Context Driven Retrieval Algorithms for Semi-Structured Personal Lifelogs

by

Liadh Kelly, B.Sc. (Hons), M.Sc. (Research)

A dissertation submitted in fulfilment of the requirements for the award of
Doctor of Philosophy (Ph.D.)
to the

Dublin City University
School of Computing
and
Centre for Digital Video Processing

Supervisor: Dr. Gareth J. F. Jones

September 2011

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: ____________ (Candidate) ID No.: ___________ Date: _______

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Advances in digital technologies for information capture combined with massive increases in the capacity of digital storage media mean that it is now possible to capture and store much of one's life experiences in a personal lifelog (PL). Information can be captured from a myriad of personal information devices including desktop computers, mobile phones, digital cameras, video recorders, and various sensors, including GPS, Bluetooth, and biometric devices. The large personal archives that can be captured using these devices create new opportunities such as the chance to gain more details on partially recalled life events, opportunities for self reflection, facilities to share experiences, the potential to find partially remembered facts, etc, but also pose new challenges to the research community, not the least of which is developing effective means of retrieval. This thesis centers on the investigation of the challenges of retrieval in this emerging domain, and the proposal and evaluation of methods and algorithms which seek to meet these challenges.

Methods to integrate implicitly recorded and derived context data types with content-based search in information retrieval (IR) algorithms for PL retrieval are developed. These algorithms focus on the use of an individual's memories of items' content and associated context data and on the use of implicit biometric indicators of items' importance. These novel retrieval algorithms are evaluated over unique multimodal PL collections of 20 months duration. We find support for the use of recalled context data in retrieval using a novel algorithm which accounts for the structure of lifelog collections and user queries. We also find support for the use of individuals' past biometric response associated with lifelog items to locate important items in lifelogs and to re-rank ranked retrieval result lists.

Do mo thuismitheoirí.

(For my parents.)

ACKNOWLEDGEMENTS

Míle buíochas do mo mhaoirseoir, Gareth Jones, as a thuiscint, treoir, comhairle agus tacaíochta i rith an tionscadail.

Buíochas freisin chuig Fondúireacht Eolaíochta Éireann as an deontas chun an tionscadal seo a dhéanamh (deontas 06/RFP/CMS023) agus chuig Microsoft Research as na SenseCams agus bogearraí Digital Memories (deontas 2007-056).

Buíochas do gach éinne sa CDVP, CNGL agus lárionad taighde eile a thug cúnamh, comhairle nó cairdeas dom. Ach go háirithe Daragh, Eamonn agus Yuki.

Míle buíochas ó chroí do mo thuistí, Eoin, Cle, Lor, Nan, AnneD, Ann, AnneF, Audrey, Katherine, Lou agus mo chairde as an grá, craic, cairdeas agus tacaíocht. Is méanar dom, cinnte.

Thanks to Science Foundation Ireland (SFI) for funding this work, as part of the iCLIPS project, under the Research Frontiers Programme 2006 (grant number 06/RFP/CMS023).

Thanks to Microsoft Research for supporting this work through the provision of SenseCams and the Digital Memories (MyLifeBits) software (grant number 2007-056).

Sincere thanks to my supervisor, Gareth Jones, for his understanding, guidance, advice and support during this project.

Thanks to the many members of the CDVP, CNGL and other research centres who provided advice, help or friendship over the course of this research, especially Daragh, Eamonn and Yuki.

Special thanks to my parents, Eoin, Cle, Lor, Nan, AnneD, Ann, AnneF, Audrey, Katherine, Lou and my friends for their continued support, love and friendship in everything I do.

Part I

Introduction

CHAPTER

ONE

Introduction

Vannevar Bush could never have envisaged the impact his 1945 article 'As We May Think' [Bush, 1945] would have on modern science. In this article he presented his vision of the *Memex* which would continue to inspire scholars decades after its conception. This article is largely credited with proposing ideas that would lead to the development of the World Wide Web. However, Bush presented far more than the idea of linking pages of information. He provided a vision for a world where all of a person's personal information could be stored and importantly retrieved at a later stage. With advances in modern technology Bush's ideas are now coming to be realized.

It is now possible to digitally record and archive for long-term preservation many details of our lives. Details recorded can include everything from items read, written and downloaded, to footage from life experiences including photographs taken, videos seen, music heard, details of places visited, details of people met, etc. These can be captured from various platforms using a range of devices such as computers, mobile phones, cameras, video recorders, audio recorders, GPS sensors, biometric monitors, etc. When collected together these items form *personal lifelog (PL)* archives[1] which provide a rich record of our experiences. The availability of cheap digital storage means that it is in theory possible to easily store the digital experiences of an individual over their entire lifetime [Bell and Gemmell, 2009]. While whole lifetime PLs are perhaps unlikely to appear in the near future, it is increasingly easy to collect PLs over a period of years. Such PLs contain a rich record of our life experiences which can potentially be used in many applications, from looking up partially remembered

[1] In this thesis the term 'lifelog' is also used in reference to a personal lifelog (PL).

details to forming personal narratives from our lives for self reflection or to share with friends. However, any such applications will only be possible if relevant items can be retrieved reliably from the PL archive. While some progress has been made in the field of personal data retrieval, for example Gemmell et al [Gemmell et al., 2002], Cutrell et al [Cutrell et al., 2006a] and Kim et al [Kim and Croft, 2009], this element of Bush's vision largely remains an open challenge. This thesis seeks to progress research in this domain.

PLs raise many new retrieval challenges. To illustrate some of these challenges consider a sample scenario where a person is looking for a particular email within her PL archive from amongst the many emails she received from a friend Jack. All she now remembers is the sun glaring in the window, as she chatted with her friend Mary, when she received Jack's email. Conventional information retrieval (IR) techniques would not be capable of retrieving the correct item based on these criteria which are unrelated to the item's content. New approaches to IR using context are required, e.g. a system that could retrieve based on the weather and people present when the email was received. This requires capture and association of potentially useful context data. Further, since the PL in our sample scenario could potentially contain many emails matching the subject's recalled context details, techniques to detect the relevant one from amongst these candidate emails are also required.

This thesis centers on the development of retrieval techniques for the PL domain. Means to integrate subjects' recalled content and context, associated with required PL items, into successful retrieval algorithms are investigated. Developed algorithms focus on the textual media within PLs, possibilities for extension to other types of PL media, e.g. audio and images, are discussed in the context of future work. To improve the performance of our algorithms, metrics to detect the importance of lifelog items and ways to integrate these metrics into our IR algorithms are explored. Our investigation into the detection of important lifelog items from amongst the possibly vast number of items within such collections also has direct utility in its own right, such as in the suggestion of interesting items when browsing a lifelog collection for example. Hence in exploring the issue of important item detection we also consider other types of media beyond textual media.

To evaluate our retrieval techniques long-term PLs were created as part of the iCLIPS project[2]. These PLs contain 20 months of personal data, annotated with rich sources of automatically derived context data, for 3 subjects. Content-and-context-based queries, along with target result sets, were also created for these PLs.

In this chapter we highlight the retrieval challenges which motivate this work (Section 1.1), provide our hypothesis for the work (Section 1.2) and a summary of the contributions and aims of the thesis (Section 1.3). The chapter concludes with an outline of the remainder of the thesis.

1.1 Motivation

IR techniques have been applied to various types of collections, from early work on structured library catalogues and text-based academic sources to more recent developments such as multimedia content (e.g. video archives) and the linked and meta tagged World Wide Web. The focus of retrieval in these types of collections is predominantly on development of techniques to facilitate the finding of information, for example finding scientific articles, related to the user's need for information. IR techniques are also now being applied to collections which are in some way personal to the individual, such as collections of items on personal computers, email archives, etc. Here the predominant focus is on finding previously seen information, e.g. re-finding conference papers written or emails received. These types of personal collections may be considered subsets of lifelogs. Both these subsets and more heterogeneous PLs present new challenges for IR. PLs are fundamentally different from traditional content archives due to the following factors:

- A PL is typically a combination of many types of media, audio, video, images, and many types of textual content;
- There is the potential for a large percentage of noisy data in these archives, e.g. data which is incomplete (such as an email in the archive, for which the sender of the email was not captured) or of no current and future interest to the user;

[2] The iCLIPS project focuses on the automated annotation of multimedia items within lifelog archives to facilitate more effective browsing and searching through understanding of people's memory for lifelog items and development of backend retrieval techniques and suggestive user interfaces. See http://www.cdvp.dcu.ie/iCLIPS/ (September 2011) for further details. iCLIPS project members: Gareth J. F. Jones, Yi Chen, Liadh Kelly and Daragh Byrne (collaborator). Dublin City University.

- Many items in the archive may be very similar, repeatedly covering the same topic, hence making it difficult to extract the specific item(s) a user requires;
- A user may not be aware that a particular piece of data was captured (or have forgotten about its existence), and is therefore available for retrieval, e.g. a user forgetting that web pages on a programming problem they are currently trying to solve were previously viewed and are stored in their lifelog;
- The user may not be able to describe clearly what they are looking for, e.g. a user not recalling terms in an item they wish to retrieve;
- Items may not have formal textual descriptions, meaning that they cannot be retrieved using standard text or meta-tag based retrieval methods;
- Items may not be joined by inter-item links, meaning link structure cannot be utilized in the retrieval process.

It is this unique combination of attributes of PLs in comparison to document collections on which IR has traditionally focused that motivates this research into the creation of retrieval techniques specifically for the personal archive domain. This domain is fundamentally distinct from traditional IR domains due to both the above combination of factors, and the fact that items in PLs are personal to the individual, in that they have been created or obtained by the individual or represent an experience of the individual (e.g., programme from a concert attended, news article relating to a sports match attended). Since this is the case, the user may have personal experiences and memories associated with the items in the archive. The combination of these factors lead to the requirement of retrieval techniques specific to PLs.

1.1.1 Memory and Context

Existing work in personal data retrieval acknowledges that memory plays a vital role in item retrieval, for example, [Cutrell et al., 2006b, Elsweiler et al., 2007]. That is, people's memories of past experience of items, such as the location at which a file was stored, play an important role in relocating items. However, we find that current IR systems do not fully exploit what people remember about items. Items for retrieval from a PL may consist of the item itself, e.g. a document or email, and available associated context information, e.g. the time and date when a document was accessed or geo-location at which an email was written, captured automatically using relevant sensor technologies, e.g. use of GPS tracking to infer geo-location. However, much of the richer non-item related forms of context used in existing systems has generally relied on manual annotation from the user. For example, the PHLAT system allows individuals to add tags to their computer items [Cutrell et al., 2006b]. Also existing backend algorithms, which allow retrieval of multiple types of personal items, appear to operate on a simple boolean premise, for example Dumais et al [Dumais et al., 2003] and Gemmell et al [Gemmell et al., 2006]. That is, the queried context data is either present or not present in the database items. People's memories are likely to be incomplete or partially inaccurate, and thus the retrieval challenge is how to best make use of the captured content and context in the PL with user queries for most effective retrieval.

We assert in this thesis that to fully harness the power of people's memories of PL items, items should be implicitly annotated with rich sources of context data and sophisticated PL retrieval algorithms developed which can exploit these annotations to achieve effective retrieval of items of interest from an individual's PL, based on partially remembered details of item content and/or context.

1.1.2 Implicit Indicators of Item Importance

Due to the difficulty in discerning important lifelog items from the possibly vast number of items contained in lifelogs any additional information which can assist in this process is potentially very important. For example, identification of such items has potential utility both in re-ranking IR result lists and in identifying items to suggest to an individual when they are browsing their collection. Previous work has shown that an individual's biometric response is related to their overall arousal levels [Lang, 1995]. Significant or important events tend to raise an individual's arousal level, causing a measurable biometric response [McGaugh, 2003]. Events that can be recalled clearly in the future are often those which were important or emotional in our lives [Gazzaniga et al., 2002]. Current technologies enable the capture of a number of biometric measures on a continuous basis. For example, using a device such as the BodyMedia SenseWear Pro II armband[3]. However, past research has not exploited user biometric response associated with previous experience with an item in IR algo-

[3] http://www.bodymedia.com/ (September 2011)

rithms. As part of this research we explore the use of this biometric information in the re-ranking of IR result sets to make items which are more significant to the individual more easily available to them.

The exploitation of biometric data combined with recalled content and context associated with items in IR algorithms for effective PL search form the main objectives of this thesis.

1.2 Hypothesis

We hypothesise that people's memory of the generation of, and interaction with, items stored in their PL can be used to improve retrieval performance in the PL domain. While re-finding can be based on memories of search terms contained within items (i.e. item content), over time individuals' memories of these terms fade, resulting in increased difficulty in locating the required information/item. It is often easier to remember non-textual elements associated with items (e.g. date of previous interaction with the item) than the content/terms of the items themselves [Aizawa et al., 2004, Kelly et al., 2008]. These non-textual memories of items (context associated with items) may include the people present when accessing a file, the location one was in when they received an important SMS message, etc. With advances in modern technology these and other types of context data can now easily be captured, and used to automatically annotate PL items with rich sources of context data.

