterPoint_lon,
                    duration: mobi[i].Place[j].duration,
                    start_time: moment(mobi[i].Place[j].start_time).format("YYYY-MM-DD"),
                    end_time: moment(mobi[i].Place[j].end_time).format("YYYY-MM-DD"),
                    time: moment(mobi[i].Place[j].start_time).format("HH:mm:ss")
                };
                var mobi_tmp = {
                    name: mobi[i].Place[j].name,
                    location: {
                        accuracy: null,
                        latitude: mobi[i].Place[j].CenterPoint_lat,
                        longitude: mobi[i].Place[j].CenterPoint_lon
                    },
                    local_id: mobi[i].Place[j].Id,
                    start_time: moment(mobi[i].Place[j].start_time).format(),
                    end_time: moment(mobi[i].Place[j].end_time).format(),
                    timestamp: mobi[i].Place[j].start_time,
                    identifier: mobi[i].Place[j].CenterPoint_lat + ',' + mobi[i].Place[j].CenterPoint_lon
                };
                tracks.push(tmp);
                gps_data.push(mobi_tmp)
            }
        }
    }
    console.timeEnd("Benchmark - Handel SmarTracker data")
}

// Moves Data formating and restructring ---->
function movesData(geo) {
    console.time("Benchmark - Handel Moves data")
    for (i in geo) {
        if (i > 1800) break;
        if (geo[i].segments) {
            for (j in geo[i].segments) {
                if (geo[i].segments[j].type !== 'move') {
                    geo[i].segments[j].start_time = moment(geo[i].segments[j].start_time).format()
                    geo[i].segments[j].end_time = moment(geo[i].segments[j].end_time).format()
                    if (geo[i].date !== moment(geo[i].segments[j].end_time).format('YYYY-MM-DD')) {
                        var newTime = moment(geo[i].segments[j].start_time).hour(23).minute(59).second(59).millisecond(999)
                        geo[i].segments[j].end_time = moment(newTime).format()
                    } else if (geo[i].date !== moment(geo[i].segments[j].start_time).format('YYYY-MM-DD')) {
                        var newT = moment(geo[i].segments[j].end_time).hour(0).minute(0).second(0).millisecond(0)
                        geo[i].segments[j].start_time = moment(newT).format()
                    }
                }
                if (geo[i].segments[j].place) {
                    if (geo[i].segments[j].place.name || (geo[i].segments[j].place.mha && geo[i].segments[j].place.mha.name)) {
                        if (geo[i].segments[j].place.location) {
                            var tmp = {
                                name: geo[i].segments[j].place.mha ? geo[i].segments[j].place.mha.name :
                                  ↪    geo[i].segments[j].place.name ? geo[i].segments[j].place.name : 'unknown',
                                lat: geo[i].segments[j].place.location.lat,
                                lng: geo[i].segments[j].place.location.lon,
                                identifier: geo[i].segments[j].place.location.lat + ',' + geo[i].segments[j].place.location.lon,
                                duration: geo[i].segments[j].duration,
                                start_time: moment(geo[i].segments[j].start_time).format("YYYY-MM-DD"),
                                end_time: moment(geo[i].segments[j].end_time).format("YYYY-MM-DD"),
                                time: moment(geo[i].segments[j].start_time).format("HH:mm:ss")
                            }
                            var timeing = new Date(geo[i].segments[j].start_time);
                            var gps_tmp = {
                                name: geo[i].segments[j].place.mha ? geo[i].segments[j].place.mha.name :
                                  ↪    geo[i].segments[j].place.name ? geo[i].segments[j].place.name : 'unknown',
                                location: {
                                    accuracy: null,
                                    latitude: geo[i].segments[j].place.location.lat,
                                    longitude: geo[i].segments[j].place.location.lon
                                },
                                local_id: geo[i].segments[j].place.local_id,
                                start_time: geo[i].segments[j].start_time,
                                end_time: geo[i].segments[j].end_time,
                                timestamp: parseInt(moment(timeing).format('x')),
                                identifier: geo[i].segments[j].place.location.lat + ',' + geo[i].segments[j].place.location.lon
                            };

                            tracks.push(tmp);
                            gps_data.push(gps_tmp)

                        } // IF location
                    }
```

```
                }
            }
        }
    }
    console.timeEnd("Benchmark - Handel Moves data");
}
// movesData(geo)
mobileData(geo);
gps_data = uniqueData(gps_data);
top.__trackFreq__ = tracks;
questionnaire = questionnaire[0];
top.output = gps_data;

// var clusters = handleGPS(gps_data,0.085,1)  // <---- for MoveData
var clusters = handleGPS(gps_data, 0.06, 1); // <---- for SmarTracker data

// Benchamrk ---->
JSLitmus.test('Pre-processing Moves Data', function (count) {
    while (count--) {
        movesData()
    }
});

JSLitmus.test('Pre-processing SmarTracker Data', function (count) {
    while (count--) {
        mobileData();
    }
});

JSLitmus.test('DBSCAN', function (count) {
    while (count--) {
        handleGPS();
    }
});

for (i in clusters) {
    var tmp_clstr = {
        name: 'cluster_' + clusters[i].id,
        lat: clusters[i].location.latitude,
        lng: clusters[i].location.longitude,
        identifier: clusters[i].location.latitude + ',' + clusters[i].location.longitude,
        duration: null,
        start_time: null,
        end_time: null,
        time: null
    };
    tracks.push(tmp_clstr)
}

// tracks = uniqueData(tracks)
tracks = removeDuplicatesBy(x => x.identifier, tracks);
top.tracks = tracks;
var matrix_profile = [];
var catList = top.catList;
for (w in weekdays) {
    var new_matrix = math.zeros(24, 10);
    new_matrix = new_matrix._data;
    //   console.log("matrix for "+weekdays[w]+"----- ", new_matrix)
    for (q in questionnaire) {
        var cal = (questionnaire[q].freq / (questionnaire[q].week).length) / questionnaire[q].freq // Normilising the value
        if (isNaN(cal)) cal = 0.0001;
        var cat_index = categories.indexOf(q);
        if ((questionnaire[q].week).indexOf(weekdays[w]) != -1) {
            for (t in questionnaire[q].time) {
                if (questionnaire[q].time[t] == 'morning') {
                    for (var m = 6; m < 12; m++) {
                        new_matrix[m][cat_index] = cal;
                    }
                } else if (questionnaire[q].time[t] == 'afternoon') {
                    for (var m = 12; m < 18; m++) {
                        new_matrix[m][cat_index] = cal;
                    }
                } else if (questionnaire[q].time[t] == 'evening') {
                    for (var m = 18; m < 24; m++) {
                        new_matrix[m][cat_index] = cal;
                    }
                }
            }
        } else {
            for (var m = 0; m < 24; m++) {
                new_matrix[m][cat_index] = cal;
            }
        }
    }
    var tmpW = {
        weekday: weekdays[w],
        matrix: new_matrix
    };
    matrix_profile.push(tmpW);
```

```
};

// Create a Foursquare developer account: https://developer.foursquare.com/
// NOTE: CHANGE THESE VALUES TO YOUR OWN:
// Otherwise they can be cycled or deactivated with zero notice.
// code for myhealthavatar.org
var CLIENT_ID = '****'
var CLIENT_SECRET = '****'

// https://developer.foursquare.com/start/search
var API_ENDPOINT = 'https://api.foursquare.com/v2/venues/search' +
    '?client_id=CLIENT_ID' +
    '&client_secret=CLIENT_SECRET' +
    '&v=20160815' +
    '&ll=LATLON' +
    '&limit=15' +
    '&radius=50'
// '&callback=?'

var cat_url = 'https://api.foursquare.com/v2/venues/categories?oauth_token=000000&v=20160116'
var all_frequency = 0
for (q in questionnaire) {
    all_frequency += questionnaire[q].freq ? questionnaire[q].freq : 0
}

// Assign the venues from all the available coordinates - Disabled now as I save Feng venues data
var allTracks = [];
var intervalIndex = 0;
$(".progressDiv").show();
var timeInterval = setInterval(getVenue, 550);
var $pb = $('.progress .progress-bar').progressbar({
    display_text: 'fill'
})

function saveFile(name, type, data) {
    if (data != null && navigator.msSaveBlob)
        return navigator.msSaveBlob(new Blob([data], {
            type: type
        }), name)

    var a = $("<a style='display: none;'/>")
    var url = window.URL.createObjectURL(new Blob([data], {
        type: type
    }))
    a.attr('href', url)
    a.attr('download', name)
    $('body').append(a)
    a[0].click()
    setTimeout(function () { // fixes firefox html removal bug
        window.URL.revokeObjectURL(url)
        a.remove()
    }, 500)
}

// GETTING VENUE FROM FOURSQUARE ---->
function getVenue() {
    if (intervalIndex >= tracks.length) {
        clearInterval(timeInterval)
        var json_str = JSON.stringify(allTracks)
        saveFile('yourfilename.json', 'data:application/json', new Blob([json_str], {
            type: ''
        }))
        $('.progressDiv').hide()
        return
    }
    (function (index) {
        var api_uri = API_ENDPOINT
            .replace('CLIENT_ID', CLIENT_ID)
            .replace('CLIENT_SECRET', CLIENT_SECRET)
            .replace('LATLON', +tracks[index].lat + ',' + tracks[index].lng);
        access_foursquare(api_uri, function (data) {
            var venue = data.response.venues
            if (venue.length > 0) {
                var percentage = Math.ceil(intervalIndex * 100 / (tracks.length))
                if (typeof $pb !== 'undefined') {
                    $pb.attr('data-transitiongoal', percentage).progressbar({
                        display_text: 'fill'
                    })
                }
                intervalIndex++
            }
            for (v in venue) {
                var cat_id = venue[v].categories.length > 0 ? venue[v].categories[0].id : 'unknown'
                var cat_ref = top.ctg
                if (venue[v].categories.length > 0) {
                    for (j in cat_ref) {
                        for (k in cat_ref[j].categories) {
                            if (cat_ref[j].categories[k].id == cat_id) {
                                (venue[v].categories[0])['cat_top_level'] = cat_ref[j].name
```

```
                            } else {
                                for (c in cat_ref[j].categories[k].categories) {
                                    if (cat_ref[j].categories[k].categories[c].id == cat_id) {
                                        (venue[v].categories[0])['cat_top_level'] = cat_ref[j].name
                                    }
                                    for (s in cat_ref[j].categories[k].categories[c].categories) {
                                        if (cat_ref[j].categories[k].categories[c].categories[s].id == cat_id) {
                                            (venue[v].categories[0])['cat_top_level'] = cat_ref[j].name
                                        }
                                    }
                                }
                            }
                        }
                    }
                } else {
                    (venue[v].categories).push([{
                        'cat_top_level': 'unknown'
                    }])
                }
            }
            tracks[index]['venues'] = venue

            var tmp = {
                name: tracks[index].name,
                identifier: tracks[index].identifier,
                venues: venue
            }
            allTracks.push(tmp)
        })
    })(intervalIndex)
} // end of the function getVenue

for (i in clusters) {
    var ThisGps = []
    var associatedPlaces = []
    for (j in clusters[i].parts) {
        for (k in venueData) {
            if (gps_data[clusters[i].parts[j]].identifier == venueData[k].identifier) {
                var tmp = {
                    real_place: gps_data[clusters[i].parts[j]],
                    candidates: venueData[k]
                }
                associatedPlaces.push(venueData[k])
                ThisGps.push(tmp)
            }
        }
    }
    var mixedVenues = []
    for (v in associatedPlaces) {
        for (a in associatedPlaces[v].venues) {
            mixedVenues.push(associatedPlaces[v].venues[a])
        }
    }
    clusters[i]['mix_venues'] = removeDuplicatesBy(x => x.id, mixedVenues)
    clusters[i]['venues'] = removeDuplicatesBy(x => x.identifier, associatedPlaces)
    clusters[i]['places'] = ThisGps
}

// Counting the identifiers ------>
for (i in clusters) {
    var get_places = [],
        get_ids = [],
        id_count = {},
        tmp_count = {},
        addressing = [],
        address_count = {},
        citying = [],
        city_count = {},
    weekdays_count = {},
        time_count = {},
        weekBulk = [],
        timeBulk = [];

    for (j in clusters[i].places) {
        get_places.push(clusters[i].places[j].real_place.name)
        get_ids.push(clusters[i].places[j].real_place.identifier)
        for (p in clusters[i].places[j].candidates.venues) {
            var simpleAdd = (clusters[i].places[j].candidates.venues[p].location.address)
            simpleAdd = simpleAdd ?
              ↪    simpleAdd.replace(/\d+|^\s+|\s+$|-|\.|\b(unit|floor|level|first|second|ground)\b\s*|\,|\/|\#/gi,
              ↪    '').trim().toLowerCase() : undefined
            simpleAdd = simpleAdd ? simpleAdd.replace(/(\b(a|b|first|second|third|fourth|fifth|ground|upper|lower|g)\b\s*)/gi,
              ↪    '').trim() : undefined
            simpleAdd = simpleAdd ? simpleAdd.replace(/(\b(rd)\b\s*)/gi, 'road') : undefined
            simpleAdd = simpleAdd ? simpleAdd.replace(/(\b(ln)\b\s*)/gi, 'lane') : undefined
            simpleAdd = simpleAdd ? simpleAdd.replace(/(\b(cntr)\b\s*)/gi, 'centre') : undefined
            simpleAdd = simpleAdd ? simpleAdd.replace(/st$/, 'street') : undefined
            simpleAdd = simpleAdd ? removeChar(simpleAdd) : undefined
            var thisCity = (clusters[i].places[j].candidates.venues[p].location.city)
```