We further hypothesise that certain PL items are more important to the individual than others, and that such items may be determined from an individual's biometric response at the time of experiencing the item. In particular, we hypothesize that adding query independent boosts (static scores) to important items in lifelog IR result lists, where important items are detected based on recorded biometric levels associated with past accesses to the items, may improve retrieval performance.

The overall goal of the work reported in this thesis then is to develop and test methods which integrate automatically recorded and derived context data types with implicit indicators of item importance into IR algorithms for improved retrieval in the PL domain. In particular, it is hypothesised that integrating remembered and partially remembered context information into IR algorithms combined with implicit indicators of item importance will prove beneficial for PL retrieval.

1.3 Thesis Aims and Contributions

EXAMPLE 1 - 'single item' query:

Recalled Content (keywords):
Content term: personal
Content term: lifelog

Recalled Context type 'weather':
Context term: sunny

Recalled Context type 'people present':
Context term: Sarah

EXAMPLE 2 - query for multiple relevant items:

Recalled Content (keywords):
Content term: hashmap

Recalled Context type 'extension':
Context term: java

Recalled Context type 'month':
Context term: November

Recalled Context type 'year':
Context term: 2009

Figure 1.1: Sample content+context-based queries.

The aim of this research is to develop IR algorithms for a PL which will effectively retrieve the required information in response to a user's information need expressed as a query combining content and context features. Figure 1.1 shows two sample content+context queries. In the first example an individual wishes to retrieve a conference paper they wrote based on recalled keywords (content) in the paper and recalled weather conditions and people present (context) associated with creation of the paper or a previous access to the paper. The second example shows a query to retrieve all code which performs a specific task based on recalled keywords (content), extension type (context) and recalled month and year (context) of previous access.

When searching, individuals may perform 'single item' searches (e.g. searching for the camera ready version of a conference paper written), or searches where there may be multiple relevant items (e.g. searching for publications by a specific author). The developed algorithms could be, for example, plugged into a PL application which allows content+context-based queries.

Within the PL domain this thesis focuses on: 1) understanding, at an observed level, the types of data people recall about required items and on the integration of recalled context data into retrieval algorithms; and 2) improving ranked retrieval performance by investigating implicit indicators of lifelog item importance and exploring means to integrate them into PL retrieval techniques. Specifically, we seek to establish the following:

1. How can remembered context data best be combined into PL retrieval algorithms?
2. Can observed biometric response be integrated into retrieval algorithms to improve performance?

Answering these questions results in the following contributions being made by this thesis:

- Review and critique of related work in personal information systems, as well as existing work on the use of context data in retrieval and the use of biometrics in retrieval.
- Design and implementation of methods to integrate personal data into lifelogs.
- Design, implementation and evaluation of PL IR algorithms which cater for content+context queries.
- Design, implementation and evaluation of techniques to extract important items from PL collections using biometric response associated with past experience of items.
- Design, implementation and evaluation of techniques to integrate biometric measures associated with past experience of items into PL retrieval algorithms.

1.4 Thesis Outline

This section describes the organisation of the thesis.

Chapter 2 - Towards Context Data for Personal Lifelog Retrieval: Existing work in personal information access, content-based search, peoples' memories of computer activity, the use of context data in retrieval and biometric response analysis inform our work. In this chapter we overview and analyse existing work in these areas, and explain how it motivates our research to integrate implicitly recorded and derived context data into retrieval techniques for the PL domain. The chapter also overviews existing test set and test case generation approaches which inform our adopted test set and test case approach described in Chapters 3 and 4.

Chapter 3 - Lifelog Test Sets: In order to analyse our proposed approaches to integrate recalled context and query independent context into retrieval algorithms for the PL domain, it was necessary to create lifelogs and test cases for experimentation purposes. The chapter describes the contents of the lifelogs created for this study and the techniques used to create them. 20 month lifelogs containing records of all accesses to computer items (e.g., web pages, word documents, emails) and SMS messages sent and received were created for 3 subjects. Lifelog data items were annotated with several rich sources of automatically derived context data types, e.g. geo-location at the time of item experience using collected GPS data. To explore the role of biometric response in lifelog retrieval one month of the lifelogs was further annotated with biometric response data and passively captured images depicting the subject's life using the Microsoft Research SenseCam[4].

Chapter 4 - Towards Information Retrieval in the Lifelog Domain: In this chapter we overview and analyse existing IR approaches which inform our work. We review both existing ranked retrieval approaches and approaches for integrating static query independent scores into ranked retrieval algorithms. We also describe how we indexed the textual data in the lifelogs described in Chapter 3 to create test sets for the retrieval experiments presented in this thesis. The process by which test cases and result sets were generated for these experiments is also described.

Chapter 5 - Queried Content-and-Context-Based Retrieval Algorithms for the Lifelogging Domain: This chapter presents our adopted approaches for integrating

[4] http://research.microsoft.com/sendev/projects/sensecam/ (September 2011)

queried content and context into IR algorithms for textual data in the PL domain. For search, the key research challenge is how to score the individual fields individually or in combination to generate the most effective overall score for retrieval. In this chapter we explore two main approaches to this. The first approach, which acts as our baseline, uses a simple data fusion approach, where the overall relevance score is the sum of the individual field scores. The second uses a more sophisticated technique to exploit field structure and importance to determine relevance scores for items. Experiments to explore the effectiveness of these algorithms and analysis of results obtained are also presented.

Chapter 6 - Extracting Important Items using Past Biometric Response: As a precursor to integrating query independent biometric levels into the ranked retrieval algorithms presented in Chapter 5, we explore the predictive power of biometric response associated with previous experience of lifelog items in detecting lifelog items individuals might want to view in the future. Identification of such potentially important items within lifelogs also has other utilities, for example in the suggestion of interesting lifelog items to an individual when they are browsing their lifelog. This chapter presents a discussion of these topics, along with the setup of experiments to explore our adopted use of biometric levels associated with lifelog items and analysis of results obtained.

Chapter 7 - Static Scores: Boosting Relevant Items in Result Lists using Past Biometric Response: Following on from Chapter 6, this chapter presents the approaches we investigated to potentially improve our ranked retrieval algorithms, presented in Chapter 5, by integrating biometric levels associated with previous experience of items into these algorithms. Experiments to establish the effectiveness of these approaches and analysis of experimental results obtained are also presented.

Chapter 8 - Conclusions and Future Work: This chapter presents a summary of the scope and findings of the thesis, highlights the contributions made and provides directions for future work.

Part II

Background, Setup and Analysis

CHAPTER

TWO

Towards Context Data for Personal Lifelog Retrieval

Chapter Overview: This thesis is concerned with the exploration of retrieval techniques for the PL domain. Such technologies will potentially allow individuals to search a PL capturing aspects of their life based on recalled content and context of items which they wish to find. The thesis further examines the potential for improving the effectiveness of accessing relevant items by boosting the scores of items using the individual's biometric response associated with past interactions with these items. Existing work in the spaces of personal information access, content-based search, people's memories of computer activity, the use of recalled context data and query independent context data in retrieval and biometric response analysis inform our work. In this chapter we overview and analyse this work, and explain how it motivates our research to explore integrating implicitly recorded and derived context data types into retrieval techniques for the PL domain. The chapter also reviews existing evaluation approaches which inform the evaluation strategy we adopt later in this thesis.

2.1 Introduction

We are currently experiencing a digital revolution which is enabling the capture, storage and transmission of previously unimagined amounts of information. One of the features of this revolution is that people are storing increasing volumes and types of personal information in digital format. A very wide range of the materials of individuals' lives can now potentially be recorded through sound recording, photograph capture, motion picture capture, etc, and through storing copies of web pages visited, personal emails, files created, music listened to, etc. Indeed it has been proposed that we are moving towards a society of total digitization and the resulting ability to relive or retrieve all details from our lives [Bell and Gemmell, 2009]. This notion of being able to relive or retrieve all details from our lives through digital media is referred to as 'total recall'. While there is still some way to go to reach the state of total recall envisaged by Bell et al [Bell and Gemmell, 2009], increasing evidence that this society of 'total recall' may be established in the not too distant future is emerging. Decreases in cost in parallel with massive increases in storage capacity coupled with existing technologies mean that it is now in theory possible to store many details of our lives in digital format for long term preservation. Coupled with this, recent advances in research in the personal information space, which we look at in Section 2.2, mean that Bell's vision may be closer than we think. Some contend however that creating a society of 'total recall' is undesirable. For example, Bannon argued that there is a human need to forget [Bannon, 2006], which digital life stores would prevent individuals from doing. While others oppose lifelogs because of their potential intrusive nature [Nack, 2005]. Security issues pertaining to lifelogged data have also been raised [Bell and Gemmell, 2007]. While we believe that these are valid concerns which need to be addressed, further discussion of these topics is beyond the scope of this thesis, which is concerned only with mechanisms for effective retrieval.

Existing research in the personal data retrieval space use personal collections containing varying types and volumes of data [Cutrell et al., 2006a, Gemmell et al., 2002, O'Hare et al., 2006]. Some collections are purely visual, e.g. containing photograph collections representing snap shots from a person's life. Others may be purely textual, containing records of the computer activity engaged in by an individual - collections of this nature may contain web pages viewed, emails sent and received, documents opened on computer, etc. Richer personal collections will contain a combination of these image and data types, and may contain other sources of personal data such as text messages, audio logs, videos from one's life, etc. Items in personal collections may additionally be annotated with rich sources of context data such as geographic location, "date-time" information, weather conditions, etc.

The personal information space is a rapidly growing area of research, with researchers exploring techniques to both manage and retrieve various types of personal information [Barreau et al., 2009, Elsweiler et al., 2010, Gemmell and Sundaram, 2004, Gemmell and Sundaram, 2005, Jones, 2006, Mase, 2006, Teevan and Jones, 2008]. Our research focuses on the backend retrieval challenge associated with this space. Existing work in personal data retrieval finds that memory plays a vital role in item retrieval, for example, [Cutrell et al., 2006b, Elsweiler et al., 2007]. In the next section we review the role of people's memories of personal items in the retrieval process and explain how this motivates our research into exploring the role of recalled context data in personal information retrieval. We believe that beyond recalled context data other implicit context found in lifelogs can act as implicit indicators of item importance and offer potential utility as a static score integrated into ranked retrieval approaches for PL retrieval. We explore this topic in Section 2.3. Finally, in Section 2.4 we overview evaluation approaches for the personal retrieval space.

2.2 Memory and Context

Large amounts of data can be captured in a PL with the result that searching through one to find individual items is difficult, particularly when the user will often only remember small amounts of information relating to them. Often remembering the context of an experience with items can be easier than remembering these crucial keywords [Aizawa et al., 2004].

People's memory, or lack thereof, of items in their PLs is acknowledged as being crucial for effective retrieval and the design of effective PL interfaces. Key benefits from a greater understanding of how people remember, or forget, items should then lead to a greater understanding of what context data and interfaces might prove beneficial in aiding people recover, rediscover, or even discover items in their personal data archives. Elsweiler et al. [Elsweiler et al., 2005, Elsweiler et al., 2007] carried out an extensive user study to help them understand how people recover from memory lapses. They put forward the hypothesis that memory lapses make it difficult for people to find items in their personal archives, using traditional IR systems which require people to remember enough information about the item they are looking for to generate a query. Based on their findings they developed a photograph browser which exploits people's remembering mechanisms. More specifically the interface created always displays a user's entire photograph collection, photographs matching a user's query are enlarged and all other photographs are shrunk in size. Additionally, when a user hovers over a photograph the photograph is enlarged and information related to the photograph is displayed. They also provide a number of filtering options which they classify as: visual filtering, semantic filtering by free-text, semantic filtering by groups, temporal filtering via date line, spatial filtering by screen location and smart filtering.

In other memory related research, Jaimes et al. [Jaimes et al., 2004] use memory cues for a meeting video retrieval system. From a user study, they found what types of items people easily remember and easily forget about meetings. They used this information in the design of an intuitive interface that uses information which will act as memory cues, for the retrieval of the desired meeting video.

Much other research exists in support of the use of people's recalled memory of past interactions with items as a means to relocate them. Research such as [Blanc-Brude and Scapi, 2007, Teevan et al., 2004] discovered many attributes individuals recall about files such as location of a file, actions performed on the file, daylight status, weather and local time, which they use as context data to retrieve them. In other work standard forms of context data such as time, date, number of accesses, etc, have proved beneficial in retrieval from various types of collections (e.g. [Elsweiler and Ruthven, 2007, O'Hare et al., 2006, Ringel et al., 2003]). There are many other examples of the use of context in simple ways in existing work. We look at this in more detail in the next section in combination with an investigation of the current state of personal data retrieval/access systems and research.

2.2.1 Personal Information Systems

As mentioned in Chapter 1, in his seminal work Bush [Bush, 1945] proposed a future where all of a person's personal information could be stored and importantly retrieved at a later stage. “A Memex is a device in which an individual stores all his books, records, and communications, and which is mechanized so that it may be consulted with exceeding speed and flexibility. It is an enlarged intimate supplement to his memory” [Bush, 1945]. With advances in modern technology, Bush’s ideas are now coming to be realized. Microsoft’s Gordon Bell has dedicated many years towards the realisation of this vision with the digital archiving of his life, as part of the MyLifeBits project [Bell, 2004, Gemmell et al., 2002]. He has captured everything from letters, books, CDs, to items viewed on computer, phone conversations, etc. Beyond the capture of data, the MyLifeBits project is also looking at the retrieval of personal data. To this end they have developed a database which organizes all a person’s personal data. Within this database items have context data associated with them. However at present this context data is limited to standard items such as file location or date information. We believe that there are other rich sources of context information associated with a file that could be exploited, for example, people present and weather conditions. The MyLifeBits system only allows retrieval from PLs using simple interfaces based on timelines, context filtering and standard text based searching (which performs ranked retrieval).

In other work at Microsoft, the “Stuff I’ve Seen” (SIS) system [Cutrell et al., 2006a, Dumais et al., 2003] created an index of items previously viewed by the user on their computer and metadata associated with these items, such as date, author, etc. SIS contains an interface for retrieval of this previously viewed personal information. This interface takes advantage of the fact that stored contextual data such as author, time, thumbnails or previews of the item can act as cues to help trigger a user’s memory. The user can perform keyword search, combined with narrowing the results to certain types of files and items containing certain types of metadata (memory cues) such as time. An extension of this system, Phlat [Cutrell et al., 2006b], allows the user to add custom tags to items, which can then be used in future retrieval. However, we believe this manual tagging places a burden on the user and may not capture all of the details that an individual may recall about a searched for item.

Another example system Haystack [Karger et al., 2005] is an application that allows users to organize all their personal information in whatever way makes most sense to them. Similar to the MyLifeBits, Stuff I’ve Seen and Phlat systems, it removes the barriers that normally exist between different types of items, such as emails and photographs for example. Another system, Lifestreams [Freeman and Gelernter, 1996], replaces the standard desktop with an interface that arranges items in time-order. Using this interface a person can filter and order items. The VisMe system [Gomes et al., 2010] also uses the notion of organising desktop items in a timeline, with a linked graph style visualization displaying key terms and context associated with items, which can be expanded to navigate through the personal collection.

Other research has looked at the automatic grouping of items into tasks and folders. In this context a task is the user defined activity that they are engaged in, e.g. writing of thesis. The system in [Stumpf et al., 2005] automatically determines what task a user is engaged in, offers the facility to correct misclassified tasks, and groups items into tasks for the user. This is achieved using a probabilistic framework which considers personal items viewed and the context in which they are used. It also uses analysis of the user's current task to predict the next folder from which the user may wish to access information.

As can be seen from these sample systems, the key idea common to all current efforts in the personal data retrieval/access domain, is to create systems that make it easier for users to (re-)access personal information. Commercial systems which help people search through their personal files, while arguably not as sophisticated as some of the systems already mentioned, are also showing signs of beginning to address the personal file retrieval problem. For example, Google Desktop[1] indexes all items a person views and all items saved on the person's computer. Users can then perform text based queries to retrieve items at a later stage. Microsoft's Windows Desktop Search (WDS)[2] also indexes files on a person's computer and allows them easily query the index from the Windows taskbar. Another system, i-sho[3], organizes all a person's (or group of people's) personal data in a horizontal time-line interface. More precisely the interface is like a digital diary, grouping items by their timestamp. A separate timeline (layer in the application) can be created for different categories. For example, users can then choose to perform a search or view all photographs taken in a given day, month or year.

However, a significant disadvantage of these current systems is that the retrieval approaches are limited, as discussed in the next section. In this thesis we will explore

[1] http://desktop.google.com/ (September 2011)

[2] http://www.microsoft.com/windows/desktopsearch/default.mspx (September 2011)

[3] http://www.i-sho.com/ (September 2011)

the development of new context-based retrieval algorithms for the PL domain (this is explored in greater detail in the next section). Additionally, much of the burden for annotating items with context features is placed on the user. Users of course are often busy people, or perhaps just lazy, and very often will not take the time to add annotations to items; but they would still like the benefits of a system enabling them to search using rich annotations derived from context data. Widespread take-up of context-based search clearly requires methods for automatic annotation of items with minimal user involvement. The research presented in this thesis seeks, among other things, to address this issue. Towards this end, Section 2.2.3 includes an exploration of existing technologies for automatically generating rich sources of context data.

2.2.2 Recalled Context in Retrieval

As we have seen, existing research using context in the personal information retrieval space has predominantly focused on the front-end retrieval challenge and how memory of context associated with required information can be harnessed in these interfaces. The existing backend algorithms supporting context-based retrieval in these systems, which allow retrieval of multiple types of personal items, generally operate using a simple Boolean technique, for example [Dumais et al., 2003]. That is, the queried context data is either present or not present in the database items, thus relying on the user having detailed and accurate recall of the required query elements to locate relevant items. This is unlikely to be the case, given that queries will often be generated for partially remembered events from a few or possibly many years previously. Further, even if the user has accurate detailed recall of these details, it is not clear how best these multi-field semi-structured documents should most effectively be scored if a more sophisticated ranked retrieval strategy were to be used.

In earlier work we demonstrated that content and context data can successfully be combined to improve ranked retrieval effectiveness for a sample PL [Kelly et al., 2008]. This pilot study investigating the utility of remembered context data in desktop retrieval using 6 weeks of one subject's desktop activity found support for the use of context data in retrieval [Fuller et al., 2008, Kelly et al., 2008]. The context types explored are shown in Table 2.1. In this study precise geo-location was manually recorded by the subject, e.g., kitchen, office, and weather conditions for the region the subject was in during data capture were obtained from a website containing hourly

General Information	
Event ID	Event content
Context Data	
Title	Month
Source e.g., Word, Firefox	Hour
Type e.g., document, Web	Minute
Location e.g., college, kitchen	Weekday e.g., Mon, Tues
Weather e.g., showers, cloudy	Surrounding Events Types
Season e.g. summer	Surrounding Events Sources
Day	Surrounding Events Content
Year	

Table 2.1: Complete set of content and context.

weather history for Dublin Airport, Ireland[4]. This experiment was a known-item search task where the user searched for single partially remembered items from within the data collection. 27 tasks were created and the subject's content and context memory of these tasks tested after the lifelog build up period and again after a 6 month interval. Significant advantage was found for combining remembered content with context as opposed to using remembered content only, for both the initially recalled and recalled 6 months later data. Interestingly, for the recalled data 6 months later this advantage is significantly greater than it is for the initially recalled data. These results suggest support for the use of context data in retrieval, and that over time as an individual's memory of content data associated with lifelog items fades, the use of context data in retrieval becomes more important. Despite the fact that this was just a pilot study using one person's personal data recorded over a period of 6 weeks, the results obtained using simple retrieval approaches are promising and indicate that over the longer term, recalled context data could be used to improve content only retrieval performance in the PL domain. These findings, using a 'simple' retrieval approach, suggest support for the use of recalled content and context in lifelog retrieval algorithms.

In the next section we examine lifelogging beyond the desktop and how rich sources of context data associated with lifelogged items might be automatically generated.

[4] http://www.freemeteo.com/ (2007)

2.2.3 Beyond the Desktop

Various forms of visual logs, video logs, audio logs, mobile phone activity logs, accelerometer data, activity sensors, GPS, brainwave data and Bluetooth devices in the vicinity of the user, have been compiled and investigated [Aizawa et al., 2004, Blum et al., 2006, Ellis and Lee, 2006, Hori and Aizawa, 2003, Mase et al., 2006, Tancharoen et al., 2005].

An emerging area of lifelogging explored in this thesis is the proactive capture of images using devices such as the Microsoft Research SenseCam[5], see Figure 2.1, which allows for the creation of a visual log of an individual's activities [Gurrin et al., 2008].

The SenseCam is a digital camera, with fish-eye lens, worn around a subject's neck, which passively captures images depicting the wearers visual perspective. Image capture is triggered based on changes in sensor data captured by the device. For example, high acceleration values, passive infrared (body heat detector) as someone walks in front of the wearer or changes in light level. If no sensor has triggered an image to be captured, the camera takes one anyway after a period of approximately 20 seconds. When worn continuously roughly 3,000 images are captured in an average day. The SenseCam stores images and a sensor file on board. The sensor file contains among other things "date-time" stamps for images. These images, due to both sheer volume and unstructured nature, are impractical for a user to search through. Doherty et al [Doherty et al., 2007] segment these images into events using SenseCam sensor data. Key images, similar to keyframes in video search, are then extracted from each event. These key images can be used as a summary of a day or as an entry point into more images of a particular event. They also explore extracting the most important SenseCam events in each day through facial analysis and image analysis of the novelty of events relative to other days' events [Doherty and Smeaton, 2008]. Other work has explored detecting the similarity of SenseCam events using Bluetooth and GPS data [Byrne et al., 2007]. These works represent promising progress in helping people wade through potentially vast quantities of images depicting their lives. However, given the potentially huge numbers of images in personal SenseCam collections, means to detect the most important or interesting SenseCam events from within these collections as a whole are required. We investigate ways to do this using implicitly

[5] http://research.microsoft.com/en-us/um/cambridge/projects/sensecam/ (September 2011)

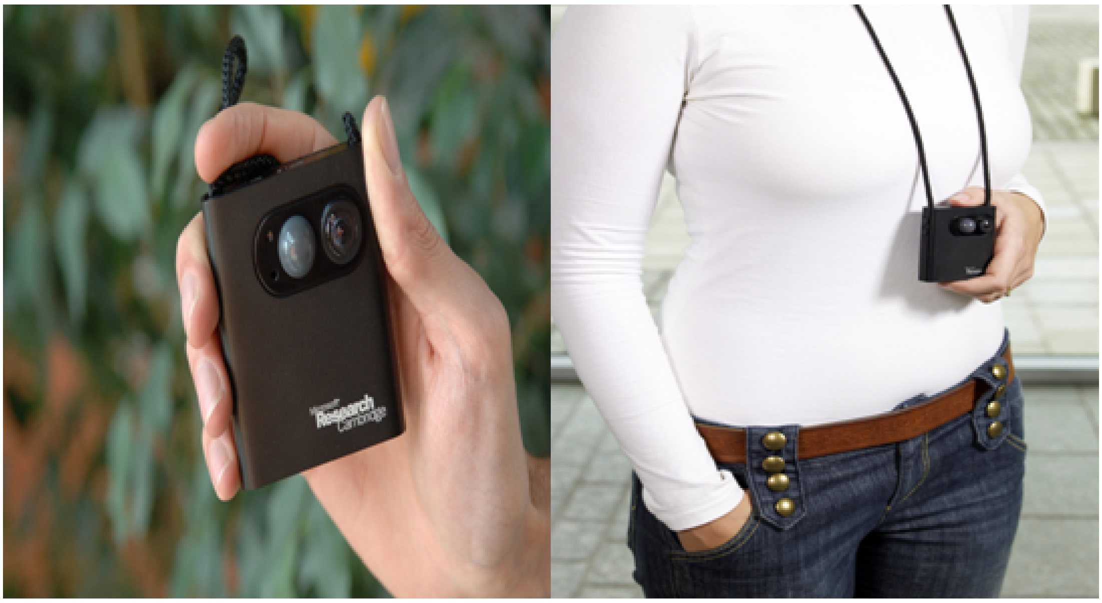


Figure 2.1: The Microsoft Research SenseCam. (Images courtesy of Microsoft Research website)

recorded context data in Chapter 6.

Ubiquitous computing also offers potential for automatically tagging lifelog items with rich sources of context data. Bluetooth tracking devices allow for the detection of other Bluetooth devices in the nearby vicinity - in today's society many people have Bluetooth technology activated on their mobile phones [Byrne et al., 2007, Lavelle et al., 2007]. This information may prove useful in subsequent search. For example, by tagging lifelog items with people present information derived from Bluetooth information, an individual may be able to recall who was present when they were working on or viewing a particular item. We investigate this in the next chapter as part of our lifelog creation process.

GPS tracking devices also offer the potential to tag lifelog items with information of potential use in ranked content+context retrieval approaches. A good example of this is provided in the MediAssist project [O'Hare et al., 2005, O'Hare et al., 2006] which provides an interface for search of digital photographs. Among other things they tag photographs with geo-location, light status and weather conditions, and allow search for images containing these tags. This is done by combining photographs timestamps with GPS information (from which geo-location is inferred), and downloading the light status and weather conditions for geo-locations from the web. This auto-context generation process could be used to tag all lifelog items with geo-location, weather conditions and light status information. We also investigate this in the next chapter as part of our lifelog creation process.

Personal information systems which stimulate remembering and allow the user to follow memory cues are required. These systems exploit humans' recognition and recall processes [Norman, 1998]. In order to create these systems, it is necessary to capture as much information as possible that people remember about items. Both suitable user interfaces, which enable the user to follow memory cues, and backend context-based search algorithms are required. In this thesis we focus on the exploration and development of user query driven context-based search algorithms, and automatic context generation to support these searches. Means to automatically tag items with rich context sources are explored in the next chapter. In Chapter 5 we investigate using context data in ranked retrieval.