```
                thisCity = thisCity ? thisCity.replace(/,/g, ' ').trim().toLowerCase() : undefined

                // var stopWordsRE = /(?:^|\s+)(?:foo|bar)(?=\s+|$)/gi
                if (simpleAdd) addressing.push(simpleAdd)
                citying.push(thisCity)
            }
            var place_time = new Date(clusters[i].places[j].real_place.start_time),
             palce_end_time = new Date(clusters[i].places[j].real_place.end_time),
             week_days = moment(place_time).format('dddd').toLowerCase(), // Wednesday
             dayOfWeek = moment(place_time).format('d'), // Sunday is 0, monday 1, ...
             s_time = get_time(place_time), // 18.10
             e_time = get_time(palce_end_time),
            time_up = Math.floor(s_time),
            time_down = Math.ceil(e_time),
             timeDif = time_down - time_up;

            if (timeDif >= 1) {
                for (var m = time_up; m < time_down; m++) {
                    timeBulk.push(m)
                }
            }
            weekBulk.push(weekdays.indexOf(week_days))
        }

        for (k in get_places) {
            tmp_count[get_places[k]] = (tmp_count[get_places[k]] || 0) + 1
        }
        for (t in get_ids) {
            id_count[get_ids[t]] = (id_count[get_ids[t]] || 0) + 1
        }

        for (h in id_count) {
            id_count[h] = id_count[h] / (get_ids.length)
        }

        for (v in addressing) {
            address_count[addressing[v]] = (address_count[addressing[v]] || 0) + (1 / addressing.length)
        }

        for (s in citying) {
            city_count[citying[s]] = (city_count[citying[s]] || 0) + (1 / citying.length)
        }

        timeBulk.sort(sortNumber)

        for (a in timeBulk) {
            time_count[timeBulk[a]] = (time_count[timeBulk[a]] || 0) + 1
        }

        var norm_time_count = math.zeros(24)
        norm_time_count = norm_time_count._data

        for (g in time_count) {
            norm_time_count[g] = time_count[g] || 0 // Normilise the weekday array
        }

        for (b in weekBulk) {
            weekdays_count[weekBulk[b]] = (weekdays_count[weekBulk[b]] || 0) + 1
        }
        var norm_week_count = math.zeros(7)
        norm_week_count = norm_week_count._data

        for (f in weekdays_count) {
            norm_week_count[f] = weekdays_count[f] || 0 // Normilise the weekday array
        }
        var tmp = {
            count_name: tmp_count,
            count_id: id_count,
            address_count: address_count,
            city_count: city_count,
            weekdays_count: norm_week_count,
            time_count: norm_time_count
        }
        clusters[i]['counts'] = tmp

    } // END OF COUNTING

    // CATEGORY VOTING ---->

    for (i in clusters) {
        var vote = {}
        var cat_matrix = {}
        var cat_hr = {}
        var unique_cat = []
        for (j in clusters[i].mix_venues) {
            var placeMatrix = math.zeros(7, 24)
            var placeMatrixWknd = math.zeros(2, 24)
            if (clusters[i].mix_venues[j].categories[0].cat_top_level) {
                unique_cat.push(clusters[i].mix_venues[j].categories[0].cat_top_level)
```

```
        }
        var thisId = clusters[i].mix_venues[j].id
        var tmp = {
            identifier: clusters[i].mix_venues[j].id,
            name: clusters[i].mix_venues[j].name,
            cluster_id: clusters[i].id,
            place_matrix: placeMatrix._data,
            place_week_weknd: placeMatrixWknd._data
        }
        vote[thisId] = tmp
    }

    unique_cat = uniqueArray(unique_cat)

    var cat_hr_all = math.zeros(10, 24)
    cat_hr_all = cat_hr_all._data
    for (x in unique_cat) {
        var cat_in = math.zeros(24)
        cat_hr[unique_cat[x]] = cat_in._data
    }
    for (f in weekdays) {
        var catMatrix = math.zeros(24, 10)
        cat_matrix[weekdays[f]] = catMatrix._data
    }
    for (p in clusters[i].places) {
        for (c in clusters[i].places[p].candidates.venues) {
            var idCan = clusters[i].places[p].candidates.venues[c].id
            // Calculate the Haversine distance between each coordinate and venues -->
            var coord_id = (clusters[i].places[p].candidates.identifier).split(',')
            var venue_coord = clusters[i].places[p].candidates.venues[c].location
            var startPoint = {
                    latitude: coord_id[0],
                    longitude: coord_id[1]
                },
                endPoint = {
                    latitude: venue_coord.lat,
                    longitude: venue_coord.lng
                }
            var haversineDistance = Math.round(haversine(startPoint, endPoint, {
                unit: 'meter'
            }))
            venue_coord['haversine_dist'] = haversineDistance
            var venueDistance = 1 / (haversineDistance) == Infinity ? 0.01 : 1 / (haversineDistance),
             place_time = new Date(clusters[i].places[p].real_place.start_time),
             palce_end_time = new Date(clusters[i].places[p].real_place.end_time),
             week_days = moment(place_time).format('dddd').toLowerCase(), // Wednesday
             dayOfWeek = moment(place_time).format('d'), // Sunday is 0, monday 1, ...
             thisCategory = clusters[i].places[p].candidates.venues[c].categories[0].cat_top_level,
            if (thisCategory == undefined) continue;
            var s_time = get_time(place_time), // 18.10
             e_time = get_time(palce_end_time);
            time_up = Math.floor(s_time)
            time_down = Math.ceil(e_time)
            for (t in vote) {
                if (idCan == vote[t].identifier) {
                    var placeIdentifier = clusters[i].places[p].real_place.identifier
                    var thisCount = clusters[i].counts.count_id[placeIdentifier]
                    var timeDif = time_down - time_up
                    if (timeDif >= 1) {
                        for (var m = time_up; m < time_down; m++) {
                            vote[t].place_matrix[(weekdays.indexOf(week_days))][m] += (venueDistance * thisCount) / timeDif
                          vote[t].place_week_weknd[(dayOfWeek == 0 || dayOfWeek == 6) ? 1 : 0][m] += (venueDistance * thisCount)
                            ↪    / timeDif
                            cat_matrix[week_days][m][categories.indexOf(thisCategory)] += (venueDistance * thisCount) / timeDif
                        }
                    } else {
                       console.log('option 2 is activated', time_down, time_up, time_down - time_up, place_time, palce_end_time)
                        vote[t].place_matrix[(weekdays.indexOf(week_days))][time_up] += 1
                    }
                }
            } // end of t vote
            if (timeDif >= 1) {
                for (var g = time_up; g < time_down; g++) {
                    cat_hr_all[categories.indexOf(thisCategory)][g] += (venueDistance * thisCount) // timeDif
                }
            }
        }
    }
    // Normilise the values ----->
    var lin = d3.scale ? d3.scale.linear : d3.scaleLinear;
    var norm_cats = lin()
        .domain([math.min(cat_hr_all), math.max(cat_hr_all)])
        .range([0, 1])
    for (h in cat_hr_all) {
        for (f in cat_hr_all[h]) {
            cat_hr_all[h][f] = norm_cats(cat_hr_all[h][f])
        }
    }
    clusters[i]['category_24hr'] = cat_hr
```

```
        clusters[i]['vote'] = vote
        clusters[i]['categories_vote'] = cat_matrix
        clusters[i]['cat_24hr_matrix'] = cat_hr_all
    }
    var allDimensions = [];
    for (i in clusters) {
        var time = clusters[i].counts.time_count;
        var week = clusters[i].counts.weekdays_count;
        allDimensions.push(clusters[i].dimension);
    }

    // Temp disabling the ClusterPoints
    clusterPoints = []
    top.voteData = []
    for (k in clusterPoints) {
        (function (index) {
            $.ajax({
                    crossDomain: true, // added in jQuery 1.5
                    headers: {
                        'Access-Control-Allow-Origin': '*'
                    },
                    type: 'GET',
                    dataType: 'jsonp',
                    async: false,
                    url: API_ENDPOINT
                        .replace('CLIENT_ID', CLIENT_ID)
                        .replace('CLIENT_SECRET', CLIENT_SECRET)
                        .replace('LATLON', +clusterPoints[index].location.latitude + ',' + clusterPoints[index].location.longitude),
                    // success: crosSucessHandler(result, status)
                })
                .done(function (data) {
                    var venue = data.response.venues
                    for (v in venue) {
                        var cat_id = venue[v].categories.length > 0 ? venue[v].categories[0].id : 'unknown'
                        var cat_ref = top.ctg
                        if (venue[v].categories.length > 0) {
                            for (j in cat_ref) {
                                for (k in cat_ref[j].categories) {
                                    if (cat_ref[j].categories[k].id == cat_id) {
                                        (venue[v].categories[0])['cat_top_level'] = cat_ref[j].name
                                    } else {
                                        for (c in cat_ref[j].categories[k].categories) {
                                            if (cat_ref[j].categories[k].categories[c].id == cat_id) {
                                                (venue[v].categories[0])['cat_top_level'] = cat_ref[j].name
                                            }
                                            for (s in cat_ref[j].categories[k].categories[c].categories) {
                                                if (cat_ref[j].categories[k].categories[c].categories[s].id == cat_id) {
                                                    (venue[v].categories[0])['cat_top_level'] = cat_ref[j].name
                                                }
                                            }
                                        }
                                    }
                                }
                            }
                        } else {
                            (venue[v].categories).push([{
                                'cat_top_level': 'unknown'
                            }])
                        }
                    }
                    var part_index = clusterPoints[index].parts
                    var start_time_avg = 0
                    var end_time_avg = 0
                    var st = []
                    var wkd = []
                    for (m in part_index) {
                        start_time_avg += get_time(output[part_index[m]].start_time)
                        end_time_avg += get_time(output[part_index[m]].end_time)
                        st.push(Math.floor(get_time(output[part_index[m]].start_time)),
                          ↪     Math.floor(get_time(output[part_index[m]].end_time)))
                        wkd.push(moment(output[part_index[m]].start_time).format('dddd'))
                        output[part_index[m]]['cluster_id'] = clusterPoints[index].id ? clusterPoints[index].id : 0
                    }
                    st = uniqueArray(st)
                    wkd = uniqueArray(wkd)
                    var dimAvg = clusterPoints[index].dimension
                    clusterPoints[index]['days'] = wkd
                    clusterPoints[index]['hours'] = st.sort(sortNumber)
                    var categoryFactor = math.zeros(10)
                    categoryFactor = categoryFactor._data
                    var voteMatrix = math.zeros(24, 10)
                    voteMatrix = voteMatrix._data
                    var categoryScore = math.zeros(10)
                    categoryScore = categoryScore._data
                    for (v in venue) {
                        var cat_top = venue[v].categories[0].cat_top_level
                        var catIndex = categories.indexOf(cat_top)
                        categoryFactor[catIndex] += (1 / venue.length)
                    }
```

```
                    for (t in venue) {
                        var cat_top = venue[t].categories[0].cat_top_level
                        if (!cat_top) continue
                        var catIndex = categories.indexOf(cat_top)
                        var delta = (1 / venue[t].location.distance) * (venue[t].stats.checkinsCount) / (categoryFactor[catIndex])

                        categoryScore[catIndex] += (isNaN(delta) ? 0 : delta)
                    }
                    norm_value = d3.extent(categoryScore, function (d) {
                        return d
                    })
                    //  Normilise min max
                    for (i in categoryScore) {
                        categoryScore[i] = ((categoryScore[i] - norm_value[0]) / (norm_value[1] - norm_value[0]))
                    }
                    var sum = categoryScore.reduce(function (a, b) {
                        return a + b;
                    }, 0)
                    //  Normilise between 0 and 1
                    for (i in categoryScore) {
                        categoryScore[i] = (categoryScore[i] / sum)
                    }
                    for (t in st) {
                        voteMatrix[st[t]] = math.dotDivide(categoryScore, (wkd.length / 7))
                    }
                    clusterPoints[index]['vote'] = voteMatrix
                })
                .fail(function () {
                    console.log('error')
                })
        })(k)
    }
    for (i in clusterPoints) {
        if (clusterPoints.hasOwnProperty(i)) {}
    }
    setTimeout(function () {
        console.time("Benchmark - Annotation Model B")
        trackPoints = output
        var all_final_values = []
        var minMaxDimension = d3.extent(clusters, function (d) {
            return d.dimension
        })
        for (i in clusters) {
            for (j in clusters[i].places) {
                if (j > 0) continue;
                var thisPlace = clusters[i].places[j]
                var place_time = new Date(thisPlace.real_place.start_time)
                place_end_time = new Date(thisPlace.real_place.end_time),
                    weekdays = moment(place_time).format('dddd').toLowerCase(), // Wednesday
                    dayOfWeeks = moment(place_time).format('d'); // Sunday is 0, monday 1, ...
                date_info = moment(place_time).format('DD-MMM-YYYY').toLowerCase(),
                    s_time = get_time(place_time), // 18.10
                    e_time = get_time(place_end_time),
                    time_up = Math.floor(s_time),
                    time_down = Math.ceil(e_time),
                    ffp = Math.ceil(e_time)

                for (p in matrix_profile) {
                    if (matrix_profile[p].weekday === weekdays) {
                        var time_of_event = matrix_profile[p].matrix[time_up],
                         time_of_event2 = math.zeros(10),
                         timeDif = time_down - time_up,
                        if (timeDif >= 1) {
                            for (var t = time_up; t < time_down; t++) {
                                var thisVector = matrix_profile[p].matrix[t]
                                time_of_event2 = time_of_event2.map(function (num, idx) {
                                    return num + thisVector[idx];
                                })
                            }
                        }
                        time_of_event2 = time_of_event2._data
                        var placeCandidates = thisPlace.candidates.venues
                        var thisBatch = []
                        for (m in placeCandidates) {
                            if (placeCandidates[m].categories.length == 0) continue; // ingnore the venues without category
                            if (placeCandidates[m].categories[0].cat_top_level == 'unknown') continue;
                            var cat_top = placeCandidates[m].categories[0].cat_top_level,
                             cat_ref = top.ctg,
                             catIndex = categories.indexOf(cat_top),
                             venue_id = placeCandidates[m].id,
                             categoryVote = 0.01,
                             venue_score = 0.01;
                            if (timeDif >= 1) {
                                for (var w = time_up; w < time_down; w++) {
                                   venue_score += clusters[i].vote[venue_id].place_week_weknd[(dayOfWeeks == 0 || dayOfWeeks == 6) ?
                                     ↪    1 : 0][w] / timeDif
                                    categoryVote += clusters[i].cat_24hr_matrix[catIndex] ? clusters[i].cat_24hr_matrix[catIndex][w]
                                     ↪    / timeDif : 0
                                }
```