2.3 Query Independent Context Data

Many systems obtain indicators of item importance from system users, either implicitly or explicitly in order to improve the performance of the system, or help locate items or information of importance/relevance to the user. Explicit relevance feedback is where the user themself informs the system of the importance of documents to their interests and needs. This explicit relevance feedback can be obtained from a number of sources [Brusilovsky, 1996], for example through user feedback where the user grades the relevance of items, a good example here is the 'thumbs up' and thumbs down' facility offered in many blogging systems for example, or through adaptation of the system at the backend (user model) or interface level, for example systems that allow users to add or remove items from their 'interested in list'. While explicit feedback can be a good way to discern the interests of users, this benefit is at the cost of the cognitive burden placed on the user [Belkin et al., 2000]. An alternative is implicit relevance feedback, or as we refer to them in this thesis implicit indicators of item importance, whereby the system attempts to infer how interesting the items are to an individual [Balabanovic, 1998, Ruthven, 2005]. Implicit indicators of interest can potentially be obtained from actions performed by individuals on items. Examples here include: actions on the web such as printing, saving, forwarding, bookmarking, replying to and posting a follow-up message to an item can indicate interest; returning to the previous document without having either saved the target document or followed further links can indicate disinterest; and since there is a tendency to browse links in a top-to-bottom, left-to-right manner a link that has been 'passed over' can be assumed to be less interesting [Lieberman, 1997]. Other examples here include: the number of times a person visits/views an item can indicate their level of interest in the item; and the time a user spends reading an item can also indicate their level of interest in the item [Nichols, 1997]. Much of this research on the use of implicit relevance feedback has been focused on web search, information filtering applications and recommender systems. Here implicit relevance feedback has been used either as a current indicator of item importance to guide a current search (e.g. the 'show me more like this notion'), or to form a user profile which provides details on the interests of users and in turn can be used to guide future search tasks (explicit relevance feedback is used in the same way). All of these implicit indicators of item importance, either future or current, are personal to the individual.

A different type of indicator of item importance, which is not personal to the individual, is the use of the web's link structure to detect the importance of web pages using algorithms such as PageRank [Page et al., 1998] and HITS [Kleinberg, 1999]. These implicit indicators of webpage importance are then used as static query independent boosts in retrieval algorithms. Use of algorithms such as PageRank and HITS has moved beyond the web into the personal file search space. [Chirita and Nejdl, 2006, Chirita et al., 2006, Kurland and Lee, 2006, Kurland, 2006, Soules and Ganger, 2005, Soules, 2006] used varying approaches to link items or result lists in desktop collections. These future personal, link based, indicators of item importance, are then used to re-rank the results of ranked retrieval result lists. More specifically, [Chirita and Nejdl, 2006, Chirita et al., 2006] factor link based indicators of item importance, derived based on access patterns between files and shared characteristics of files (e.g. linking files in the same folder) into the ranked retrieval score; [Kurland and Lee, 2006, Kurland, 2006] retrieve a set of documents in response to a user query using traditional methods and then use inter-document relationships to rank order the retrieved documents; and [Soules and Ganger, 2005, Soules, 2006] reranks the results of text-based queries using link based indicators of item importance derived from past access patterns between files.

Exploration of the many facets of implicit and explicit user feedback and user modelling is beyond the scope of this thesis. In this thesis we are interested in exploring a new type of implicit indicator of future item importance and its potential utility as a static, query independent, score integrated into ranked retrieval approaches for the lifelogging domain, namely biometric response associated with past experience of lifelog items. For the remainder of this section we overview biometric response and its existing uses in the digital environment as an implicit indicator of current item importance.

2.3.1 Biometric Response

As mentioned in Chapter 1.1.2, previous work has shown an individual's biometric response to be related to their overall arousal levels [Lang, 1995]. Significant or important events tend to raise an individual's arousal level, causing a measurable biometric response [McGaugh, 2003]. Events that can be recalled clearly in the future are often those which were important or emotional in our lives [Gazzaniga et al., 2002]. It has been demonstrated that the strength of the declarative or explicit memory for such emotionally charged events has a biological basis within the brain, specifically involving interaction between the amygdala and the hippocampal memory system [Ferry et al., 1999]. Variations in arousal level elicit physiological responses such as changes in heart rate (HR) or increased sweat production. Thus one way of observing an arousal response is by measuring the skin conductance response (SCR) (also referred to as galvanic skin response (GSR)). The GSR reflects a change in the electrical conductivity of the skin as a result of variation in the activity of the sweat glands. It can be measured even if this change is only subtle and transient, and the individual concerned is not obviously sweating [W. Boucsein, 1992, Gazzaniga et al., 2002]. The rate of heat exchange from a person's body to the outside environment, called heat flux (HF), also provides an indicator of an individual's arousal levels. Arousal response can also be observed through skin temperature (ST). With increased arousal levels, sympathetic nervous activity increases, resulting in a decrease of blood flow in peripheral vessels. This blood flow decrease causes a decrease in ST [Kataoka et al., 1998, Sakamoto et al., 2006]. Current technologies enable the capture of a number of biometric measures on a continuous basis. For example using a device such as the BodyMedia SenseWear Pro II armband[6] [Andre et al., 2006] which can continuously record

[6] http://www.bodymedia.com/ (September 2011)

the wearer's GSR, ST and HF, or using the Polar heart rate monitor[7] which can continuously record the wearer's HR.

A problem for arousal level detection using biometric response is that many factors, such as defective sensors and food intake, can cause noise in biometric data [Jain and Ross, 2004]. Noise in biometric data when attempting to use it to infer arousal levels is also caused by external factors such as physical activity, which also causes changes in biometric levels [Nakayama et al., 1977, Torii et al., 1992]. One way to measure levels of physical activity is through an energy expenditure calculation which considers a person's motion, age, weight and height. Energy expenditure is a calculation of the energy used by the human body, based on physical activity, resting metabolic rate and the thermic effect of food (cost of processing food for storage and use) [Ainsworth et al., 1993, Ainsworth et al., 2000, Black et al., 1996, Brockway, 1987, Denzer and Young, 2003]. Devices such as the BodyMedia SenseWear Pro 2 armband record, in addition to biometric readings, a person's acceleration and provide the option to enter one's weight, age and height. Using this data the BodyMedia device can calculate energy expenditure readings at a rate of once per minute, using proprietary algorithms which calculate energy expenditure based on the activity of the user, inferred from the on device data [Andre et al., 2006]. The validity of the BodyMedia SenseWear Pro 2 armband's energy expenditure calculation has been shown in various studies [Cole et al., 2004, Fruin and Rankin, 2004, Jakicic et al., 2004, King et al., 2005, Mealey et al., 2007, St-Onge et al., 2007]. While not, to our knowledge, explored to date, we believe that consideration of energy expenditure levels when attempting to infer arousal levels from biometric data may remove the noise in biometric data caused by some factors.

2.3.2 Biometric Response and the Digital Environment

Much research exists on exploring the relationship between biometric response and individuals' arousal and emotional levels, for example [Bradley et al., 2001a, Bradley et al., 2001b, Kim et al., 2004, Kim and Andre, 2008a, Kim and Andre, 2008b, Lang et al., 1993, Lang, 1995, Lisetti et al., 2003, Lisetti and Nasoz, 2004, Maltzman and Boyd, 1984]. Researchers have also begun looking at how an individual's biometric response may be used in emotion detec-

[7] http://www.polarusa.com/ (September 2011)

tion for HCI systems, for example [Anttonen, 2002, Anttonen and Surakka, 2005, Klein et al., 2002, Partala and Surakka, 2004, Picard, 2000, Picard et al., 2001, Scheirer et al., 2002, Ward et al., 2002, Ward and Marsden, 2003] and in eliciting of emotional response to movies and movie scenes, for example [Canini et al., 2010, Chen and Segall, 2009, Hettema et al., 2000, Mooney et al., 2006, Rothwell et al., 2006, Smeaton and Rothwell, 2009, Soleymani et al., 2008].

Research has also been carried out looking at the use of observed biometric response to detect tasks or items in different test sets which are of current relevance or importance to the individual. To our knowledge, at present there is only one example of work in this domain, that of selection and elicition of topical relevance for impersonal multimedia collections (TRECVid [Smeaton et al., 2007] and TREC Web track [Bailey et al., 2003] collections) [Arapakis et al., 2009]. In this work the authors show a relationship between the topical relevance of search results and an individual's emotional response, where emotional response is detected by passing biometric measures through a Support Vector Machine (SVM). This work represents exciting and promising progress in support of biometric response as an implicit indicator of current item relevance (or importance) for retrieval systems. However, to our knowledge previous research has not investigated the exploitation of observed biometric response as an implicit indicator of future item importance, nor has it looked at personal lifelog collections. This we believe is an important previously unexploited opportunity to gain passive feedback from subjects to potentially improve the retrieval performance of future searches in both lifelogging and other domains. In Chapter 6 we investigate our hypothesis that there is a relationship between biometric response at the time of experiencing items and the future importance of the items. Following this, in Chapter 7 we investigate the utility of these biometric response measures in re-ranking ranked retrieval result lists by adding the biometric measures as static implicit indicators of item importance to ranked retrieval algorithms. Our studies on the use of biometrics in retrieval are not comparable with those of Arapakis et al [Arapakis et al., 2009]. Their studies recorded biometric response in a controlled lab environment, whereas ours record biometric response 'in the wild'. Further, they used impersonal data, whereas we are dealing with personal collections; and they are examining the use of observed biometrics in the detection of current importance of items - that is they are attempting to detect, using biometric response, the items which are relevant to a given search, whereas we are examining the use of observed biometric response as a future indicator of item importance - that is we are attempting to determine whether biometric response associated with previous experience of items indicates the importance of items in collections as a whole.

2.4 Evaluation in the Lifelogging Domain

IR experiments often use test collections to evaluate techniques. These test collections typically contain a set of documents (or items), queries to perform on the set and a list of relevant target result items for each query. Collections of this nature follow the now standard Cranfield model [Cleverdon and Keen, 1966]. Various evaluation workshops, such as TREC[8], offer standardized IR evaluation tasks, by providing test collections, for various types of datasets, such as news archives and the linked and meta tagged World Wide Web. The focus of retrieval in these types of collections is predominantly on development of techniques to facilitate the finding of relevant information, for example finding news items on a topic of interest. A common characteristic of data sources that have been used for such standardized IR evaluation to date is that, the test set data can be shared with workshop participants (subject to copyright agreement). Further, the search requirements and information needs of target user groups of the collections can generally be captured and studied, and used to develop experimental search topics for evaluation purposes. The success of these systems is then assessed based on manual post hoc assessment of the data by analysing documents retrieved or submitted for assessment by workshop participants. Personal collections differ from the data sources used for existing standardized IR evaluation tasks in a number of ways. Firstly, these collections are personal to the individual, in that they have been created or obtained by the individual or represent experiences of the individual including for example, emails and SMS messages relating to concerts attended, news articles relating to sports matches attended. Since this is the case, individuals will generally be unwilling to share these collections. The collection owner may have personal experiences and memories associated with the items in their archive, which will inform, depending on their information needs at given moments in time, the items they wish to retrieve from the archive and the query terms

[8] http://trec.nist.gov/ (September 2011)

they will use in this retrieval process. This means that only the personal collection owner can provide their real re-finding information needs and the terms they would use for queries to retrieve relevant items. Further, only this individual can determine the relevance of items retrieved for a given information need from their personal collection. These key differences between personal collections and collections for which TREC-like IR tracks exist are important, make evaluation in this domain challenging[9] and to date have hindered formation of shared Cranfield style collections for personal search evaluation.

To conduct experiments in personal data search, researchers largely need to create their own test collections consisting of individuals' data, queries and result sets. There are a number of problems with this approach: 1) the effort required to create these collections; 2) the difficulty in gaining large volumes of subjects for such experiments; and 3) lack of comparability across research efforts. In an effort to overcome these problems Kim et al [Kim and Croft, 2009] have created pseudo desktop test collections for the desktop search space. The authors proposed amassing and creating 3 pseudo desktop collections by extracting emails of 3 individuals prominent in the TREC Enterprise Track collection [Craswell and Vries, 2005] and locating web pages, word documents, PDF files and PowerPoint presentations related to these people by a web search query consisting of the person's name, organization and area of speciality (provided by TREC expert search track). They randomly chose known items from these collections and used a modification of the approach proposed by [Azzopardi et al., 2007], for simulated query generation for webpage re-finding, to generate simulated queries across multi-field personal items. This approach presents a promising direction towards larger scale test collection creation to support research in desktop search, and provides a means to support research into the utility of desktop retrieval approaches without the need for real users and their collections. However, these collections do not represent the diversity of real users' collections, and hence may not provide a reliable way to evaluate the performance of retrieval algorithms intended for personal desktop collections. The created collections contain a limited number of item types and the same volume of each provided item type across the three collections (with the exception of emails the number of which showed large variation across the collec-

[9] Indeed the need to move towards standardization in this domain was highlighted at the recent SIGIR 2010 Desktop Search workshop [Elsweiler et al., 2010] and ECIR 2011 Evaluating Personal Search workshop [Elsweiler et al., 2011].

tions). Given the personal nature of desktop collections, we can expect individuals to have different types of collections, with varying volumes and types of content, covering varying volumes of topics. Further it is not known to what extent the query formulation approach used reflects what collection owners will actually recall about required items and hence the query terms they will use. Indeed the query generation approach of Azzopardi et al [Azzopardi et al., 2007] which forms the core part of this multi-field query formation approach was developed for webpage refinding, and is acknowledged by its authors to require further analysis and refinement to exhibit more of the characteristics observed by individuals in the webpage re-finding space in which they were working.