```
                        }
                        var prob_day_place = time_of_event2[catIndex] != 0 ? time_of_event2[catIndex] : 0.01,
                         place_prob = (cat_top == undefined) ? 0.01 : (questionnaire[cat_top].freq / all_frequency),
                        place_prob = isNaN(place_prob) ? 0.01 : place_prob != 0 ? place_prob : 0.01,
                         distance = placeCandidates[m].location.haversine_dist == 0 ? 1 :
                          ↪    placeCandidates[m].location.haversine_dist,
                         address = placeCandidates[m].location.formattedAddress,
                         finalProbability = (1 / distance) * 1000 * (prob_day_place == undefined ? 0.01 : prob_day_place) *
                          ↪    venue_score * categoryVote;

                        var tmp = {
                            venue_info: placeCandidates[m],
                            probability: isNaN(finalProbability) ? 0 : finalProbability,
                            name: placeCandidates[m].name,
                            category_top_level: cat_top,
                            category_details: placeCandidates[m].categories[0].name ? placeCandidates[m].categories[0].name :
                             ↪    undefined,
                            distance: distance,
                            start_time: place_time,
                            end_time: place_end_time
                        }
                        thisBatch.push(tmp)
                    }
                }
            }

            var fin_tmp = thisPlace.real_place
            var cityTmp = []
            for (var index in clusters[i].counts.city_count) {
                if (index == undefined) continue
                var ctm = {
                    city: index,
                    probability: clusters[i].counts.city_count[index]
                }
                cityTmp.push(ctm)
            }

            var addTmp = []
            for (var add in clusters[i].counts.address_count) {
                var atm = {
                    address: add,
                    probability: clusters[i].counts.address_count[add]
                }
                addTmp.push(atm)
            }
            fin_tmp['enrichment'] = thisBatch;
            fin_tmp['address_probability'] = addTmp;
            fin_tmp['city_probability'] = cityTmp;
            fin_tmp['weekdays'] = clusters[i].counts.weekdays_count;
            fin_tmp['hours'] = clusters[i].counts.time_count;
            fin_tmp['assumption'] = clusters[i].dimension == minMaxDimension[1] ? 'Home' : clusters[i].dimension ==
             ↪    allDimensions[1] ? 'Workplace' : null;
            all_final_values.push(fin_tmp)
        }
    }

    var timeIntervalGeo = setInterval(GeoCluster, 200);
    var intervalIndexGeo = 0;
    function GeoCluster() {
        if (intervalIndexGeo >= all_final_values.length) {
            clearInterval(timeIntervalGeo)
            return
        }
        (function (index) {
            var lat = all_final_values[intervalIndexGeo].location.latitude,
                lng = all_final_values[intervalIndexGeo].location.longitude;
            var urlGeo = 'https://maps.googleapis.com/maps/api/geocode/json?latlng=' + lat + ',' + lng + '&sensor=true';
            $.getJSON(urlGeo, function (geo) {
                if (geo.status == 'OK') {
                    intervalIndexGeo++;
                }
                var content = geo.results[0].formatted_address;
                all_final_values[index]['geo_address'] = content;
            })
        })(intervalIndexGeo)
    }
    setTimeout(function () {
        for (i in all_final_values) {
            var thisdata = all_final_values[i].enrichment
            var minMax = d3.extent(thisdata, function (d) {
                return d.probability
            })
            norm_cats.domain(minMax)
            var sumAllProbability = d3.sum(thisdata, function (d) {
                return d.probability
            })
            var tableProb = "",
                tableAddress = "",
                tabelCity = "";
```

```
                for (j in all_final_values[i].enrichment) {}
                (thisdata).sort(function (a, b) {
                    return d3.descending(a.probability, b.probability)
                });
                (all_final_values[i].address_probability).sort(function (a, b) {
                    return d3.descending(a.probability, b.probability)
                });
                (all_final_values[i].city_probability).sort(function (a, b) {
                    return d3.descending(a.probability, b.probability)
                })
                var enrch = all_final_values[i].enrichment;
                for (c in enrch) {

                }
                var addp = all_final_values[i].address_probability;
                var citp = all_final_values[i].city_probability;
                for (d in addp) {
                    tableAddress += "<tr><td>" + addp[d].address + "</td><td>" + addp[d].probability + "</td></tr>"
                }

                for (g in citp) {
                    tabelCity += "<tr><td>" + citp[g].city + "</td><td>" + citp[g].probability + "</td></tr>"
                }


                $("#result_in").append(content_res)
                initClusterMap(all_final_values[i], i);
                drawingCharts(all_final_values[i].weekdays, i)
                drawingChartshr(all_final_values[i].hours, i)
            }
            $('.process').hide()
        }, 6000)
    }, 4000)
})
```

C.2 Significance ranking pseudo code

```
function handle_data(data, option) {
    var places_name = [];
    var tf_data = [];

    for (i in data) {
        var tmp = [];
        for (j in data[i]) {
            for (k in data[i][j]) {
                if (k === 'date') continue;
                if (data[i][j][k].place && data[i][j][k].place.mha) {
                    places_name.push(data[i][j][k].place.mha.name);
                    var b_cat = (data[i][j][k].place && data[i][j][k].place.mha !== null) ? data[i][j][k].place.mha.category :
                     ↪    "unknown";
                    var cat_index = categories.indexOf(Cat_maping(b_cat));
                    var temp_location = {
                        duration: moment(data[i][j][k].end_time).diff(moment(data[i][j][k].start_time), 'seconds'),
                        name: data[i][j][k].place.mha.name,
                        location: data[i][j][k].place.location,
                        ref: data[i][j][k].mid,
                        start_time: data[i][j][k].start_time,
                        end_time: data[i][j][k].end_time,
                        cat: Cat_maping(b_cat)
                    }
                    tmp.push(temp_location);
                }
            }
        }
        tf_data.push(tmp)
    }

    if (option && option.w3) {
        return tf_data
    } else {
        var arr = places_name;
        var list_places = uniqueArray(places_name);
        var tmp_p = [];
        for (j in tf_data) {
            var tmp_count = {};
            for (k in tf_data[j]) {
                tmp_count[tf_data[j][k].name] = (tmp_count[tf_data[j][k].name] || 0) + 1;
            }
            tmp_p.push(tmp_count)
        }
```

```
var df_freq = [];
var occ_res = {};
var occ_in_month = [];
for (i in tmp_p) {
    if (Object.keys(tmp_p[i]).length <= 0) continue;
    var place_occ = {};
    var occurance = {};

    for (n in list_places) {
        place_occ[list_places[n]] = (tmp_p[i][list_places[n]]) ? 1 : 0;
        occ_res[list_places[n]] = 0;
        occurance[list_places[n]] = (tmp_p[i][list_places[n]]) ? (tmp_p[i][list_places[n]]) : 0;
    }
    df_freq.push(place_occ)
    occ_in_month.push(occurance)
}

//document frequency
for (i in df_freq) {
    for (j in df_freq[i]) {
        occ_res[j] += (df_freq[i][j]);
    }
}

var N_df = df_freq.length
var idf_dic = {};

//calculate the IDF Dictionary
for (i in occ_res) {
    idf_dic[i] = Math.log10(N_df / occ_res[i]) + 0.00001;
}

top.__idf_dic__ = idf_dic;
for (i in tf_data) {
    for (j in tf_data[i]) {
        tf_data[i][j]['tf_idf'] = Math.pow(tf_data[i][j].duration, top.w2_frequency) * idf_dic[tf_data[i][j].name] *
         ↪    top.influence_factor[tf_data[i][j].cat]
    }
}

var tmp_duration = [];
var total_dr = {};
var occ_final = [];
for (j in tf_data) {
    var tmp_dr = {};
    var tmp_occ = {};
    for (k in tf_data[j]) {
        tmp_dr[tf_data[j][k].name] = (tmp_dr[tf_data[j][k].name] || 0) + tf_data[j][k].duration;
        total_dr[tf_data[j][k].name] = (total_dr[tf_data[j][k].name] || 0) + tf_data[j][k].duration;
        tmp_occ[tf_data[j][k].name] = (tmp_occ[tf_data[j][k].name] || 0) + 1;
    }
    tmp_duration.push(tmp_dr)
    occ_final.push(tmp_occ)
}

for (i in tf_data) {
    tf_data[i].sort(function (a, b) {
        return d3.descending(a.tf_idf, b.tf_idf)
    })
}

var obj = {};
for (var i = 0, j = arr.length; i < j; i++) {
    obj[arr[i]] = (obj[arr[i]] || 0) + 1;
}

//Calculate the Variance
var raw_variance = [];
for (i in obj) {
    var variance = 0;
    var avg = (obj[i]) / (df_freq.length);
    for (j in occ_in_month) {
        variance += Math.pow((occ_in_month[j][i] - avg), 2);
    }
    raw_variance[i] = (((idf_dic[i] / variance) == Infinity ? 0 : (idf_dic[i] / variance) / occ_res[i]) + idf_dic[i]) *
     ↪    total_dr[i]
}

//Scaling the Results between 0 - 1
var max_variance = d3.max(d3.values(raw_variance))
var variance_scale = d3.scale.linear()
    .domain([0, max_variance])
    .range([0, 1]);
var final_variance_norm = []
for (i in raw_variance) {
    final_variance_norm[i] = variance_scale(raw_variance[i]);
}
```

```
        //Adding the variance to the final return data
        for (i in tf_data) {
            for (j in tf_data[i]) {
                tf_data[i][j]['variance'] = final_variance_norm[tf_data[i][j].name]
                tf_data[i][j]['score'] = (1 - top.w1_gradient) * (tf_data[i][j].tf_idf) + (top.w1_gradient *
                 ↪    final_variance_norm[tf_data[i][j].name])
            }
        }
        return tf_data;
    }
}

function significantRanking(data, sensor, start, end, option, constrains) {

    var i, j, date = [],
        output = [],
        places = [],
        moving = [],
        item = [],
        momo = [],
        dataset = data,
        fitbit = [],
        withings = [],
        segData = [],
        segData_filtered = [];
    constrains = constrains || {};
    top.__monthMatrix__ = top.__monthMatrix__ || [];
    top.t__mha__a = top.t__mha__a || 7e8;
    top.t__mha__p = top.t__mha__p || 8e8;
    // maybe should clear each time
    // object of date yyyy-MM-dd, which it self will be a 24 hourly array, which contains another array of
    // arrays of categories, which contains array of ids
    top.t__mha__m = top.t__mha__m || {};
    // a reference back to the segment data
    top.t__mha__m.data = top.t__mha__m.data || {};
    // store the places information
    top.t__mha__m.poi = top.t__mha__m.poi || {};
    var isConstrain = Object.keys(constrains).length;

    top.search__places__all = []
    if (dataset != null) {
        for (i in dataset) {
            if (sensor === "moves" && dataset[i].summary.source === sensor) {
                if (dataset[i].segments && start <= dataset[i].date && dataset[i].date <= end) {
                    date.push(dataset[i].date);
                    var seg_temp = [],
                        seg_temp_filter = [];
                    for (j in dataset[i].segments) {
                        //get all the place name for search
                        if (!constrains.cat && dataset[i].segments[j].place && dataset[i].segments[j].place.mha) {
                            top.search__places__all.push(dataset[i].segments[j].place.name ? dataset[i].segments[j].place.name :
                             ↪    dataset[i].segments[j].place.mha.name);
                        } else if (constrains.cat && dataset[i].segments[j].place && dataset[i].segments[j].place.mha &&
                         ↪    (dataset[i].segments[j].place.mha !== null ?
                         ↪    (constrains.cat).indexOf(Cat_maping(dataset[i].segments[j].place.mha.category)) != -1 : {})) {
                            top.search__places__all.push(dataset[i].segments[j].place.name ? dataset[i].segments[j].place.name :
                             ↪    dataset[i].segments[j].place.mha.name);
                        }
                        if (dating(dataset[i].segments[j].end_time) != dataset[i].date) {
                            dataset[i].segments[j].start_time = moment(dataset[i].segments[j].start_time).format("YYYY-MM-DD
                             ↪    HH:mm:ssZZ");
                            dataset[i].segments[j].end_time = moment(new Date(dataset[i].segments[j].start_time).setHours(23, 59,
                             ↪    59)).format("YYYY-MM-DD HH:mm:ssZZ");
                        }
                        if (dating(dataset[i].segments[j].start_time) != dataset[i].date) {
                            dataset[i].segments[j].start_time = moment(new Date(dataset[i].segments[j].end_time).setHours(0, 0,
                             ↪    0)).format("YYYY-MM-DD HH:mm:ssZZ");
                        }
                       dataset[i].segments[j].start_time = moment(dataset[i].segments[j].start_time).format("YYYY-MM-DD HH:mm:ssZZ");
                        dataset[i].segments[j].end_time = moment(dataset[i].segments[j].end_time).format("YYYY-MM-DD HH:mm:ssZZ");