Other researchers in conducting retrieval experiments for personal search have used 'real' users and some form of their collection. In [Ringel et al., 2003], 30 queries with target email items which had been sent to a large number of people in a company were manually created. These queries were then entered by test subjects into the Stuff I've Seen (SIS) interface, resulting in the retrieval of various items from their personal collections including the target email. Subjects were then required to locate the target email, thus allowing for testing of various versions of the SIS interface. While this technique proved useful for testing features of personal search interfaces, it is not appropriate for evaluating PL retrieval algorithms where a user is required to use recalled content and context to form a query. Even taking a modification of this approach where we generate tasks by giving a subject a task description of emails that had been sent to a group would not be appropriate, as the subject may have no recollection of these emails, and therefore could not recall content or context data with which to form a query. Further, by providing individuals with the required target item, we are removing the real refinding requirements of individuals from the evaluation process. Finally, and most obviously, the test sets used in this approach lack the rich context sources we wish to explore in retrieval experiments.

Elsweiler's work [Elsweiler et al., 2007], similar to the SIS evaluations, used a static email collection for PL experimentation. That is, the interactions individuals have with PL items were not recorded. This information is required to assign rich item access related context types to items. For example, to tag an item with the months an individual accessed an item, each access to the item needs to be recorded. However Elsweiler's work did adopt a more personal approach to task generation. This work presented a framework for a task based approach to PL user evaluation. Specifically, over a period of approximately three weeks subjects recorded web and email related tasks (task = email viewed and the purpose for which it was viewed, e.g. relocate email which contains Joe Smith’s phone number). They also allowed for the generation of additional tasks by the experiment investigator through observation of the type of tasks recorded by subjects and by having a number of subjects in a given group (e.g. students in the same class) provide a tour of their collections. This allowed them form task descriptions which simulated a ‘real world’ task the subject may engage in. Tasks were then categorized into three distinct types: tasks requiring a specific piece of information from within a computer item (lookup task); tasks requiring a specific computer item (known-item tasks); and tasks requiring information from multiple computer items (multi-item tasks). This approach allows for comparison of retrieval performance across different task types (i.e., across ’lookup tasks’, ’known-item tasks’ and ’multi-item tasks’), and importantly the evaluation of a personal information retrieval application in a structured manner using the personal information owners themselves. Elsweiler et al. examined this task generation approach only on emails and web pages, and thus its portability to other item types is not guaranteed.

Moving beyond emails and web pages and into the space of personal computer items and of particular interest to us into the space of backend retrieval algorithm evaluation, [Soules and Ganger, 2005, Soules, 2006] logged the computer activity (including all accesses to items) of 6 subjects over a period of 6 months. The subjects submitted 3-5 content only queries, which they appeared to freely generate from their memory with respect to the collection period. These were multi-item standard ad hoc type queries, e.g. “locate all items associated with writing of thesis”. To create oracle results for these queries the results for each query across different search engines were pooled together, and the subject rated the relevance of the pooled result set. This evaluation approach is similar to standard TREC type pooling strategy and is of particular interest to us as it allows for a means to create Cranfield type test collections containing accesses to personal items, which once created can be used for exploration of and development of unlimited numbers of backend retrieval techniques without requirement for user interaction.

2.5 Conclusions

In this chapter we reviewed existing work in personal information retrieval and provided support for the retrieval techniques that we will explore in this thesis. These retrieval techniques will explore the integration of recalled content and context with query independent biometric context in retrieval techniques for the lifelogging domain. In order to explore such retrieval techniques test collections are required for evaluation purposes. As we saw in this chapter evaluation in the personal domain is challenging. Further PLs generally span a long time period, potentially up to a lifetime's worth of digital data could be contained in a PL. PLs will additionally typically contain many types of digital media, for example, computer activity, mobile phone activity. This means that to reliably evaluate retrieval approaches in the PL domain long term multimodal PLs are required. Further, in order to facilitate the automatic annotation of rich context sources to PL items, both sensor readings (such as GPS to allow auto detection of geo-location) and individuals' interactions (accesses) with lifelog items (e.g. to facilitate tagging items with all geo-locations experienced by the individual when accessing an item) need to be recorded. In the next chapter we describe how we created such collections for our evaluation purposes.

CHAPTER

THREE

Lifelog Test Sets

Chapter Overview: In order to investigate approaches to integrating recalled context and query independent biometric context into retrieval algorithms for the PL domain it was necessary to create test datasets for experimentation purposes. This chapter describes the design and contents of the PL test sets created for this study and the techniques used to create them. We first introduce our lifelog test sets which consist of lifelog items, including computer files interacted with, emails sent and received, webpages viewed, SMS messages sent and received, and SenseCam images captured, and the context data associated with these items (i.e., title; path to file; URL; extension type; to; and from) and context data associated with each access to these items (i.e., "date-time" related information; geo-location when accessing the item; light status when accessing the item; weather conditions when accessing the item; people present when accessing the item; and biometric response when accessing the item). We then describe the database structure used to archive these test sets. This is followed by technical details of the means used to log subjects activities to populate these lifelog data sets and a detailed analyses of their contents.

3.1 Introduction

PLs generally span a long time period, potentially up to a lifetime's worth of digital data could be contained in an individual's PL. They typically contain many types of digital media, for example computer activity, mobile phone activity, digital photographs, etc. This means that in order to begin to explore and seek to evaluate retrieval approaches in what begins to approximate a PL of realistic size, personal data collections gathered over a substantial period of time are required. With these points in mind, as part of the iCLIPS project[1], personal digital data was recorded over an extended period by 3 postgraduate students within our research group (1 male, 2 females; from Asian and Caucasian ethnic groups)[2]. Specifically PC and laptop activity, SMS messages sent and received, passively captured images depicting their life (using the SenseCam device, described in Chapter 2.2.3), digital photographs taken, and location and social context (i.e. geo-location and co-present Bluetooth devices from which people present can be inferred, described in Section 3.3.4.3) were recorded over a period of 20 months by the 3 postgraduate students. Biometric data was also recorded for a one month period (September 2008). It was not possible to capture biometric data for a longer time period due to the cumbersome nature of the biometric devices and psychological burden placed on subjects recording biometric data, described in Section 3.3.4.3. These recorded personal digital data types and the means used to record them are described in detail in Section 3.3.

These personal data collections are larger and richer than any others we know of and while collected for only 3 subjects their long term nature and richness provide us with unparalleled real lifelogs for experimentation in this emerging domain. While these lifelogs, to our knowledge, are unrivaled, it should be noted that this is not a commercial system, rather a research driven approach to creating lifelogs for experimentation purposes. The lifelogs consist of real data, gathered from real sensors and are sub-

[1] http://www.cdvp.dcu.ie/iCLIPS/ (September 2011).

[2] One of these test subjects was the thesis author. The other two subjects also conducted their own research experiments using the generated lifelogs. In the investigations presented in this thesis, the author was not an outlier in experimentation. Subjects, the author included, were provided instructions for generation of queries on their personal collections (as will be described in Chapter 4). The author, as a subject in the study, followed the same guidelines as the other two subjects. Subjects also rated the relevance of retrieved results for their user queries, using a simple Boolean relevant/irrelevant rating scale (also described in Chapter 4), which didn't afford opportunity for removal of objectivity on the author's, as subject, part. As will be described in Chapter 5, subjects were also required to rate SenseCam images and computer files, with varying associated biometric response, on a number of scales. Subjects, the author included, were not aware of the biometric response associated with images and files, hence objectivity was not removed by the use of the author in this study.

ject to failures of hardware, software, input and output at various points in the data processing chain [Byrne et al., 2010]. This is reflective of a real environment, although hopefully a commercial system would be more reliable. In our personal data gathering we are pushing the limits of the available data collection devices and software. Further, this is real data from individuals, as such it is subject to the individual's occasional need for privacy, need for mental breaks from the lifelogging process, and subjects' forgetfulness in turning several lifelogging devices on, charging devices and in downloading data from mobile devices before device memory fills (this forgetfulness results in inability to continue data recording until a download is made and in extreme cases corruption of data onboard the device) [Byrne et al., 2010]. These issues and the resulting implications on the make up of the personal data collections are discussed in Sections 3.3 and 3.4.

This thesis centers on the development of IR algorithms to enable individuals to search for items in their PL collections. We investigate the utility of using a subject's recalled content and context associated with required PL items for lifelog item retrieval and investigate development of retrieval algorithms designed to exploit these recalled features. Development of ranked retrieval techniques for the many types of media which may be present in a lifelog, e.g., audio, images, textual items, etc, is beyond the scope of one thesis. In this thesis we focus on the development of ranked retrieval algorithms for the textual media within PLs. Possibilities for extension to other types of media are discussed in the context of future work in Chapter 8.2.2. For the textual ranked retrieval experiments presented in this thesis the 20 months of PC and laptop activity and SMS messages sent and received were organised into lifelog test sets. These test sets contain the content of items (i.e., content of SMS messages and content of the computer files, webpages and emails interacted with) annotated with rich sources of context data derived using the location and social context and information contained within the items. Our choice of context data types was motivated by existing memory studies investigating individuals recall of context associated with items, discussed in Chapter 2.2. Specifically, each item was annotated with the following context types: words in item title (for computer files title = filename, for emails title = email subject field, for webpages title = title of webpage); extension type; path to file; URL (for webpages only); and to/from (for SMS messages and emails only). Each access to these items was also annotated with the following additional sources of context data: year;

season; month; day of week; weekday or weekend; beginning of week, mid-week or end week; part of day (i.e., morning, afternoon, evening and night); begin date and time; end date and time; device (e.g., laptop, mobile phone); light status (i.e. daylight and dark); weather; geo-location; and people present. Table 3.1 provides the complete list of context types used. The derivation of context types and organising of recorded personal digital data into lifelog test sets is described in Sections 3.2 and 3.3.

General Information	
Item ID	Item content
Item Specific Context Data	
Title i.e., computer filename, email subject, webpage title	Extension Type e.g., Excel, Web
Path to File	URL
To (for emails & SMS messages only)	From (for emails & SMS messages only)
Context Data Assigned to Each Access to an Item	
Begin Date & Time	End Date & Time
Year	Season
Month	Day of Week
Weekday or Weekend	Part of Week i.e., begin week, midweek, end week
Part of Day e.g., morning, afternoon	device e.g., PC, mobile phone
Geo-Location	Light Status i.e., daylight, dark
Weather e.g., raining, cloudy	People Present e.g. Joe Smith

Table 3.1: Complete set of content and context.

This thesis also explores the possibility of improving retrieval effectiveness using biometric metrics captured (for one month) in the context of textual items in a PL and methods by which these might be integrated into our IR algorithms. As part of our investigation into improving ranked retrieval effectiveness for the textual items in PLs using biometric metrics, we explore the utility of captured biometric metrics in detecting important lifelog items. Using biometric significance measures to locate important items from amongst the possibly vast number of items within PL collections also has direct utility in its own right. For example, in the suggestion of interesting items when browsing a lifelog collection. Hence, given the possibly vast number of SenseCam images in a lifelog and difficulty in locating interesting images from within such a collection, described in Chapter 2.2.3, in our investigation of important lifelog item detection using biometrics we also consider, in addition to textual media, SenseCam images as an example of the utility of our approach for other lifelog media types. To facilitate experimentation into extracting important items from lifelogs using biometric response and into re-ranking ranked retrieval result lists using biometric data, one month of the generated 20 month lifelog test set was further annotated with biometric data (described in Sections 3.2 and 3.3). Specifically GSR, HF, ST, HR and energy expenditure (described in Chapter 2.3.1) levels associated with accesses to lifelog items. SenseCam images for this one month period were also added to the lifelog test set (described in Sections 3.2 and 3.3) to examine important SenseCam event extraction using biometrics.

Subjects' lifelog test sets were stored locally on their PCs in an SQL database. The structure of these lifelog databases is described in detail in Section 3.2. Following the description of subjects' lifelog databases, in Section 3.3 we describe how we logged, derived and wrote subjects computer activity (Section 3.3.1), SMS messages (Section 3.3.2), SenseCam images (Section 3.3.3) and context data (Section 3.3.4) to the lifelog databases. We then provide an analysis of each subjects resulting lifelog database in Section 3.4. Finally, we conclude the chapter with pointers to the use of the test sets in the remainder of the thesis.