                        //try to remove the duplicate segments which has start today and end tomorrow - Farzad
                        if (isConstrain != 0 && constrains.place && dataset[i].segments[j].type !== "move" &&
                         ↪    dataset[i].segments[j].place && (dataset[i].segments[j].place.name ||
                         ↪    dataset[i].segments[j].place.mha) && dataset[i].date === dating(dataset[i].segments[j].start_time)
                         ↪    && ((dataset[i].segments[j].place.name ?
                         ↪    (constrains.place).indexOf(dataset[i].segments[j].place.name) != -1 :
                         ↪    dataset[i].segments[j].place.mha !== null ?
                         ↪    (constrains.place).indexOf(dataset[i].segments[j].place.mha.name) != -1 : {}))) {

                            //Check for any category in search
                            if (constrains.cat && constrains.time.length == 0 && (dataset[i].segments[j].place.mha !== null ?
                             ↪    (constrains.cat).indexOf(Cat_maping(dataset[i].segments[j].place.mha.category)) != -1 : {})) {

                                var place_filter = dataset[i].segments[j];
                                seg_temp_filter.push(place_filter);

                                //Check if there is any time input in the search
```

```
        } else if (!constrains.cat && constrains.time.length != 0 && dataset[i].segments[j].start_time &&
         ↪    dataset[i].segments[j].end_time) {

            for (k in constrains.time) {
                if (timing(dataset[i].segments[j].start_time) >= constrains.time[k].start_time &&
                 ↪    timing(dataset[i].segments[j].end_time) <= constrains.time[k].end_time) {

                    if (dating(dataset[i].segments[j].end_time) != dataset[i].date) {
                        dataset[i].segments[j].start_time =
                         ↪    moment(dataset[i].segments[j].start_time).format("YYYY-MM-DD HH:mm:ssZZ");
                        dataset[i].segments[j].end_time = moment(new
                         ↪    Date(dataset[i].segments[j].start_time).setHours(23, 59,
                         ↪    59)).format("YYYY-MM-DD")
                    }
                    var place_filter = dataset[i].segments[j];
                    seg_temp_filter.push(place_filter);
                }
            }
        } else if (constrains.cat && constrains.time.length != 0 && (dataset[i].segments[j].place.mha !== null ?
         ↪    (constrains.cat).indexOf(Cat_maping(dataset[i].segments[j].place.mha.category)) != -1 : {}) &&
         ↪    dataset[i].segments[j].start_time && dataset[i].segments[j].end_time) {
            for (h in constrains.time) {
                if (timing(dataset[i].segments[j].start_time) >= constrains.time[h].start_time &&
                 ↪    timing(dataset[i].segments[j].end_time) <= constrains.time[h].end_time) {
                    if (dating(dataset[i].segments[j].end_time) != dataset[i].date) {
                        dataset[i].segments[j].start_time =
                         ↪    moment(dataset[i].segments[j].start_time).format("YYYY-MM-DD HH:mm:ssZZ");
                        dataset[i].segments[j].end_time = moment(new
                         ↪    Date(dataset[i].segments[j].start_time).setHours(23, 59,
                         ↪    59)).format("YYYY-MM-DD HH:mm:ssZZ");
                    }
                    var place_filter = dataset[i].segments[j];
                    seg_temp_filter.push(place_filter);
                }
            }
        } else if (!constrains.cat && constrains.time.length == 0) {
            var place_filter = dataset[i].segments[j];
            seg_temp_filter.push(place_filter);
        }

        // For no constrain when first load --->
    } else if (isConstrain != 0 && !constrains.place && !constrains.cat && constrains.time.length != 0 &&
     ↪    dataset[i].segments[j].start_time && dataset[i].segments[j].end_time && dataset[i].date ===
     ↪    dating(dataset[i].segments[j].start_time)) {
        for (t in constrains.time) {
            if (timing(dataset[i].segments[j].start_time) >= constrains.time[t].start_time &&
             ↪    timing(dataset[i].segments[j].end_time) <= constrains.time[t].end_time) {
                if (dating(dataset[i].segments[j].end_time) != dataset[i].date) {
                    dataset[i].segments[j].start_time =
                     ↪    moment(dataset[i].segments[j].start_time).format("YYYY-MM-DD HH:mm:ssZZ");
                    dataset[i].segments[j].end_time = moment(new
                     ↪    Date(dataset[i].segments[j].start_time).setHours(23, 59, 59)).format("YYYY-MM-DD
                     ↪    HH:mm:ssZZ");
                }
                var place_filter = dataset[i].segments[j];
                seg_temp_filter.push(place_filter);
            }
        }
    } else if (isConstrain != 0 && constrains.cat && !constrains.place && constrains.time.length == 0 &&
     ↪    dataset[i].segments[j].place && dataset[i].segments[j].place.mha && dataset[i].date ===
     ↪    dating(dataset[i].segments[j].start_time) && (dataset[i].segments[j].place.mha !== null ?
     ↪    (constrains.cat).indexOf(Cat_maping(dataset[i].segments[j].place.mha.category)) != -1 : {})) {
        var place_filter = dataset[i].segments[j];
        seg_temp_filter.push(place_filter);
    } else if (isConstrain != 0 && !constrains.place && constrains.cat && constrains.time.length != 0 &&
     ↪    dataset[i].segments[j].place && dataset[i].segments[j].place.mha &&
     ↪    (dataset[i].segments[j].place.mha !== null ?
     ↪    (constrains.cat).indexOf(Cat_maping(dataset[i].segments[j].place.mha.category)) != -1 : {}) &&
     ↪    dataset[i].segments[j].start_time && dataset[i].segments[j].end_time) {
        for (h in constrains.time) {
            if (timing(dataset[i].segments[j].start_time) >= constrains.time[h].start_time &&
             ↪    timing(dataset[i].segments[j].end_time) <= constrains.time[h].end_time) {
                if (dating(dataset[i].segments[j].end_time) != dataset[i].date) {
                    dataset[i].segments[j].start_time =
                     ↪    moment(dataset[i].segments[j].start_time).format("YYYY-MM-DD HH:mm:ssZZ");
                    dataset[i].segments[j].end_time = moment(new
                     ↪    Date(dataset[i].segments[j].start_time).setHours(23, 59, 59)).format("YYYY-MM-DD
                     ↪    HH:mm:ssZZ");

                }
                var place_filter = dataset[i].segments[j];
                seg_temp_filter.push(place_filter);
            }
        }
    } else if ((isConstrain == 0 || !constrains.place && constrains.time.length == 0 && !constrains.activity &&
     ↪    !constrains.cat) && dataset[i].segments[j].type !== "move" && dataset[i].date ===
     ↪    dating(dataset[i].segments[j].start_time)) {
        // console.log('-------- No Constrains ----------')
        var place_filter = dataset[i].segments[j];
```

```
                        seg_temp_filter.push(place_filter);
                    }
                }
                segData_filtered.push(seg_temp_filter);
                for (var d in segData_filtered) {
                    if (segData_filtered.hasOwnProperty(d)) segData_filtered[d]["date"] = date[d];
                }
                segData.push(seg_temp);
                for (var t in segData) {
                    if (segData.hasOwnProperty(t)) segData[t]["date"] = date[t];
                }
            } else {
                momo.push(dataset[i]);
            }
        }
    }
}

var obj = {};
for (var i = 0, j = (top.search__places__all).length; i < j; i++) {
    obj[top.search__places__all[i]] = (obj[top.search__places__all[i]] || 0) + 1;
}

var searchEngineData = [];
for (i in obj) {
    var tmp_eng = {
        value: i,
        occurence: obj[i]
    }

    searchEngineData.push(tmp_eng)
}

searchEngineData = searchEngineData.sort(function (a, b) {
    return d3.descending(a.variance, b.variance);
});

top._searchEngineData_ = searchEngineData;

var pureData = uniqueData(segData);
var pureData_filtered = uniqueData(segData_filtered);

if (!option) {
    var temp_mat = [];
    for (var j = 0; j < 12; j++) {
        var temp = {};
        temp_mat.push(temp);
    }
    var month_matrix = [],
        month_matrix_filtered = [];
    for (m in temp_mat) {
        var t_m = [];
        var t_f = [];
        for (i in pureData) {
            if (!pureData.hasOwnProperty(i)) continue;
            var date = pureData[i].date;
            var month = new Date(date).getMonth();
        }
        month_matrix.push(t_m);

        for (k in pureData_filtered) {
            if (!pureData_filtered.hasOwnProperty(k)) continue;
            var date_f = pureData_filtered[k].date;
            var month_f = new Date(date_f).getMonth();
            if (month_f == m) t_f.push(pureData_filtered[k]);
        }
        month_matrix_filtered.push(t_f)
    }
} else if (option === 'year') {

    var temp_mat = [];
    for (var j = 0; j < 11; j++) {
        var temp = {};
        temp_mat.push(temp);
    }
    var year_matrix = [];
    for (n in temp_mat) {
        var t_y = [];
        for (i in pureData) {
            var date = pureData[i].date;
            var year = new Date(date).getFullYear();
            if (year == (parseInt(n) + 2005)) t_y.push(pureData[i]);
        }
        year_matrix.push(t_y);
    }
}

var checkIfEmpty = 0;
for (e in month_matrix_filtered) {
```

```
        for (t in month_matrix_filtered[e]) {
            if (month_matrix_filtered[e][t].length > 0) checkIfEmpty++;
        }
    }

    month_matrix_filtered = (checkIfEmpty != 0) ? month_matrix_filtered : top._prev_month_matrix_filtered;
    top._prev_month_matrix_filtered = month_matrix_filtered;
    top.__monthMatrix__ = month_matrix;

    return option === 'year' ? year_matrix : month_matrix_filtered;
};
```

Bibliography

[1] Aggarwal, C. C. and Han, J. [2014], *Frequent Pattern Mining*, Springer International Publishing, Cham.

[2] Agrawal, R. and Srikant, R. [1995], Mining Sequential Patterns, *in* 'Proceedings of the Eleventh International Conference on Data Engineering', ICDE '95, IEEE Computer Society, Washington, DC, USA, pp. 3–14.

[3] Aigner, W., Miksch, S., Müller, W., Schumann, H. and Tominski, C. [2007], 'Visualizing time-oriented data - A systematic view', *Computers and Graphics (Pergamon)* 31(3), 401–409.

[4] Aigner, W., Miksch, S., Muller, W., Schumann, H. and Tominski, C. [2008], 'Visual methods for analyzing time-oriented data', *Visualization and Computer Graphics, IEEE Transactions on* 14(1), 47–60.

[5] Aigner, W., Miksch, S., Schumann, H. and Tominski, C. [2011], *Visualization of time-oriented data*, Springer.

[6] Al-Hajj, S., Pike, I., Riecke, B. and Fisher, B. [2013], Visual Analytics for Public Health: Supporting Knowledge Construction and Decision-Making, *in* 'System Sciences (HICSS), 2013 46th Hawaii International Conference on', pp. 2416–2423.

[7] Allen, R. B. [2005], A focus-context browser for multiple timelines, *in* 'Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries - JCDL '05', pp. 260–261.

[8] Alsallakh, B., Aigner, W., Miksch, S. and Groller, M. E. [2012], 'Reinventing the Contingency Wheel: Scalable Visual Analytics of Large Categorical Data', *Visualization and Computer Graphics, IEEE Transactions on* 18(12), 2849–2858.

[9] Alvares, L. O., Bogorny, V., Kuijpers, B., de Macedo, J. A. F., Moelans, B. and Vaisman, A. [2007], A model for enriching trajectories with semantic geographical information, *in* 'Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems - GIS '07', GIS '07, ACM, New York, NY, USA, p. 1.

[10] André, P., Wilson, M. L., Russell, A., Smith, D. A., Owens, A. and Schraefel, M. C. [2007], Continuum: Designing Timelines for Hierarchies, Relationships and Scale, *in* 'Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology', UIST '07, ACM, New York, NY, USA, pp. 101–110.

[11] Andrienko, G. and Andrienko, N. [2008], Spatio-temporal aggregation for visual analysis of movements, *in* 'VAST'08 - IEEE Symposium on Visual Analytics Science and Technology, Proceedings', IEEE, pp. 51–58.

[12] Andrienko, G., Andrienko, N., Bak, P., Keim, D. and Wrobel, S. [2013], *Visual analytics of movement*, Springer.

[13] Andrienko, G., Andrienko, N., Hurter, C., Rinzivillo, S. and Wrobel, S. [2013], 'Scalable analysis of movement data for extracting and exploring significant places', *IEEE Transactions on Visualization and Computer Graphics* 19(7), 1078–1094.

[14] Andrienko, G., Andrienko, N., Rinzivillo, S., Nanni, M., Pedreschi, D. and Giannotti, F. [2009], Interactive visual clustering of large collections of trajectories, *in* 'VAST 09 - IEEE Symposium on Visual Analytics Science and Technology, Proceedings', IEEE, pp. 273–274.

[15] Andrienko, G., Andrienko, N., Wrobel, S. and Augustin, S. [2007], 'Visual analytics tools for analysis of movement data', *ACM SIGKDD Explorations Newsletter* 9(2), 38–46.

[16] Andrienko, N. and Andrienko, G. [2013], 'A visual analytics framework for spatio-temporal analysis and modelling', *Data Mining and Knowledge Discovery* 27(1), 55–83.

[17] Andrienko, N., Andrienko, G. and Fuchs, G. [2013], 'Towards Privacy-Preserving Semantic Mobility Analysis', *In EuroVis workshop on visual analytics. The Eurographics Association* pp. 19–23.

[18] Andrienko, N., Andrienko, G., Fuchs, G. and Jankowski, P. [2016], 'Scalable and privacy-respectful interactive discovery of place semantics from human mobility traces', *Information Visualization* 15(2), 117–153.

[19] Andrienko, N., Andrienko, G. and Gatalsky, P. [2000], Visualization of spatio-temporal information in the Internet, *in* 'Proceedings - International Workshop on Database and Expert Systems Applications, DEXA', Vol. 2000-Janua, pp. 577–585.

[20] Asgary, A., Ghaffari, A. and Levy, J. [2010], 'Spatial and temporal analyses of structural fire incidents and their causes: A case of Toronto, Canada', *Fire Safety Journal* 45(1), 44–57.

[21] Bade, R., Schlechtweg, S. and Miksch, S. [2004], Connecting Time-oriented Data and Information to a Coherent Interactive Visualization, *in* 'Proceedings of the 2004 conference on Human factors in computing systems - CHI '04', CHI '04, ACM, New York, NY, USA, pp. 105–112.

[22] Bailey, M. and Cunningham, S. [2008], *Introduction to Computer Graphics*, Addison-Wesley.

[23] Barata, G., Nicolau, H. and Gonçalves, D. [2012], 'AppInsight: What have I been doing?', *Proceedings of the International Working Conference on Advanced Visual Interfaces - AVI '12* pp. 465–472.

[24] Basole, R. C., Braunstein, M. L., Kumar, V., Park, H., Kahng, M., Chau, D. H. P., Tamersoy, A., Hirsh, D. A., Serban, N., Bost, J., Lesnick, B., Schissel, B. L. and Thompson, M. [2015], 'Understanding variations in pediatric asthma care processes in the emergency department using visual analytics', *Journal of the American Medical Informatics Association* 22(2), 318–323.

[25] Baur, D., Seiffert, F., Sedlmair, M. and Boring, S. [2010], 'The Streams of Our Lives: Visualizing Listening Histories in Context', *IEEE Transactions on Visualization and Computer Graphics* 16(6), 1119–1128.

[26] Beecham, R., Wood, J. and Bowerman, A. [2014], 'Studying commuting behaviours using collaborative visual analytics', *Computers, Environment and Urban Systems* 47, 5–15.

[27] Beel, J., Gipp, B., Langer, S. and Breitinger, C. [2016], 'Research-paper recommender systems: a literature survey', *International Journal on Digital Libraries* 17(4), 305–338.

[28] Berger, A., Caruana, R., Cohn, D., Freitag, D. and Mittal, V. [2000], Bridging the lexical chasm: statistical approaches to answer-finding, *in* 'Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval', ACM Press, New York, New York, USA, pp. 192–199.

[29] Bibeault, B. and Katz, Y. [2010], *jQuery In Action*, Manning Publications, CT, USA.

[30] Bögl, M., Aigner, W., Filzmoser, P., Gschwandtner, T., Lammarsch, T., Miksch, S. and Rind, A. [2014], Visual Analytics Methods to Guide Diagnostics for Time Series Model Predictions, *in* 'Proceedings of the 2014 IEEE VIS Workshop on Visualization for Predictive Analytics', Vol. 1, Paris, France.

[31] Borgo, R., Abdul-Rahman, A., Mohamed, F., Grant, P. W., Reppa, I., Floridi, L. and Chen, M. [2012], 'An empirical study on using visual embellishments in visualization', *Visualization and Computer Graphics, IEEE Transactions on* 18(12), 2759–2768.

[32] Borgo, R., Kehrer, J., Chung, D. H. S., Maguire, E., Laramee, R. S., Hauser, H., Ward, M. and Chen, M. [2013], Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications, *in* 'Eurographics State of the Art Reports', Leipzig, Germany, pp. 39–63.

[33] Bostock, M., Ogievetsky, V. and Heer, J. [2011], 'D Data-Driven Documents', *Visualization and Computer Graphics, IEEE Transactions on* 17(12), 2301–2309.

[34] Brodlie, K., AllendesOsorio, R. and Lopes, A. [2012], A Review of Uncertainty in Data Visualization, *in* J. Dill, R. Earnshaw, D. Kasik, J. Vince and P. C. Wong, eds, 'Expanding the Frontiers of Visual Analytics and Visualization', Springer London, pp. 81–109.

[35] Brunsdon, C., Corcoran, J. and Higgs, G. [2007], 'Visualising space and time in crime patterns: A comparison of methods', *Computers, Environment and Urban Systems* 31(1), 52–75.

[36] Bryant, F. B., Smart, C. M. and King, S. P. [2005], 'Using the Past to Enhance the Present: Boosting Happiness Through Positive Reminiscence', *Journal of Happiness Studies* 6(3), 227–260.

[37] Buchin, M., Dodge, S. and Speckmann, B. [2014], 'Similarity of trajectories taking into account geographic context', *Journal of Spatial Information Science* 9(9), 101–124.

[38] Buhmann, J., Fellner, D., Held, M., Ketterer, J. and Puzicha, J. [1998], 'Dithered Color Quantization', *Computer Graphics Forum* 17(3), 219–231.

[39] Buono, P., Aris, A., Plaisant, C., Khella, A. and Shneiderman, B. [2005], Interactive pattern search in time series, *in* 'Electronic Imaging 2005', International Society for Optics and Photonics, pp. 175–186.

[40] Burkhardt, D., Stab, C., Steiger, M., Breyer, M. and Nazemi, K. [2012], Interactive Exploration System: A User-Centered Interaction Approach in Semantics Visualizations, *in* 'Cyberworlds (CW), 2012 International Conference on', pp. 261–267.

[41] Byrne, D., Kelliher, A. and Jones, G. J. [2011], Life Editing: Third-Party Perspectives on Lifelog Content, *in* 'Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11', ACM Press, New York, New York, USA, pp. 1501–1510.

[42] Carlis, J. V. and Konstan, J. A. [1998], 'Interactive visualization of serial periodic data', *Proceedings of the 11th annual ACM symposium on User interface software and technology - UIST '98* pp. 29–38.

[43] Chen, C. and Yang, J. [2011], Essence of Two-Dimensional Principal Component Analysis and Its Generalization: Multi-dimensional PCA, *in* 'Innovations in Bio-inspired Computing and Applications (IBICA), 2011 Second International Conference on', pp. 85–90.

[44] Chen, L., Özsu, M. T. and Oria, V. [2004], Symbolic Representation and Retrieval of Moving Object Trajectories, *in* 'Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval', MIR '04, ACM, New York, NY, USA, pp. 227–234.

[45] Chen, S., Yuan, X., Wang, Z., Guo, C., Liang, J., Wang, Z., Zhang, X. L. and Zhang, J. [2016], 'Interactive Visual Discovering of Movement Patterns from Sparsely Sampled Geo-tagged Social Media Data', *IEEE Transactions on Visualization and Computer Graphics* 22(1), 270–279.

[46] Chen, Z., Shen, H. T. and Zhou, X. [2011], Discovering popular routes from trajectories, *in* 'Proceedings - International Conference on Data Engineering', ICDE '11, IEEE Computer Society, Washington, DC, USA, pp. 900–911.

[47] Chin, J. P., Diehl, V. A. and Norman, L. K. [1988], Development of an instrument measuring user satisfaction of the human-computer interface, *in* 'Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '88', ACM Press, New York, New York, USA, pp. 213–218.

[48] Choo, J. and Park, H. [2013], 'Customizing Computational Methods for Visual Analytics with Big Data', *Computer Graphics and Applications, IEEE* 33(4), 22–28.

[49] Chung, D. H. S., Legg, P. A., Parry, M. L., Bown, R., Griffiths, I. W., Laramee, R. S. and Chen, M. [2013], 'Glyph Sorting: Interactive Visualization for Multi-dimensional Data', *Information Visualization* 14(1), 76–90.

[50] Cichosz, S. L., Johansen, M. D. and Hejlesen, O. [2015], 'Toward Big Data Analytics: Review of Predictive Models in Management of Diabetes and Its Complications', *Journal of Diabetes Science and Technology* 10(12), 1–8.

[51] Cockburn, A., Karlson, A. and Bederson, B. B. [2008], 'A review of overview+ detail, zooming, and focus+ context interfaces', *ACM Computing Surveys (CSUR)* 41(1), 1–42.

[52] Collins, C., Carpendale, S. and Penn, G. [2009], 'DocuBurst: Visualizing document content using language structure', *Computer Graphics Forum* 28(3), 1039–1046.

[53] Consolvo, S., McDonald, D. W. and Landay, J. A. [2009], Theory-driven design strategies for technologies that support behavior change in everyday life, *in* 'Proceedings of the 27th international conference on Human factors in computing systems - CHI 09', number May, ACM Press, New York, New York, USA, pp. 405–414.

[54] Correa, C., Chan, Y.-H. and Ma, K.-L. [2009], A framework for uncertainty-aware visual analytics, *in* 'Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on', pp. 51–58.

[55] Cressie, N. and Wikle, C. K. [2011], *Statistics for spatio-temporal data*, John Wiley & Sons.

[56] Cui, W., Wu, Y., Liu, S., Wei, F., Zhou, M. X. and Qu, H. [2010], Context preserving dynamic word cloud visualization, *in* 'Pacific Visualization Symposium (PacificVis), 2010 IEEE', IEEE, pp. 121–128.

[57] Dachselt, R. and Weiland, M. [2006], TimeZoom: A Flexible Detail and Context Timeline, *in* 'CHI '06 Extended Abstracts on Human Factors in Computing Systems', CHI EA '06, ACM, New York, NY, USA, pp. 682–687.

[58] Deitrick, S. A. [2007], Uncertainty visualization and decision making: Does visualizing uncertain information change decisions, *in* 'Proceedings of the XXIII International Cartographic Conference', pp. 4–10.

[59] Deng, Z., Zhao, Y., Parvinzamir, F., Zhao, X., Wei, H., Liu, M., Zhang, X., Dong, F., Liu, E. and Clapworthy, G. [2016], MyHealthAvatar: A lifetime visual analytics companion for citizen well-being, *in* 'Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)', Vol. 9654, pp. 345–356.