3.2 Test Set Structure

As described in the introduction of this chapter, 20 month textual lifelogs annotated with several rich sources of context data test sets were created to evaluate our ranked retrieval approaches. To facilitate our investigations into the utility of biometric response in retrieval, one of these months (September 2008) was further annotated with biometric data. This means that all items which were created or/and accessed during this one month period were assigned the biometric levels associated with the creation or/and access(es) to the items observed during this period. SenseCam images were also included in the lifelog for this month (September 2008). That is while subjects wore the SenseCam for the 20 month lifelogging period, in the experiments presented in this thesis we only use the one month of their SenseCam collections for which we also have biometric data and hence only this month of SenseCam images are added to subjects' lifelog databases. In this section we describe the structure of these lifelogs. The means by which the database tables were populated through lifelog data capture and derivation is described in the subsequent section, Section 3.3.

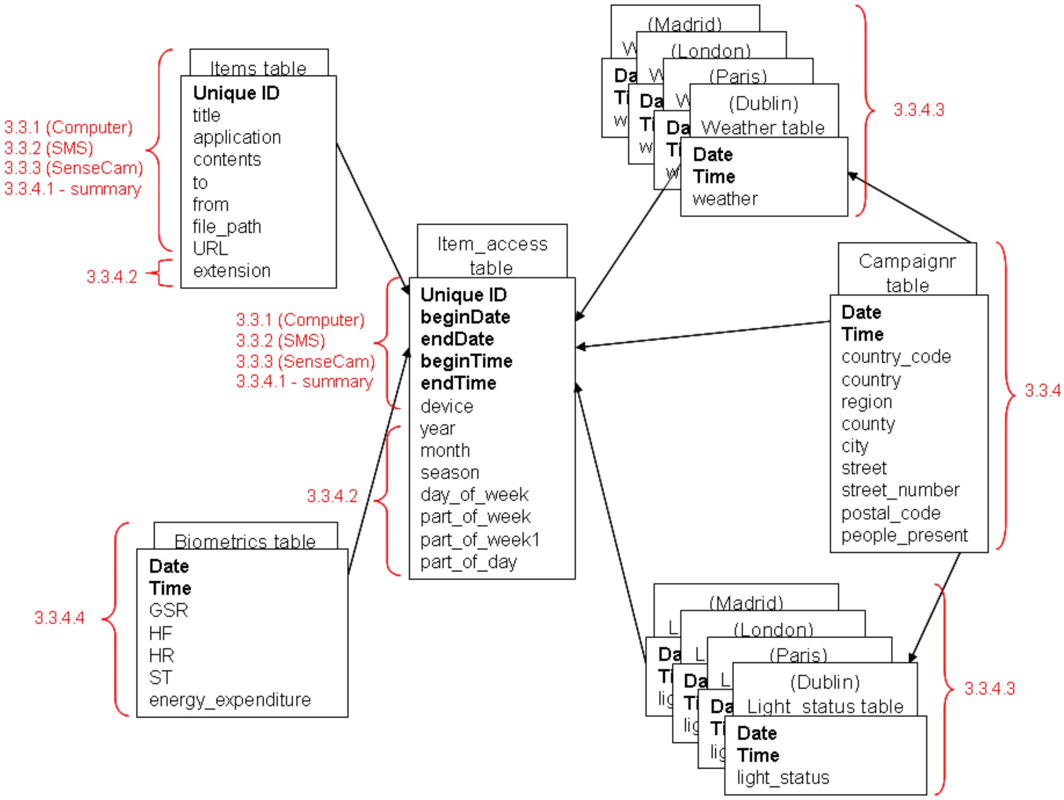


Figure 3.1: Lifelog database structure showing database tables, fields in each table, and links between tables. Fields in bold font are used to link tables. Section numbers, shown in red font, indicate the sections which describe how the lifelog data stored in each field was captured or derived.

Each of our 3 subjects lifelogs was stored locally on their PC in a SQL database. Figure 3.1 depicts the structure of the database and provides pointers to the sections describing the logging and derivation of the personal data stored in the database tables. The database contained an items table, item_access table, Campaignr table[3], weather tables (one for each geo-location visited by the subject during lifelogging period), light_status tables (one for each geo-location visited by subject during lifelogging period) and a biometrics table. The structure of these tables, along with sample fictitious data are shown in Tables 3.2 - 3.7. The remainder of this section describes these tables.

Items Table:

The items table maintains a list of all distinct items accessed by subjects, see Table 3.2. Specifically a unique_id is stored for each distinct item, along with title, application used to open the item (e.g., WINWORD.EXE, EXCEL.EXE, IExplorer), item content, to

[3] The Campaignr table stores geo-location and people present context data. This table derives it's name from the 'Campaignr' software which logs the data from which geo-location and people present context data were derived, described in Section 3.3.4.3.

Unique_ID	528
title	paper09
application	WINWORD.EXE
contents	In this paper we discuss....
to	
from	
file_path	C:\Documents and Settings\Jack\My Documents\paper09.doc
URL	
extension	word

Table 3.2: *Items* table fields and sample data. This table holds all unique items accessed by a subject.

Unique_ID	528	528
beginDate	2009-08-24	2009-09-09
endDate	2009-08-24	2009-09-09
beginTime	14:31:53	23:44:09
endTime	14:40:04	23:48:42
year	2009	2009
month	August	September
season	Autumn	Autumn
day_of_week	Monday	Wednesday
part_of_week	Weekday	Weekday
part_of_week1	beginWeek	midWeek
part_of_day	Afternoon	Night
device	laptop	laptop

Table 3.3: *Item_access* table fields and sample data. This table holds all accesses to items by a subject.

and from information (for emails and SMS messages), file path, URL (for webpages) and extension type (e.g., java (for java code file), excel (for XLS and XLSX files), word (for DOC and DOCX files), web (for webpage viewed), email (for emails sent and received), SMS (for text messages sent and received on mobile phone), SenseCam (for SenseCam images captured)).

Item_access Table:

The item_access table, see Table 3.3, maintains information related to the date and time for each individual access to computer items which can be linked to the items table based on unique item_ids. Specifically, the following date and time related information is held in the item_access table for items: "begin" date and time of the access to the item, "end" date and time of the access, along with year, month, season, day of week, whether the access happened at the weekend or during the week (part_of_week) and whether the access happened at the beginning-of-week, mid-week or at end-of-week (part_of_week1). The item_access table also records the device on

Date	2009-08-24	2009-10-01	2009-11-09
Time	14:32:00	16:01:25	19:51:22
country_code	IE	IE	IE
country	Republic of Ireland	Republic of Ireland	Republic of Ireland
region	Dublin City	Wicklow	Dublin City
county	Dublin	Wicklow	Dublin
city	Dublin	NULL	Dublin
street	NULL	NULL	Kinvara Rd
street_number	NULL	NULL	51-101
postal_code	Dublin9	NULL	Dublin7
people_present	My Nokia,Fiona		Sinead,Malcom

Table 3.4: *Campaignr* table fields and sample data. This table holds all geo-locations and people present experienced by a subject.

Date	2009-08-24	2009-08-24	2009-08-24
Time	13:00:00	14:00:00	15:00:00
weather	Light Rain	Clear	Scattered Clouds

Table 3.5: *Weather* table fields and sample data. A weather table was created for each geo-location visited by a subject. Weather tables hold the weather conditions experienced by a subject.

which the item access occurred. In our lifelogs the device is set to laptop or PC for computer items. We do not actually maintain accesses to SMS messages sent/received and SenseCam images captured in our lifelogs, rather just the "date-time" stamp for SMS message send/receive and SenseCam image capture is stored. The item_access table also holds these timestamps and sets the device field to 'mobile_phone' for SMS messages sent/received and to 'SenseCam' for SenseCam images captured. Sections 3.3.2 and 3.3.3 discuss this topic in greater detail.

Campaignr Table:

The Campaignr table records details on the geo-locations in which a person was and the people who were in their presence, along with date and time information, see Table 3.4. Geo-location and people present data is indicated for every 20 seconds (polled once every 20 seconds) in time over the 20 month lifelogging period (described in Section 3.3.4.3). Specifically the following fields are in the Campaignr table: date, time, country_code, country, region, county, city, street, street_number, postal_code, people_present. This table can be linked to the item_access table based on date and time information present in both tables.

Weather Tables:

A separate weather table was created for each country or region (e.g. regions in the

Date	2009-08-24	2009-08-24	2009-08-24
Time	14:32:00	14:33:00	14:34:00
light_status	daylight	daylight	daylight

Table 3.6: *Light_status* table fields and sample data. A light_status table was created for each geo-location visited a subject. Light_status tables hold the ambient light status experienced by a subject.

Date	2008-09-03	2008-09-03	2008-09-03
Time	14:13:04	14:13:05	14:13:06
GSR	0.213277742	0.220610112	0.219143629
HF	160.03681530000003	160.13633590000003	160.23585650000004
HR	62	62	62
ST	26.24159524	26.24159524	26.24159524
energy_expenditure	1.312743545	1.312743545	1.312743545

Table 3.7: *Biometrics* table fields and sample data. This table holds all raw biometric data captured for a subject.

USA = Arizona, Chicago) visited by the subject over the 20 month lifelogging period. Hourly weather information was available for each country or region (described in Section 3.3.4.3). Each weather table holds date, time and weather conditions information, see Table 3.5. The weather table for the appropriate region can be linked to the item_access table based on date and time information and geo-location information present in the Campaignr table.

Light_status Tables:

At the end of the lifelogging period, a separate ambient light_status table was created for each country or region visited by the subject over the 20 month lifelogging period, see Table 3.6 for the structure of a light_status table. Each light_status table contains date, time and light_status fields. Light status is represented as 'daylight' and 'dark' as two classes. Light status is present for every minute in a day (described in Section 3.3.4.3). Similar to weather data, light status data can be linked to the item_access table based on date and time and geo-location information held in the Campaignr table. Note, since we only require weather and light status information for each region for the periods in time when the subject was in these regions, we only stored light status and weather information in each region's light status and weather tables for the periods in which the subject was in these regions, as determined from the geo-location and date and time information in the Campaignr table.

Biometrics Table:

Biometric data is held in the biometrics table. The biometrics table contains date, time, GSR, HF, HR, ST and energy_expenditure fields (described in Section 3.3.4.4). This tables structure, along with sample data is shown in Table 3.7. The biometrics table can be linked to the item_access table based on date and time stamps.

3.3 Test Set Creation

In this section we describe how our three subjects' personal data was logged and used to populate the lifelog databases described in Section 3.2. Specifically, Section 3.3.1 describes how emails, webpages and textual files interacted with by our three subjects on their PCs and laptops were logged, using a combination of proprietary and bespoke software, and used to populate the title, application, contents, to, from, file_path and URL fields of their lifelog database items table and the beginDate, endDate, beginTime, endTime and device fields of their item_access table. Section 3.3.2 describes how SMS messages sent and received by subjects on their mobile phones were logged, using bespoke software, and used to populate the contents, to and from fields of their lifelog database items table and the beginDate, beginTime, endDate, endTime and device fields of their item_access table. Section 3.3.3 describes the capturing of SenseCam images and how details on these SenseCam images were written to the title and file_path fields of subjects' lifelog databases items table and beginDate, endDate, beginTime, endTime and device fields of their item_access table. Finally, in Section 3.3.4 we describe how context data was derived and used to populate the remaining fields of subjects' items and item_access tables and all fields of their Campaignr table, weather tables, light_status tables and biometrics table.

It should be noted that organising, time-aligning and joining lifelog data from diverse sources is challenging [Byrne et al., 2010]. Timestamps and data across devices are in different formats (text files, XML, CSV files, etc). Coupled with this, since our experiments rely on accurate data alignment, particularly in the case of biometric readings which can change on a second by second basis, it is imperative that timestamps across data sources are fully in synch. Data from different sources had different timestamps, e.g., laptop 5 minutes behind GMT, location data in UTC time, laptop and mobile data timestamps changing between GMT and PST as a subject travels for example, and some subjects mobile phones always recording time stamps in UTC time while others recording in the local time of their given location, etc. All data was time-aligned by

Field	Windows OS	Mac OS X
title	Slife & Digital Memories	Slife & Local script
application	Slife	Slife
contents	Java script	Java script
to	Java script	Local script
from	Java script	Local script
file_path	Digital Memories	Local script
URL	Slife	Slife
beginDate	Slife	Slife
beginTime	Slife	Slife
endDate	Slife	Slife
endTime	Slife	Slife

Table 3.8: Source of information (for computer items) for title, application, contents, to, from, file_path and URL fields in items table and beginDate, beginTime, endDate and endTime fields in item_access table.

converting it to the local time in the person's given location at any moment in time. This allowed for accurate context annotations to be made.

3.3.1 Computer Activity Monitoring

In this section we describe how the PC and laptop activity of our 3 subjects was logged[4]. Specifically, we describe how information for the title, application, contents, to, from, file_path and URL fields in the items table (see Table 3.2) and the beginDate, beginTime, endDate, endTime and device fields in the item_access table (see Table 3.3) was obtained for emails, webpages and computer files interacted with.