[60] Dias, R., Fonseca, M. and Gonçalves, D. [2012], 'Interactive exploration of music listening histories', *Proceedings of the Workshop on Advanced Visual Interfaces AVI* pp. 415–422.

[61] D.Manning, C., Prabhakar, R. and Schtze, H. [2008], *An Introduction to Information Retrieval*, Vol. 1, Cambridge University Press, Cambridge.

[62] Dodge, S. [2011], Exploring Movement Using Similarity Analysis, PhD thesis, Mathematisch-naturwissenschaftlichen Fakultät der Universität Zürich.

[63] Draper, G., Livnat, Y. and Riesenfeld, R. F. [2009], 'A survey of radial methods for information visualization', *Visualization and Computer Graphics, IEEE Transactions on* 15(5), 759–776.

[64] Drossis, G. and Grammenos, D. [2013], 3D Visualization and Multimodal Interaction with Temporal Information Using Timelines, *in* P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson and M. Winckler, eds, 'Human-Computer Interaction-INTERACT 2013', Vol. 8119 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 214–231.

[65] Dunne, C., Shneiderman, B., Gove, R., Klavans, J. and Dorr, B. [2012], 'Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization', *Journal of the American Society for Information Science and Technology* 63(12), 2351–2369.

[66] Ellard, C. G. [1995], Context and consciousness, *in* B. A. Nardi, ed., 'Behavioral and Brain Sciences', Vol. 18, Massachusetts Institute of Technology, Cambridge, MA, USA, chapter Studying C, pp. 69–102.

[67] Ellis, G. and Dix, A. [2007], 'A taxonomy of clutter reduction for information visualisation', *Visualization and Computer Graphics, IEEE Transactions on* 13(6), 1216–1223.

[68] Endert, A., Fiaux, P. and North, C. [2012], 'Semantic Interaction for Sensemaking: Inferring Analytical Reasoning for Model Steering', *Visualization and Computer Graphics, IEEE Transactions on* 18(12), 2879–2888.

[69] Ester, M., Kriegel, H. P., Sander, J. and Xu, X. [1996], A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, *in* 'Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining', KDD'96, AAAI Press, pp. 226–231.

[70] Euler, L. [n.d.], 'The Euler Timelines'.
URL: *http://eulerarchive.maa.org/historica/euler-timeline.html*

[71] Fails, J. A., Karlson, A., Shahamat, L. and Shneiderman, B. [2006], A Visual Interface for Multivariate Temporal Data: Finding Patterns of Events across

Multiple Histories, *in* 'IEEE Symposium on Visual Analytics Science and Technology 2006, VAST 2006 - Proceedings', IEEE, pp. 167–174.

[72] Fan, C., Forlizzi, J. and Dey, A. [2012], 'A Spark Of Activity: Exploring Informative Art As Visualization For Physical Activity', *The 14th International Conference on Ubiquitous Computing* pp. 81–84.

[73] Fekete, J. D. [2013], 'Visual Analytics Infrastructures: From Data Management to Exploration', *Computer* 46(7), 22–29.

[74] Fellner, D. W. and Helmberg, C. [1993], 'Robust Rendering of General Ellipses and Elliptical Arcs', *ACM Transactions on Graphics* 12(3), 251–276.

[75] Few, S. [2017], 'Data Visualization Effectiveness Profile', *Perceptual Edge* pp. 1–11.

[76] Filatova, E. and Hatzivassiloglou, V. [2004], 'Event-based extractive summarization', *Proceedings of ACL Workshop on Summarization* pp. 104–111.

[77] Flanagan, D. [2011], *JavaScript: The Definitive Guide*, 6th edn, O'Reilly, California, US.

[78] Frank, A. U. [1998], 'Different Types of "Times " in GIS', *Spatial and Temporal reasonning in Geographic Information Systems* pp. 40–61.

[79] Frohlich, D., Kuchinsky, A., Pering, C., Don, A. and Ariss, S. [2002], Requirements for Photoware, *in* 'Proceedings of the 2002 ACM conference on Computer supported cooperative work - CSCW '02', CSCW '02, ACM, New York, NY, USA, pp. 166–175.

[80] Frost, J. H. and Massagli, M. P. [2008], 'Social Uses of Personal Health Information Within PatientsLikeMe, an Online Patient Community: What Can Happen When Patients Have Access to One Another's Data', *Journal of Medical Internet Research* 10(3), e15.

[81] Fruchterman, T. M. J. and Reingold, E. M. [1991], 'Graph Drawing by Forc-directed Placement', *Software-Practice and Experience* 21(1), 1129–1164.

[82] Furletti, B., Cintia, P., Renso, C. and Spinsanti, L. [2013], Inferring human activities from GPS tracks, *in* 'Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing - UrbComp '13', UrbComp '13, ACM, New York, NY, USA, pp. 5:1–5:8.

[83] Giannotti, F., Nanni, M., Pinelli, F. and Pedreschi, D. [2007], Trajectory pattern mining, *in* 'Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07', ACM, New York, NY, USA, pp. 330–339.

[84] Gorg, C., Liu, Z., Kihm, J., Choo, J., Park, H. and Stasko, J. [2013], 'Combining computational analyses and interactive visualization for document exploration and sensemaking in jigsaw', *Visualization and Computer Graphics, IEEE Transactions on* 19(10), 1646–1663.

[85] Gotz, D. and Stavropoulos, H. [2014], 'DecisionFlow: Visual Analytics for High-Dimensional Temporal Event Sequence Data', *IEEE Transactions on Visualization and Computer Graphics* 20(12), 1783–1792.

[86] Green, T. M., Ribarsky, W. and Fisher, B. [2008], Visual Analytics for Complex Concepts Using a Human Cognition Model, *in* 'VAST'08 - IEEE Symposium on Visual Analytics Science and Technology, Proceedings', IEEE, pp. 91–98.

[87] Griethe, H. and Schumann, H. [2006], The Visualization of Uncertain Data: Methods and Problems, *in* S. Schulze, Thomas and Horton, Graham and Preim, Bernhard and Schlechtweg, ed., 'Simulation und Visualization', SCS Publishing House, pp. 143–156.

[88] Han, J., Cheng, H., Xin, D. and Yan, X. [2007], 'Frequent Pattern Mining: Current Status and Future Directions', *Data Mining and Knowledge Discovery* 15(1), 55–86.

[89] Harper, B., Slaughter, L. and Norman, K. [1998], Questionnaire administration via the WWW : A validation & reliability study for a user satisfaction questionnaire, *in* 'WebNet', Vol. 97, Toronto, Canada, pp. 1–4.

[90] Harrower, M. and Brewer, C. A. [2003], 'ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps', *The Cartographic Journal* 40(1), 27–37.

[91] Hasan, K. T., Rashed, S., Noori, H., , A. S. and Kabir, A. [2011], Making sense of time: timeline visualization for public transport schedule, *in* 'Symposium on HumanComputer Interaction and Information Retrieval', ACM, Cambridge, California, USA, pp. 45–52.

[92] Hilliges, O. and Kirk, D. S. [2009], Getting Sidetracked: Display Design and Occasioning Photo-talk with the Photohelix, *in* 'Proceedings of the 27th international conference on Human factors in computing systems - CHI 09', CHI '09, ACM, New York, NY, USA, pp. 1733–1736.

[93] Hofmann, T. [2001], 'Unsupervised learning by probabilistic latent semantic analysis', *Machine Learning* 42(1-2), 177–196.

[94] Holten, D. [2006], 'Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data', *Visualization and Computer Graphics, IEEE Transactions on* 12(5), 741–748.

[95] Hsieh, H.-F. and Wang, J.-J. [2003], 'Effect of reminiscence therapy on depression in older adults: a systematic review', *International Journal of Nursing Studies* 40(4), 335–345.

[96] Huang, D., Tory, M., Adriel Aseniero, B., Bartram, L., Bateman, S., Carpendale, S., Tang, A. and Woodbury, R. [2015], 'Personal Visualization and Personal Visual Analytics', *IEEE Transactions on Visualization and Computer Graphics* 21(3), 420–433.

[97] Inselberg, A. and Dimsdale, B. [1990], Parallel coordinates: a tool for visualizing multi-dimensional\ngeometry, *in* 'Proceedings of the First IEEE Conference on Visualization: Visualization '90', IEEE, pp. 361–378.

[98] Ivezić, Ž., Connolly, A. J., VanderPlas, J. T. and Gray, A. [2014], *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data*, Princeton University Press.

[99] Jain, A. K. [2010], 'Data clustering: 50 years beyond K-means', *Pattern Recognition Letters* 31(8), 651–666.

[100] Janssens, D., Wets, G., De Beuckeleer, E. and Vanhoof, K. [2004], Collecting activity-travel diary data by means of a new computer-assisted data collection tool, *in* '11th European Concurrent Engineering Conference 2004: Worldwide Partnerships and Mergers', Hasselt, Belgium, pp. 85–89.

[101] Javed, W. and Elmqvist, N. [2010], Stack zooming for multi-focus interaction in time-series data visualization, *in* 'Pacific Visualization Symposium (PacificVis), 2010 IEEE', IEEE, pp. 33–40.

[102] Jin, H. and Liu, H. [2009], Research on Visualization Techniques in Data Mining, *in* 'Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference on', pp. 1–3.

[103] Jo, H. and hee Ryu, J. [2010], 'Placegram: A Diagrammatic Map for Personal Geotagged Data Browsing.', *IEEE Trans. Vis. Comput. Graph.* 16(2), 221–234.

[104] Johnson, B. and Shneiderman, B. [1991], Tree-maps: A space-filling approach to the visualization of hierarchical information structures, *in* 'Visualization, 1991. Visualization'91, Proceedings., IEEE Conference on', IEEE, pp. 284–291.

[105] Joy, K. [2009], Massive Data Visualization: A Survey, *in* T. Möller, B. Hamann and R. Russell, eds, 'Mathematical Foundations of Scientific Visualization, Computer Graphics, and Massive Data Exploration', Mathematics and Visualization, Springer Berlin Heidelberg, pp. 285–302.

[106] Kalnikaite, V. and Whittaker, S. [2011], 'A saunter down memory lane: Digital reflection on personal mementos', *International Journal of Human Computer Studies* 69(5), 298–310.

[107] Kandel, S., Paepcke, A., Hellerstein, J. M. and Heer, J. [2012], 'Enterprise Data Analysis and Visualization: An Interview Study', *Visualization and Computer Graphics, IEEE Transactions on* 18(12), 2917–2926.

[108] Kapler, T. and Wright, W. [2004], GeoTime information visualization, *in* 'Proceedings - IEEE Symposium on Information Visualization, INFO VIS', pp. 25–32.

[109] Kayyali, B., Knott, D. and Kuiken, S. V. [2013], 'The big-data revolution in US health care : Accelerating value and innovation'.
URL: *https://digitalstrategy.nl/wp-content/uploads/E2-2013.04-The-big-data-revolution-in-US-health-care-Accelerating-value-and-innovation.pdf*

[110] Keim, D. a., Mansmann, F., Thomas, J. and Keim, D. [2010], 'Visual Analytics : How Much Visualization and How Much Analytics?', *ACM SIGKDD Explorations Newsletter* 11(2), 5–8.

[111] Keim, D., Mansmann, F., Schneidewind, J., Thomas, J. and Ziegler, H. [2008], Visual Analytics: Scope and Challenges, *in* S. Simoff, M. Bohlen and A. Mazeika, eds, 'Visual Data Mining', Vol. 4404 of *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, pp. 76–90.

[112] Keim, D., Mansmann, F., Schneidewind, J. and Ziegler, H. [2006], Challenges in Visual Data Analysis, *in* 'Tenth International Conference on Information Visualisation (IV'06)', IEEE, pp. 9–16.

[113] Keim, D., Mansmann, F., Stoffel, A. and Ziegler, H. [2009], Visual Analytics, *in* L. LIU and M. T. ÖZSU, eds, 'Encyclopedia of Database Systems', Springer US, Boston, MA, pp. 3341–3346.

[114] Kindlmann, G. and Durkin, J. W. [1998], Semi-Automatic Generation of Transfer Functions for Direct Volume Rendering, Master's thesis, Cornell University.

[115] Klimov, D., Shahar, Y. and Taieb-Maimon, M. [2010], 'Intelligent visualization and exploration of time-oriented data of multiple patients', *Artificial Intelligence in Medicine* 49(1), 11–31.

[116] Kobbelt, L., Stamminger, M. and Seidel, H.-P. [2008], 'Using Subdivision on Hierarchical Data to Reconstruct Radiosity Distribution', *Computer Graphics Forum* 16(3), C347–C355.

[117] Koh, L. C., Slingsby, A., Dykes, J. and Kam, T. S. [2011], Developing and Applying a User-Centered Model for the Design and Implementation of Information Visualization Tools, *in* 'Information Visualisation (IV), 2011 15th International Conference on', pp. 90–95.

[118] Kohlhammer, J., Keim, D., Pohl, M., Santucci, G. and Andrienko, G. [2011], 'Solving Problems with Visual Analytics', *Procedia Computer Science* 7, 117–120.

[119] Krueger, R., Thom, D. and Ertl, T. [2014], Visual analysis of movement behavior using web data for context enrichment, *in* I. Fujishiro, U. Brandes, H. Hagen and S. Takahashi, eds, 'IEEE Pacific Visualization Symposium', IEEE Computer Society, pp. 193–200.

[120] Krueger, R., Thom, D. and Ertl, T. [2015], 'Semantic Enrichment of Movement Behavior with Foursquare - A Visual Analytics Approach', *IEEE Transactions on Visualization and Computer Graphics* 21(8), 903–915.

[121] Kumar, P., Krishna, P. R., Limited, I. and Raju, S. B. [2012], *Pattern discovery using sequence data mining: applications and studies*, IGI Global, Hershey, PA, USA.

[122] Kurzhals, K., Heimerl, F. and Weiskopf, D. [2014], ISeeCube: Visual Analysis of Gaze Data for Video, *in* 'Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '14', ETRA '14, ACM, New York, NY, USA, pp. 351–358.

[123] Lafortune, E. P. F., Foo, S.-C., Torrance, K. E. and Greenberg, D. P. [1997], Non-Linear Approximation of Reflectance Functions, *in* 'Proceedings of the 24th annual conference on Computer graphics and interactive techniques - SIGGRAPH '97', Vol. 31, pp. 117–126.

[124] Lam, H., Bertini, E., Isenberg, P., Plaisant, C. and Carpendale, S. [2012], 'Empirical studies in information visualization: Seven scenarios', *IEEE Transactions on Visualization and Computer Graphics* 18(9), 1520–1536.

[125] Leskovec, J., Rajaraman, A. and Ullman, J. D. [2014], *Mining of Massive Datasets*, 2 edn, Cambridge University Press.

[126] Lesselroth, B. J. and Pieczkiewicz, D. S. [2011], Data visualization strategies for the electronic health record, *in* 'Data Visualization Strategies for the Electronic Health Record', Vol. 16, Nova Science Publishers, Portland, Oregon, United States, chapter 3, pp. 107–140.

[127] Levenshtein, V. [1966], 'Binary codes capable of correcting deletions, insertions, and reversals', *Cybernetics and Control Theory* 10(8), 707–710.

[128] Levoy, M. [1988], Display of Surfaces from Volume Data, PhD thesis, University of North Carolina at Chapel Hill.

[129] Lewis, C. N. [1971], 'Reminiscing and Self-Concept in Old Age', *Journal of Gerontology* 26(2), 240–243.

[130] Lewis, J. R. [1995], 'IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use', *International Journal of Human-Computer Interaction* 7(1), 57–78.

[131] Li, I., Dey, A. and Forlizzi, J. [2011], 'Understanding My Data, Myself: Supporting Self-Reflection with Ubicomp Technologies', *Proceedings of the 13th international conference on Ubiquitous computing* pp. 405–414.

[132] Li, I., Dey, A. K. and Forlizzi, J. [2012], 'Using context to reveal factors that affect physical activity', *ACM Transactions on Computer-Human Interaction* 19(1), 1–21.

[133] Li, J., Fan, Q. and Zhang, K. [2007], 'Keyword extraction based on tf/idf for Chinese news document', *Wuhan University Journal of Natural Sciences* 12(5), 917–921.

[134] Li, X., Han, J., Lee, J.-G. and Gonzalez, H. [2007], Traffic Density-Based Discovery of Hot Routes in Road Networks., *in* D. Papadias, D. Zhang and G. Kollios, eds, 'Proceedings of the 10th International Symposium on Spatial and Temporal Databases (SSTD '07)', Vol. 4605 of *Lecture Notes in Computer Science*, Springer, pp. 441–459.

[135] Li, Z. [2014], *Spatiotemporal Pattern Mining: Algorithms and Applications*, Springer International Publishing, Cham, chapter 12, pp. 283–306.

[136] Li, Z., Wang, J. and Han, J. [2015], 'EPeriodicity: Mining event periodicity from incomplete observations', *IEEE Transactions on Knowledge and Data Engineering* 27(5), 1219–1232.

[137] Lin, M.-Y., Lee, P.-Y. and Hsueh, S.-C. [2012], Apriori-based frequent itemset mining algorithms on MapReduce, *in* 'Proceedings of the 6th International Conference on Ubiquitous Information Management and Communication - ICUIMC '12', ACM Press, New York, New York, USA, p. 1.

[138] Lipsa, D. [2011], 'Techniques for Large Data Visualization', *International Journal of Research and Reviews in Computer Science* 2(2), 1–18.

[139] Liu, F., Janssens, D., Wets, G. and Cools, M. [2013], 'Annotating mobile phone location data with activity purposes using machine learning algorithms', *Expert Systems with Applications* 40(8), 3299–3311.

[140] Liu, S., Cui, W., Wu, Y. and Liu, M. [2014], 'A survey on information visualization: recent advances and challenges', *The Visual Computer* 30(12), 1373–1393.

[141] Loorak, M. H., Perin, C., Kamal, N., Hill, M. and Carpendale, S. [2016], TimeSpan: Using Visualization to Explore Temporal Multi-dimensional Data of Stroke Patients, *in* 'IEEE Transactions on Visualization and Computer Graphics', Vol. 22, pp. 409–418.

[142] Lorensen, W. E. and Cline, H. E. [1987], Marching Cubes: A High Resolution 3D Surface Construction Algorithm, *in* 'Computer Graphics (Proceedings of SIGGRAPH 87)', Vol. 21, pp. 163–169.

[143] Lous, Y. L. [1990], 'Report on the First Eurographics Workshop on Visualization in Scientific Computing', *Computer Graphics Forum* 9(5), 371–372.

[144] Mabroukeh, N. R. and Ezeife, C. I. [2010], 'A Taxonomy of Sequential Pattern Mining Algorithms', *ACM Computing Surveys* 43(1), 1–41.

[145] MacEachren, A. M., Roth, R. E., O'Brien, J., Li, B., Swingley, D. and Gahegan, M. [2012], 'Visual Semiotics & Uncertainty Visualization: An Empirical Study', *Visualization and Computer Graphics, IEEE Transactions on* 18(12), 2496–2505.

[146] Mackinlay, J. D., Robertson, G. G. and Card, S. K. [1991], The Perspective Wall: Detail and Context Smoothly Integrated, *in* 'Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology - CHI '91', CHI '91, ACM, New York, NY, USA, pp. 173–176.

[147] Max, N. [1995], 'Optical Models for Direct Volume Rendering', *IEEE Transactions on Visualization and Computer Graphics* 1(2), 99–108.

[148] McLachlan, P., Munzner, T., Koutsofios, E. and North, S. [2008], LiveRAC: interactive visual exploration of system management time-series data, *in* 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', ACM, pp. 1483–1492.

[149] Mennis, J. and Guo, D. [2009], 'Spatial data mining and geographic knowledge discovery-An introduction', *Computers, Environment and Urban Systems* 33(6), 403–408.

[150] Meyer, M., Munzner, T. and Pfister, H. [2009], 'MizBee: a multiscale synteny browser', *Visualization and Computer Graphics, IEEE Transactions on* 15(6), 897–904.

[151] Monroe, M., Lan, R., Lee, H., Plaisant, C. and Shneiderman, B. [2013], Temporal event sequence simplification, *in* 'IEEE Transactions on Visualization and Computer Graphics', Vol. 19, pp. 2227–2236.

[152] Mooney, C. H. and Roddick, J. F. [2013], 'Sequential Pattern Mining – Approaches and Algorithms', *ACM Computing Surveys* 45(2), 1–39.

[153] Nielson, G. and Hamann, B. [1991], The asymptotic decider: resolving the ambiguity in marching cubes, *in* 'Proceeding Visualization '91', IEEE Comput. Soc. Press, pp. 83–91.

[154] Noh, Y.-K., Park, F. and Lee, D. D. [2012], Diffusion Decision Making for Adaptive k-Nearest Neighbor Classification, *in* 'Advances in Neural Information Processing Systems', pp. 1925–1933.

[155] Nunes, M., Greenberg, S. and Neustaedter, C. [2008], Sharing digital photographs in the home through physical mementos, souvenirs, and keepsakes, *in* 'Proceedings of the 7th ACM conference on Designing interactive systems - DIS '08', DIS '08, ACM, New York, NY, USA, pp. 250–260.

[156] Parent, C., Pelekis, N., Theodoridis, Y., Yan, Z., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., Damiani, M. L., Gkoulalas-Divanis, A. and Macedo, J. [2013], 'Semantic trajectories modeling and analysis', *ACM Computing Surveys* 45(4), 1–32.

[157] Parvinzamir, F., Zhao, Y., Deng, Z., Zhao, X., Ersotelos, N., Dong, F., Liu, E. and Clapworthy, G. [2015], MyHealthAvatar: A Case Study of Web-Based Interactive Visual Analytics of Lifestyle Data, *in* '2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing', IEEE, pp. 2335–2339.

[158] Petrelli, D., Bowen, S. and Whittaker, S. [2014], 'Photo mementos: Designing digital media to represent ourselves at home', *International Journal of Human Computer Studies* 72(3), 320–336.

[159] Petrelli, D. and Whittaker, S. [2010], 'Family memories in the home: Contrasting physical and digital mementos', *Personal and Ubiquitous Computing* 14(2), 153–169.

[160] Petrelli, D., Whittaker, S. and Brockmeier, J. [2008], AutoTypography, *in* 'Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08', CHI '08, ACM Press, New York, New York, USA, pp. 53–62.

[161] Pirolli, P. and Card, S. [2005], 'The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis', *Proceedings of International Conference on Intelligence Analysis* 2005, 2–4.

[162] Plaisant, C., Heller, D., Li, J., Shneiderman, B., Mushlin, R. and Karat, J. [1998], 'Visualizing medical records with LifeLines', *CHI 98 conference summary on Human factors in computing systems CHI 98* (May 1997), 28–29.

[163] Plaisant, C., Milash, B., Rose, A., Widoff, S. and Shneiderman, B. [1996], LifeLines: visualizing personal histories, *in* 'Proceedings of the SIGCHI conference on Human factors in computing systems common ground - CHI '96', ACM Press, New York, New York, USA, pp. 221–227.

[164] Plaisant, C., Mushlin, R., Snyder, A., Li, J., Heller, D. and Shneiderman, B. [1998], 'LifeLines: using visualization to enhance navigation and analysis of patient records', *Proceedings / AMIA ...Annual Symposium.AMIA Symposium* pp. 76–80.

[165] Plug, C., Xia, J. and Caulfield, C. [2011], 'Spatial and temporal visualisation techniques for crash analysis', *Accident Analysis and Prevention* 43(6), 1937–1946.

[166] Pöthkow, K. and Hege, H.-C. [2013], 'Nonparametric Models for Uncertainty Visualization', *Computer Graphics Forum* 32(3pt2), 131–140.

[167] Potter, K., Gerber, S. and Anderson, E. W. [2013], 'Visualization of Uncertainty without a Mean', *Computer Graphics and Applications, IEEE* 33(1), 75–79.

[168] Pousman, Z., Stasko, J. T. and Mateas, M. [2007], 'Casual information visualization: Depictions of data in everyday life', *IEEE Transactions on Visualization and Computer Graphics* 13(6), 1145–1152.

[169] Pressman, R. S. and Maxim, B. R. [2010], *Software engineering: A Practitioner's Approach*, 7th edn, McGraw-Hill, New York, NY, USA.

[170] Rajaraman, A. and Ullman, J. D. [2011], *Mining of Massive Datasets*, Cambridge University Press, New York, NY, USA.

[171] Reddy, C. K. and Aggarwal, C. C. [2016], *Healthcare Data Analytics*, Chapman & Hall/CRC, Boca Raton, FL, USA.

[172] Renso, C., Baglioni, M., de Macedo, J. A. F., Trasarti, R. and Wachowicz, M. [2013], 'How you move reveals who you are: Understanding human behavior by analyzing trajectory data', *Knowledge and Information Systems* 37(2), 331–362.

[173] Reumers, S., Liu, F., Janssens, D., Cools, M. and Wets, G. [2013*a*], 'Semantic Annotation of Global Positioning System Traces', *Transportation Research Record: Journal of the Transportation Research Board* 2383(2383), 35–43.

[174] Reumers, S., Liu, F., Janssens, D., Cools, M. and Wets, G. [2013*b*], 'Semantic Annotation of GPS Traces: Activity Type Inference', *92nd Annual Meeting of the Transportation Research Board* 32, 1–14.

[175] Riehmann, P., Hanfler, M. and Froehlich, B. [2005], Interactive Sankey Diagrams, *in* 'Proc. IEEE Symposium on Infomation Visualization, InfoVis 2005', IEEE, pp. 233–240.

[176] Rind, A. [2013], 'Interactive Information Visualization to Explore and Query Electronic Health Records', *Foundations and Trends in HumanComputer Interaction* 5(3), 207–298.

[177] Ringel, M., Cutrell, E., Dumais, S. and Horvitz, E. [2003], Milestones in time: The value of landmarks in retrieving information from personal stores, *in* 'Human-computer interaction: INTERACT'03; IFIP TC13 International Conference on Human-Computer Interaction, 1st-5th September 2003, Zurich, Switzerland', Zurich, Switzerland, pp. 184–192.

[178] Riveiro, M. [2007], Evaluation of uncertainty visualization techniques for information fusion, *in* 'Information Fusion, 2007 10th International Conference on', IEEE, pp. 1–8.

[179] Robertson, G., Fernandez, R., Fisher, D., Lee, B. and Stasko, J. [2008], 'Effectiveness of Animation in Trend Visualization', *IEEE Transactions on Visualization and Computer Graphics* 14(6), 1325–1332.

[180] Robertson, S. [2004], 'Understanding inverse document frequency: on theoretical arguments for IDF', *Journal of Documentation* 60(5), 503–520.

[181] Roddick, J. F., Roddick, J. F., Society, I. C., Spiliopoulou, M. and Society, I. C. [2002], 'A Survey of Temporal Knowledge Discovery Paradigms and Methods', *Ieee Transactions on Knowledge and Data Engineering* 14, 750–767.

[182] Sacharidis, D., Patroumpas, K., Terrovitis, M., Kantere, V., Potamias, M., Mouratidis, K. and Sellis, T. [2008], On-line discovery of hot motion paths, *in* A. Kemper, ed., 'Proceedings of the 11th International Conference on Extending Database Technology: Advances in Database Technology (EDBT '08)', Vol. 261 of *ACM International Conference Proceeding Series*, ACM, pp. 392–403.

[183] Sander, J., Ester, M., Kriegel, H.-P. and Xu, X. [1998], 'Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications', *Data Mining and Knowledge Discovery* 2(2), 169–194.

[184] Sanderson, M. [2010], *Introduction to Information Retrieval*, Vol. 16, Cambridge University Press, New York, NY, USA.

[185] Sanyal, J., Zhang, S., Bhattacharya, G., Amburn, P. and Moorhead, R. J. [2009], 'A User Study to Compare Four Uncertainty Visualization Methods for 1D and 2D Datasets', *Visualization and Computer Graphics, IEEE Transactions on* 15(6), 1209–1218.