Our subjects were users of the Windows operating system (OS) and Mac OS X. Logging was carried out using a combination of the Slife package[5] (details of this package follow), Digital Memories software[6] (details of this package follow), locally written scripts[7] and created Java scripts to record subject's computer and laptop activity. Table 3.8 provides a break down of the fields in our items and item_access tables which were populated from information derived from each of these sources.

Slife Package

Slife was the core component in our computer activity monitoring. Slife is a productivity management tool, for example, showing people how much time they spend

[4] Two of the test subjects used Windows OS on their PCs and the other used Mac OS X. On their laptops one subject used Windows OS and the other two used Mac OS X.

[5] http://www.slifeweb.com/ (September 2011). We used the early 2008 version of the Slife application which was available under licence for Windows OS and Mac OS X without source code.

[6] http://research.microsoft.com/en-us/projects/mylifebits/ (September 2011)

[7] Locally written scripts were created by Daragh Byrne, DCU.

```
<?xml version=”1.0” encoding=”utf-8” ?>
- <EventHeader xmlns:xsd=”http://www.w3.org/2001/XMLSchema” xmlns:xsi=”http://www.w3.org/2001/XMLSchema-instance”>
<Activities />
<Date>2009-04-27T17:57:02.5000000+01:00</Date>
<EndDate>2009-04-27T18:02:02.4687500+01:00</EndDate>
<Name>Microsoft Office Word</Name>
<Title>instruction manual - Microsoft Word</Title>
<Type>application</Type>
<Url />
<Subtitle />
<To />
<From />
<Path />
<Application>WINWORD.EXE</Application>
<Content />
</EventHeader>
```

Figure 3.2: Sample Slife XML file.

using different computer applications. It monitors computer activity and records the event of a window being brought to the foreground (this we refer to as an *item access*). For each event it records: type of application (e.g. iexplore.exe), document source (e.g., Microsoft Word, Microsoft Outlook), window title (e.g. instruction manual - Microsoft Word), begin and end time of the event, URLs of webpages, and content of webpages viewed using the Internet Explorer application. Each event captured by Slife is stored in a separate XML file. Figure 3.2 shows a sample Slife XML file. While these Slife XML files provided details of accesses to computer items, they did not provide the 'raw' titles[8], path to files, to and from information for emails or the content of items (with exception of webpages opened through the Internet Explorer application).

Titles were extracted for computer items from the window title provided by Slife. For example, 'Adobe Reader - [thesis.pdf]' and 'thesis.pdf' become 'thesis.pdf', 'Macromedia Dreamweaver 8 - [C:\Program Files\Apache Software Foundation\Apache2.2\htdocs\diarystudy\databaseSettings.php (XHTML)]' and 'Macromedia Dreamweaver 8 - [C:\Program Files\Apache Software Foundation\Apache2.2\htdocs\diarystudy\databaseSettings.php (XHTML)*]' become 'databaseSettings.php'. Title tags in Slife files were then updated with these extracted

[8] Slife captures titles in the window title tag, however 'raw' titles are not always captured, e.g., capturing of window title 'instruction manual - Microsoft Word' as opposed to the title 'instruction manual'.

titles.

Path to files, to/from, and content information was obtained from the Digital Memories software[9], locally written scripts and created scripts and augmented to the Slife XML files (these we refer to as *augmented Slife files*). Sections 3.3.1.1 - 3.3.1.3 describe this process. Specifically, Section 3.3.1.1 describes how 'to', 'from' and 'content' information was obtained for Slife files representing accesses to emails, and how this information was augmented to Slife files; Section 3.3.1.2 describes how file paths were obtained for and augmented to Slife files which represented accesses to computer files using the Digital Memories software for Windows OS users and using locally written scripts for Mac OS X users; and Section 3.3.1.3 describes how the content was obtained for and augmented to Slife files representing accesses to files and webpages using file path and URL information.

After augmenting content, file path and to/from information to Slife files, the augmented Slife files were written to the items and item_access tables of our lifelog databases. The items table, as shown in Section 3.2, holds details on each specific item interacted with by the subject during the lifelogging period. The item_access table which links with the items table based on unique_ID contains information on each access to the specific items during the lifelogging period, also shown in Section 3.2. To transfer a subject's augmented Slife files, which contained all accesses to computer items during the lifelogging period, into the items and item_access tables of their lifelog database bespoke code grouped all accesses to specific items together. More specifically, augmented Slife files containing the same information for the title, URL, path, to and from tags were grouped together. Each of these groupings constituted an item (entry) in the items table and was assigned an unique ID, along with the common title, application, to, from, file_path and content[10] information of the grouping. The begin and end date and time information contained in each augmented Slife file in the grouping was written as an entry to the item_access table and assigned the unique ID of the grouping. The device field in the item_access table was assigned the value 'laptop' or 'PC' depending on which device the Slife file originated from.

[9] Digital Memories is the name used for the MyLifeBits software that was made available under research grant to a number of research centres. MyLifeBits (aka Digital Memories) runs on the Windows OS and is described in Chapter 2.2.1. Details of the use of this software in our lifelogging process follow later in this section.

[10] In the case of webpage or file content changing over the 20 month lifelogging period the most recent version of the content is used.

Problems Encountered:

Since Slife records all events of windows being brought to the foreground, much noise in the form of events which were not accesses to computer items, emails or webpages were present in the Slife collections. For example, opening folders in applications, opening applications, unsaved Word documents, system functions, saving files, etc, were captured as Slife events. This noise was removed from the Slife collections. To determine if an Slife file represented an access to a computer file, email or webpage generated scripts examined the application and title fields of Slife files (shown in Figure 3.2) and inferred from this information, using an extensive list of generated rules, whether the Slife file represented an access to a valid lifelog item. For example, *application = Microsoft Word* & *title = paper2009.doc* would represent an access to a computer file [using the rule: ***if****(application = 'Microsoft Word' AND title = *.doc)* ***then*** *the Slife file is not noise*], whereas *application = Microsoft Word* & *title = Save As* would not represent an accesses to a valid lifelog item [using the rule: ***if****(application = 'Microsoft Word' AND title = 'Save As')* ***then*** *the Slife file is noise*].

3.3.1.1 Determining Sender/Receiver and Contents of Emails

Since Slife does not capture 'to' and 'from' (i.e., receiver, sender) information for emails accessed, we obtained this information from a different source and augmented it to the Slife files. In this section we describe this process.

In order to augment Slife files with 'to', 'from' and 'content' details for emails, a copy of the emails sent and received by subjects was downloaded from the email applications used by subjects[11]. For subjects who used the Microsoft Outlook email application, emails were downloaded to a structured Microsoft Access database using the download emails facility available in Microsoft Outlook. All Slife files which represented an access to an email (i.e. Slife files where the *application* field = *Microsoft Outlook*) were augmented with the appropriate 'to', 'from' and 'content' information available in the Microsoft Access database of emails. More specifically, to augment these Slife files with 'to', 'from' and 'content' information we matched the "date-time" information and 'title' information in the Slife files with the subject (title) and "date-

[11] Recall that Slife captures each access (interaction) to emails, i.e. "date-time" stamps for each time an email was accessed. In contrast to this, the emails we downloaded from subjects email applications only contain a copy of the emails sent and received by subjects, which only provide the date and time that an email was sent or received, as opposed to "date-time" stamps for accesses (interactions) to the emails.

time” information in the Microsoft access database of emails[12]. The same process was used to augment Slife files with ‘to’, ‘from’ and ‘content’ information for subjects who used the Mac Mail application, except that in these cases we were looking for Slife files where the application field = Mail and emails were downloaded to a CSV file using locally written scripts.

Problems Encountered:
It should be noted that Slife did not always capture the title of emails viewed by subjects. Emails must be opened in a new window when using 'Microsoft Outlook' for the email title to be captured by Slife. If this is not done the title of viewing the window is simply recorded as 'Microsoft Outlook' by Slife. Subjects did not always open emails in a new window meaning Slife could not capture the title of the emails in these cases. While Slife still captures “date-time” stamps in these instances, with the ambiguous title of ‘Microsoft Outlook’ we have no way of knowing whether the subject was interacting with an email and if so which email. This means that Slife files with title field containing ‘Microsoft Outlook’ could not be augmented with ‘to’, ‘from’ and ‘content’ information.

This problem combined with Slife not running on subjects computers on some occasions, due to Slife crashes and subjects occasionally stopping the Slife application for privacy reasons, resulted in the fact that some emails imported to Microsoft Access and CSV files (as described previously) did not feature in subjects Slife collections. Hence to obtain full coverage of the emails interacted with by subjects during the lifelogging period we imported emails not captured by Slife directly from Microsoft Access and CSV files into subject’s lifelog database. More specifically, these emails were imported directly into the title, application, contents, to and from fields of the items table (see Table 3.2) and the beginDate, endDate, beginTime, endTime and device fields of the item_access table (see Table 3.3) of subjects lifelog databases. These emails are referred to as *IMPORTED_EMAILS*. Section 3.4 provides statistics on the volume of *IMPORTED_EMAILS* present in each subjects lifelog database. It should be noted that, since only the exact date and time of email send or receive is available for *IMPORTED_EMAILS* (as described previously), in subject’s lifelog database both the beginDate and endDate fields are populated with the date of email send/receive

[12] In matching Slife files with emails contained in the Microsoft Access database based on “date-time” information we took the email in the Microsoft Access database which had a “date-time” stamp less than or equal to the Slife files “date-time” stamp to be the matching email.

and both the beginTime and endTime fields are populated with the time of email send/receive. For received emails this timestamp does not necessarily always correspond to the time that a subject actually read an email. Further for both sent and received emails, it means we have no way of knowing how long the subject spent reading or writing emails nor how often they re-accessed emails.

3.3.1.2 Determining Path to Files

Since Slife does not capture file paths, to obtain file paths we also ran a second logging application on each subject's computer. For Windows OS users we used the Digital Memories software to capture the file paths of computer files interacted with and for Mac OS X users we used locally written scripts. This section describes the Digital Memories software and the locally written scripts used, and how path information was extracted from them and augmented to Slife XML files.

The Digital Memories software (aka MyLifeBits) is a personal information access system which runs on Windows OS, described in Chapter 2.2.1. Of interest to us in this chapter, Digital Memories maintains, in a Microsoft SQL server database, among other things the last access each day to files. Details stored include filename (title) and date time information for the *last* access each day to computer files. The Digital Memories database also holds the file paths for these computer files across several tables.

To augment Slife files representing accesses to computer files with file path information we matched the "date-time" information and title information in the Slife files with the date, filename (title) and path to file information extracted from Digital Memories. In matching Slife files with Digital Memories database entries based on "date-time" information we matched based on exact date match and Slife end timestamp less than or equal to Digital Memories timestamp.

The locally written Shell scripts generated for the Mac OS X, which were scheduled to run on a nightly basis, wrote details on files accessed each day to a structured text file. Details written to structured text file included filename (title), path to file and "date-time" information for the last access each day to the computer file. To augment Slife files with file path information we matched the "date-time" information and title information in the Slife files with the date, filename (title) and path to file information extracted from the structured text files. In matching Slife files with information contained in the structured text files based on “date-time” information we matched based on exact date match and Slife end timestamp less than or equal to the timestamp of an entry in a structured text file.

Problems Encountered:
Following addition of the file path information to Slife files, it was discovered that for many file accesses captured by Slife, path to file information was missing for Windows OS users. This was caused by the fact that subjects regularly stopped the Digital Memories software on the Windows XP operating system due to computer speed issues. This meant that interactions with files were not captured for Windows OS users in many instances, and hence path to file information could not be augmented to Slife files during these periods. To locate missing file paths generated Java functions searched subjects computers for files which matched the title (filename) captured by Slife and which had a date of last modification subsequent or equal to the “date-time” of access captured by Slife. This allowed for the addition of missing file paths to Slife files for computer files which had not been deleted by the subject in the intervening time or originally accessed from an external device.

3.3.1.3 Obtaining File and Webpage Content

Since Slife did not capture the content of items accessed (with the exception of webpages viewed using the Microsoft Internet Explorer application), we generated a series of Java scripts to obtain file content, webpage content and email content. Section 3.3.1.1 described the means by which email content was obtained and augmented to Slife files. In this section we describe how file and webpage content data was obtained and added to Slife files.

For Slife files representing accesses to computer files, the file paths in our augmented Slife files were used to extract the content of files located at the given file paths. Functions were written to extract the content of text files (which includes, code files (e.g. .java), .tex, .dat, etc), PDF’s and files in the Open XML standards (OOXML) format (which includes Microsoft Office 2007 and 2008 XLSX, DOCX and PPTX files) and OLE 2 Compound Document (OLE2) format (which includes Microsoft Office 1997-2007 XLS, DOC and PPT files)[13]. The obtained file content was then added to the

[13] Java libraries to extract content from OOXML and OLE2 format files are available at: http://poi.apache.org/ (September 2011)

content tag of the augmented Slife file.

For webpages, the URLs in the Slife files were used to extract the textual content of (non-dynamic) webpages at the given URL. HTML/XML tags were removed from the extracted webpage content and the resulting content added to the Slife file's *content* tag.