[186] Sedlmair, M., Meyer, M. and Munzner, T. [2012], 'Design Study Methodology: Reflections from the Trenches and the Stacks', *IEEE Transactions on Visualization and Computer Graphics* 18(12), 2431–2440.

[187] Seidel, H.-P. [1993], 'Polar Forms for Geometrically Continuous Spline Curves of Arbitrary Degree', *ACM Transactions on Graphics* 12(1), 1–34.

[188] Shahar, Y. and Cheng, C. [1999], Intelligent visualization and exploration of time-oriented clinical data, *in* 'Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences', Vol. Track4, IEEE Computer, pp. 1–12.

[189] Shahar, Y., Goren-Bar, D., Boaz, D. and Tahan, G. [2006], 'Distributed, Intelligent, Interactive Visualization and Exploration of Time-oriented Clinical Data and Their Abstractions', *Artificial Intelligence in Medicine* 38(2), 115–135.

[190] Shen, Z. and Kwan-Liu, M. [2008], MobiVis: A visualization system for exploring mobile data, *in* 'IEEE Pacific Visualisation Symposium 2008, PacificVis - Proceedings', IEEE, pp. 175–182.

[191] Shneiderman, B. [1996], The eyes have it: a task by data type taxonomy for information visualizations, *in* 'Proceedings 1996 IEEE Symposium on Visual Languages', pp. 336–343.

[192] Shneiderman, B., Plaisant, C. and Hesse, B. W. [2013], 'Improving Healthcare with Interactive Visualization', *IEEE Computer* 46(5), 58–66.

[193] Shrestha, A., Miller, B., Zhu, Y. and Zhao, Y. [2013], Storygraph: Extracting Patterns from Spatio-temporal Data, *in* 'Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics', IDEA '13, ACM, New York, NY, USA, pp. 95–103.

[194] Simons, D. J. and Rensink, R. A. [2005], 'Change blindness: past, present, and future', *Trends in Cognitive Sciences* 9(1), 16–20.

[195] Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F. and Vangenot, C. [2008], 'A conceptual view on trajectories', *Data and Knowledge Engineering* 65(1), 126–146.

[196] Spinsanti, L., Celli, F. and Renso, C. [2010], 'Where you stop is who you are: understanding people's activities by places visited', *The proceedings of Behaviour Monitoring and Interpretation (BMI) workshop* .

[197] Spretke, D., Bak, P., Janetzko, H., Kranstauber, B., Mansmann, F. and Davidson, S. [2011], Exploration Through Enrichment: A Visual Analytics Approach for Animal Movement, *in* 'Proceedings of the 19th ACM SIGSPATIAL International

Conference on Advances in Geographic Information Systems', GIS '11, ACM, New York, NY, USA, pp. 421–424.

[198] Sun, G.-D., Wu, Y.-C., Liang, R.-H. and Liu, S.-X. [2013], 'A Survey of Visual Analytics Techniques and Applications: State-of-the-Art Research and Future Challenges', *Journal of Computer Science and Technology* 28(5), 852–867.

[199] Suntinger, M., Obweger, H., Schiefer, J. and Groller, E. [2008], The event tunnel: Interactive visualization of complex event streams for business process pattern analysis, *in* 'Visualization Symposium, 2008. PacificVIS'08. IEEE Pacific', IEEE, pp. 111–118.

[200] Swinburne, R. [2004], 'Bayes' Theorem', *Revue Philosophique de la France et de l'Etranger* 194(2), 250–251.

[201] Third, A., Kaldoudi, E., Gkotsis, G., Roumeliotis, S., Pafili, K., Domingue, J. and Keynes, M. [2010], Capturing Scientific Knowledge on Medical Risk Factors, *in* 'K-CAP2015: 8th International Conference on Knowledge Capture', ACM, pp. 53–62.

[202] Thiry, E., Lindley, S., Banks, R. and Regan, T. [2013], 'Authoring personal histories: Exploring the timeline as a framework for meaning making', *Proceedings of the 2013 SIGCHI conference on Human Factors in computing systems (CHI 2013)* pp. 1619–1628.

[203] Thomas, J. J. and Cook, K. A. [2006], 'A visual analytics agenda', *Computer Graphics and Applications, IEEE* 26(1), 10–13.

[204] Thudt, A., Baur, D., Huron, S. and Carpendale, S. [2016], 'Visual Mementos: Reflecting Memories with Personal Data', *IEEE Transactions on Visualization and Computer Graphics* 22(1), 369–378.

[205] Tominski, C., Abello, J. and Schumann, H. [2004], Axes-based visualizations with radial layouts, *in* 'Proceedings of the 2004 ACM symposium on Applied computing', ACM, pp. 1242–1247.

[206] Tominski, C., Schulze-Wollgast, P. and Schumann, H. [2005], 3D information visualization for time dependent data on maps, *in* 'Proceedings of the International Conference on Information Visualisation', Vol. 2005, pp. 175–181.

[207] Torres, S. O., Eicher-Miller, H., Boushey, C., Ebert, D. and Maciejewski, R. [2012], Applied Visual Analytics for Exploring the National Health and Nutrition Examination Survey, *in* 'System Science (HICSS), 2012 45th Hawaii International Conference on', pp. 1855–1863.

[208] van den Hoven, E. [2014], 'A future-proof past: Designing for remembering experiences', *Memory Studies* 7(3), 370–384.

[209] Van Kleek, M., Smith, D. A., Packer, H. S., Skinner, J. and Shadbolt, N. R. [2013], Carpé data: supporting serendipitous data integration in personal information management, *in* 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13', ACM Press, New York, New York, USA, pp. 2339–2348.

[210] Van Wijk, J. J. [2013], 'Evaluation: A challenge for visual analytics', *Computer* 46(7), 56–60.

[211] Van Wijk, J. and Van Selow, E. [1999], Cluster and calendar based visualization of time series data, *in* 'Proceedings 1999 IEEE Symposium on Information Visualization (InfoVis'99)', IEEE Comput. Soc, pp. 4–9.

[212] Vapnik V.N. [1998], *Statistical learning theory.*, Vol. 1, Wiley New York.

[213] Victor, B. [2013], 'Drawing Dynamic Visualizations'.
URL: *http://worrydream.com/DrawingDynamicVisualizationsTalkAddendum/*

[214] Viégas, F., Golder, S. and Donath, J. [2006], 'Visualizing email content: portraying relationships from conversational histories', *Proceedings of the SIGCHI conference on Human Factors in computing systems* pp. 979–988.

[215] von Landesberger, T., Bremm, S., Andrienko, N., Andrienko, G. and Tekusova, M. [2012], Visual analytics methods for categoric spatio-temporal data, *in* '2012 IEEE Conference on Visual Analytics Science and Technology (VAST)', IEEE, pp. 183–192.

[216] Von Landesberger, T., Brodkorb, F., Roskosch, P., Andrienko, N., Andrienko, G. and Kerren, A. [2016], MobilityGraphs: Visual Analysis of Mass Mobility Dynamics via Spatio-Temporal Graphs and Clustering, *in* 'IEEE Transactions on Visualization and Computer Graphics', Vol. 22, pp. 11–20.

[217] Vrotsou, K., Ellegård, K. and Cooper, M. [2009], 'Exploring time diaries using semi-automated activity pattern extraction', *electronic International Journal of Time Use Research* 6(1), 1–25.

[218] Vrotsou, K., Johansson, J. and Cooper, M. [2009], 'ActiviTree: Interactive visual exploration of sequences in event-based data using graph similarity', *IEEE Transactions on Visualization and Computer Graphics* 15(6), 945–952.

[219] Wang, T. D., Plaisant, C., Shneiderman, B., Spring, N., Roseman, D., Marchand, G., Mukherjee, V. and Smith, M. [2009], 'Temporal summaries: Supporting temporal categorical searching, aggregation and comparison', *IEEE Transactions on Visualization and Computer Graphics* 15(6), 1049–1056.

[220] Ware, C. [2004], *Information Vizualization: Perception for Design*, second edn, Morgan Kaufman.

[221] Weber, M., Alexa, M. and Muller, W. [2001], Visualizing time-series on spirals, *in* 'IEEE Symposium on Information Visualization, 2001. INFOVIS 2001.', IEEE, pp. 7–13.

[222] Webster, J. D. and McCall, M. E. [1999], 'Reminiscence Functions Across Adulthood: A Replication and Extension', *Journal of Adult Development* 6(1), 73–85.

[223] Wei, H., Wu, S., Zhao, Y., Deng, Z., Ersotelos, N., Parvinzamir, F., Liu, B., Liu, E. and Dong, F. [2016], Data mining, management and visualization in large scientific corpuses, *in* 'Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)', Vol. 9654, pp. 371–379.

[224] Wei, H., Zhao, Y., Wu, S., Deng, Z., Parvinzamir, F., Dong, F., Liu, E. and Clapworthy, G. [2016], Management of Scientific Documents and Visualization of Citation Relationships using Weighted Key Scientific Terms, *in* 'Proceedings of the 5th International Conference on Data Management Technologies and Applications', number October, pp. 135–143.

[225] West, V. L., Borland, D. and Hammond, W. E. [2015], 'Innovative information visualization of electronic health record data: a systematic review.', *Journal of the American Medical Informatics Association : JAMIA* 22(2), 330–339.

[226] Whittaker, S., Bergman, O. and Clough, P. [2010], 'Easy on that trigger dad: A study of long term family photo retrieval', *Personal and Ubiquitous Computing* 14(1), 31–43.

[227] Whittaker, S., Kalnikaite, V., Petrelli, D., Sellen, A., Villar, N., Bergman, O., Clough, P. and Brockmeier, J. [2012], 'Socio-technical lifelogging: Deriving design principles for a future proof digital past', *Human-Computer Interaction* 27(1-2), 37–62.

[228] Wijk, J. J. V. [2005], 'The Value of Visualization', *VIS 05. IEEE Visualization, 2005.* pp. 79–86.

[229] Wills, G. [2012], *Visualizing Time*, Statistics and Computing, Springer New York, New York, NY.

[230] Wittenbrink, C. M., Pang, A. T. and Lodha, S. K. [1996], 'Glyphs for visualizing uncertainty in vector fields', *Visualization and Computer Graphics, IEEE Transactions on* 2(3), 266–279.

[231] Wongsuphasawat, K. and Gotz, D. [2012], 'Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization', *IEEE Transactions on Visualization and Computer Graphics* 18(12), 2659–2668.

[232] Wongsuphasawat, K., Guerra Gómez, J. A., Plaisant, C., Wang, T. D., Taieb-Maimon, M. and Shneiderman, B. [2011], LifeFlow: Visualizing an Overview of Event Sequences, *in* 'Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11', CHI '11, ACM, New York, NY, USA, pp. 1747–1756.

[233] Wood, J., Slingsby, A. and Dykes, J. [2011], 'Visualizing the Dynamics of London's Bicycle-Hire Scheme', *Cartographica: The International Journal for Geographic Information and Geovisualization* 46(4), 239–251.

[234] Wu, H. C., Luk, R. W. P., Wong, K. F. and Kwok, K. L. [2008], 'Interpreting TF-IDF term weights as making relevance decisions', *ACM Transactions on Information Systems* 26(3), 1–37.

[235] Yan, X., Qiao, M., Li, J., Simpson, T. W., Stump, G. M. and Zhang, X. L. [2012], A Work-Centered Visual Analytics Model to Support Engineering Design with

Interactive Visualization and Data-Mining, *in* '2012 45th Hawaii International Conference on System Sciences', IEEE, pp. 1845–1854.

[236] Yan, Z., Chakraborty, D., Parent, C., Spaccapietra, S. and Aberer, K. [2011], SeMiTri, semantic trajectory, trajectory annotation, *in* 'Proceedings of the 14th International Conference on Extending Database Technology - EDBT/ICDT '11', EDBT/ICDT '11, ACM, New York, NY, USA, pp. 259–270.

[237] Yang, J., Ward, M. O. and Rundensteiner, E. A. [2002], Interring: An interactive tool for visually navigating and manipulating hierarchical structures, *in* 'Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on', IEEE, pp. 77–84.

[238] Yu, L., Wu, W., Li, X., Li, G., Ng, W. S., Ng, S. K., Huang, Z., Arunan, A. and Watt, H. M. [2015], IVizTRANS: Interactive visual learning for home and work place detection from massive public transportation data, *in* M. Chen and G. L. Andrienko, eds, '2015 IEEE Conference on Visual Analytics Science and Technology, VAST 2015 - Proceedings', IEEE Computer Society, pp. 49–56.

[239] Zhang, Z., Wang, B., Ahmed, F., Ramakrishnan, I. V., Zhao, R., Viccellio, A. and Mueller, K. [2013], 'The five Ws for information visualization with application to healthcare informatics', *IEEE Transactions on Visualization and Computer Graphics* 19(11), 1895–1910.

[240] Zhao, J., Chevalier, F. and Balakrishnan, R. [2011], Kronominer: using multi-foci navigation for the visual exploration of time-series data, *in* 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', ACM, pp. 1737–1746.

[241] Zhao, Y., Parvinzamir, F., Deng, Z., Wei, H., Zhao, X., Liu, E., Dong, F., Clapworthy, G., Lukoševičius, A., Marozas, V. and Kaldoudi, E. [2016], 'MyHealthAvatar and CARRE: case studies of interactive visualisation for internet-enabled sensor-assisted health monitoring and risk analysis', *IET Networks* 5(5), 114–121.

[242] Zhao, Y., Parvinzamir, F., Wei, H., Liu, E., Deng, Z., Dong, F., Third, A., Lukoševičius, A., Marozas, V., Kaldoudi, E. and Clapworthy, G. [2016], *Visual Analytics for Health Monitoring and Risk Management in CARRE*, Vol. 9654, Springer International Publishing, Cham, pp. 380–391.

[243] Zheng, Y. U. [2015], 'Trajectory Data Mining : An Overview', *ACM Transaction on Intelligent Systems and Technology* 6(3), 1–41.

[244] Zheng, Y., Zhang, L., Xie, X. and Ma, W.-Y. [2009], Mining interesting locations and travel sequences from GPS trajectories, *in* 'Proceedings of the 18th international conference on World wide web - WWW '09', ACM, ACM Press, New York, New York, USA, pp. 791–800.