Problems Encountered:

Following addition of content information to Slife files, it was discovered that for many file accesses captured by Slife, the content information was missing. This was caused by files having been moved to a different location than that of the captured file path. To obtain the content for these Slife files, functions were created to search the computers of Windows OS users and local scripts searched the computers of Mac OS users for files based on their title/filename. Slife files were then updated with their new file paths. The content was also extracted from these new file paths and added to the Slife files. This allowed for the addition of missing file paths and content to Slife files, for computer files which had not been deleted in the intervening time or originally accessed from an external device.

One subject used the Internet Explorer web browser (version IExplorer 7) on their PC. Slife did not store webpage accesses correctly for the IExplorer 7 application, which uses tabbed browsing. For webpages viewed in the first tab of an IExplorer 7 window, Slife recorded the correct webpage information. However, for webpages viewed in subsequent tabs in the same IExplorer 7 window, incorrect webpage details were stored. Specifically, when the subject was viewing webpages in these subsequent tabs Slife incorrectly recorded that they were viewing the URL and content of the webpage in the first tab. Slife did however record the correct "date-time" stamps and title information for webpages viewed in these subsequent tabs. To overcome this problem we consulted the subject's Digital Memories database (see previous section, Section 3.3.1.2, for a description of Digital Memories). Digital Memories contains a table which holds a record of all webpages accessed by the subject using the IExplorer application. Information stored includes "date-time" of webpage access, webpage content, webpage title and URL. The Slife files were updated with the correct webpage content and URL by matching the Slife files with the webpage accesses stored in Digital Memories based on "date-time" stamp and title match. Since the facility to record webpage accesses using Digital Memories slowed down the speed at which webpages were retrieved from the web, logging of webpage accesses by Digital Memories was sometimes disabled by the subject on their PC. This meant that not all Slife files could be updated with the correct webpage content and URL using the approach described. Hence, the webpage content and URL information was deleted for Slife files for which the correct webpage content and URL was not available in the Digital Memories database. Statistics on the percentage of Slife files representing accesses to files which are missing webpage content are provided in Section 3.4.

3.3.2 Mobile Phone Activity Monitoring

In this section we describe how mobile phone SMS messages sent and received by our 3 subjects were logged and used to populate the contents, to and from fields of the items table (see Table 3.2) and the beginDate, endDate, beginTime, endTime and device fields of the item_access table (see Table 3.3).

In order to record the SMS messages sent and received by subjects on their mobile phones we required subjects to use a mobile phone with an operating system which would allow a script to be written to download SMS messages from the mobile phone. Further, as will be described in Section 3.3.4, the mobile phone used also needed to have GPS and Bluetooth capabilities and to use the Symbian 60 operating system for context data logging. The Nokia N95 mobile phone[14] which runs the Series 60 platform enabling mobile applications to be developed and deployed on it met these capabilities. Hence all subjects used the Nokia N95 mobile phone.

Logs of SMS messages sent and received were generated using a created Python script installed on the N95s. These logs contained SMS message content, name of SMS message sender/receiver, and "date-time" of SMS message send/receive stored in structured text files. A separate log file was created for SMS messages sent and received. These log files were generated by subjects running the created Python script on their mobile phone. Figure 3.3 provides an artificial sample log file for SMS messages received, showing the format of information stored in the log file. The log files were stored onboard subjects mobile phones and transfered to PC via USB cable, from where they were written to the subject's lifelog database. Specifically, the log data was used to populate the contents, to and from fields of the items table (see Table 3.2) and

[14] http://www.forum.nokia.com/devices/N95/ (September 2011)

```
.
.
Jill
Sun Sep 21 09:04:03 2008
“See you later!”
Fred
Fri Sep 19 21:09:57 2008
“OK”
Fred
Fri Sep 19 20:25:30 2008
“Thanks for this. Talk later!”
.
.
```

Figure 3.3: Extract of an artificial sample mobile phone log file for received SMS messages.

the beginDate, endDate, beginTime and endTime of the item_access table (see Table 3.3). The device field of the item_access table was assigned the value 'mobile_phone' for these entries.

Problems Encountered:

At the time of writing the Python scripts (early 2008), capabilities did not exist to log the date and time that subjects read SMS messages on their mobile phones. This meant that when we logged SMS messages the only “date-time” information available was the date and time that a subject received or sent the SMS message[15]. For received SMS messages this timestamp did not necessarily always correspond to the time that a subject actually read the SMS message. Further for both sent and received SMS messages, it meant that we have no way of knowing how long the subject spent reading or writing SMS messages nor how often they re-accessed them.

3.3.3 Microsoft SenseCam Images

A visual log of our 3 subjects' activities was created using a Microsoft Research SenseCam, described in Chapter 2.2.3. In this section we describe how SenseCam image information was written to our lifelog databases. Specifically we describe how the title and file_path fields of the items table (see Table 3.2) and the beginDate, endDate,

[15] Note: Since only the exact date and time of SMS message send or receive is available this means that in subjects' lifelog databases both the beginDate and endDate fields are populated with the date of SMS message send/receive and both the beginTime and endTime fields are populated with the time of SMS message send/receive.

beginTime, endTime and device fields of the item_access table (see Table 3.3) were populated with information for SenseCam images captured during the month of biometric data capture.

The SenseCam stores images and a sensor file (in CSV format) containing among other things title/filename (e.g. 02763.jpg) and date time stamp recording when the image was captured. SenseCam images and associated sensor file were regularly downloaded by subjects to a folder on their PCs hard drive. Generated Java code read the sensor files and used the "date-time" stamp of an image to populate the beginDate, endDate, beginTime, endTime fields of the item_access table[16] and used the title/filename to populate the title field of the items table. The file_path field was populated with the location of the SenseCam image on the subject's hard drive (e.g. C:\SenseCam_images\July2009\02763.jpg). The device field of the item_access table was assigned the value 'SenseCam'.

Problems Encountered:

Due to space limitations onboard the SenseCam it was necessary for our subjects to download SenseCam images from the device to a computer roughly every two weeks. After each download the SenseCam resets its internal clock to synch with the clock of the computer on which the download of SenseCam images occurred. Given that subjects downloaded their SenseCam images to different computers, which were not time aligned, at different times (e.g. downloading a set of SenseCam images to their laptop at the beginning of September 2009 and downloading the next set of SenseCam images to their PC in mid September 2009) this meant that SenseCam images were not all time aligned, and further that timestamps did not always reflect the exact moment in time when the images were captured. To overcome this issue all SenseCam images were converted to the exact local time of image capture. Given that our subjects lived in Ireland, for the most part this involved converting the majority of images to GMT. However, our subjects also travelled to different time zones, e.g. PST. Images captured in such different time zones were also converted to the local time zone of image capture, e.g. PST.

A further issue observed with SenseCam image timestamps, which was sporadically

[16] Note: an exact "date-time" stamp is provided for SenseCam images which represents the time that an image was captured. This means that in our lifelog databases both the beginDate and endDate fields are populated with the date stamp of the SenseCam image and both the beginTime and endTime fields are populated with the timestamp of the SenseCam image.

caused by file corruption onboard the SenseCam, was the conversion of images timestamps to a default date in the year 2000. These images timestamps were similarly converted to GMT, PST, etc as appropriate.

It should be noted that all details of subjects lives are not captured in SenseCam images. SenseCam images are missing due to device crashes, corrupt images, data loss on board the device, battery failure, subjects' need for privacy, subjects' need for psychological break from the recording process, unwillingness of some people encountered to be captured by the SenseCam, and subjects' feeling of being uncomfortable wearing the SenseCam in some social settings.

3.3.4 Context Data Generation

In this section we describe how context data was obtained for lifelog items and added to the lifelog databases (described in Section 3.2). Context data was firstly obtained from information directly available in our augmented Slife files, SMS message logs and SenseCam image collections. These context types are described in Section 3.3.4.1. Additional types of context data were also derived from information available from these sources. These context types, their derivation and how they were written to the items and item_access tables of subjects' lifelog databases are described in Section 3.3.4.2. GPS data and co-present Bluetooth devices were also logged by our subjects over the 20 month lifelogging period. In Section 3.3.4.3 we describe how this data was logged, the context data types derived from it and how these context types were added to the Campaignr, weather and light_status tables of subjects lifelog databases. Finally, in Section 3.3.4.4 we describe how biometric data was captured and added to the biometric table of the subjects' lifelog databases.

3.3.4.1 Context Data Available in Augmented Slife Files, SMS Message Logs and SenseCam Image Collections

Context data associated with distinct items in our lifelogs consisted of title, application, to, from, file_path, URL and extension type. These context types were stored in the items table of each subject's lifelog database (see Table 3.2, which is described in Section 3.2). Of these context types title, application, to, from, file_path and URL information was directly available for computer items in the augmented Slife files (as described in Section 3.3.1), for SMS messages in the SMS message logs (as described in Section 3.3.2) and for SenseCam collections from the information downloaded from the SenseCam device (as described in Section 3.3.3). Sections 3.3.1 - 3.3.3 describe how distinct items and these available context types were written to the items table of the subject's lifelog database.

The item_access table of the subject's lifelog database held "date-time" related information and the device type information for each access to lifelog items (see Table 3.3, which is described in Section 3.2). Of the "date-time" related information associated with an access to an item beginDate, beginTime, endDate and endTime were available in the augmented Slife files for computer items (as described in Section 3.3.1), in the SMS messages logs (as described in Section 3.3.2) and for SenseCam collections from the information downloaded from the SenseCam device itself (as described in Section 3.3.3). The context data type 'device' was also obtained from these sources. For computer items interacted with, this context type provided information as to the device from which the access to the lifelog data item occurred on (i.e. laptop or PC), described in Section 3.3.1. For SMS messages the device was always 'mobile_phone' (described in Section 3.3.2) and for SenseCam images the device was always 'SenseCam' (described in Section 3.3.3). Sections 3.3.1 - 3.3.3 describe how this device context data and "date-time" of item access context data was written to the item_access table of a subject's lifelog database.

In the next section we describe how extension type and "date-time" related context data was derived for lifelog items and added to the items and item_access tables of a subject's lifelog database. That is, we describe how the remaining fields of the items table (i.e. extension field) and of the item_access table (i.e., year, month, season, day_of_week, part_of_week, part_of_week1 and part_of_day fields) were populated.

3.3.4.2 Context Data Derived from Information Available in Augmented Slife Files, SMS Message Logs and SenseCam Image Collections

As discussed in the previous section, each access to lifelog items was annotated with several context types. This data is stored in the item_access table of a subject's lifelog database (see Table 3.3, which is described in Section 3.2). In this section we describe how "date-time" related information was derived from the beginDate, endDate, beginTime and endTime information available in the item_access table of a subject's lifelog database. We also describe how extension types were derived for the distinct lifelog items stored in the items table (see Table 3.2, which is described in Section 3.2) of a subject's lifelog database.

- **Date time related information:** Using the beginDate, beginTime, endDate and endTime data available for item accesses in the item_access table (see Table 3.3) functions were written to determine, the year, season, month, day of week, part-of-week (i.e., beginning of week, midweek or end of week), whether the item was accessed during the week or at the weekend (i.e. weekend or weekday), and period of the day (i.e., morning, afternoon, evening, night), in which the event took place. This "date-time" related information was then added to the item accesses in the item_access table. This information allows for the possibility to search based on "date-time" related recall beyond exact date and time. For example, to search based on recall of the season an item was accessed in.

- **SMS message and SenseCam image extension types:** SMS messages sent and received were assigned the extension type 'SMS'. Similarly images originating from the SenseCam were assigned the extension type 'SenseCam'. These extension types were written to the items table of a subject's lifelog database at the same time as writing of SMS message and SenseCam image details to the database, described in Sections 3.3.2 and 3.3.3.

- **Computer items extension types:** Using the title/filename and application data available for items in the items table functions were written to determine computer items extension types. Specifically, items opened using an internet browser, as determined using the application field data, were assigned the extension type 'web'; items opened using an email application, as determined using the application field data, were assigned the extension type 'email'; and computer files extension types (e.g., pdf, doc) were determined from both the data in the application field and the title field (e.g., application = 'Adobe Reader' AND title = 'paper1.pdf', means that the item receives the extension type 'pdf'; application = 'Eclipse' AND title = 'test.java', means that the item receives the extension type 'java'). Extension type data was added to computer items in the items table.

3.3.4.3 Context Data Derived from GPS and Co-present Bluetooth Device Logging

In this section we describe how the Campaignr table (see Table 3.4), weather tables (see Table 3.5) and light status tables (see Table 3.6) of a subject's lifelog database were populated.

Determining geo-location: Wireless network presence (Wifi) and Global System for Mobile Communications (GSM) location data was captured by constantly running the Campaignr software which runs on the Symbian 60 operating system, provided to us by UCLA (USA) [Joki et al., 2007], on the N95 mobile phones carried by the subjects. This data was polled once every 20 seconds. It should be noted that while it was also possible to capture GPS data using the Campaignr software, due to battery life issues GPS polling was deactivated. The Campaignr software stores polled data onboard the mobile phone and provides the facility for subjects to upload the data to a remote database. Geo-location was derived from the available Wifi and GSM data captured by Campaignr using in-house scripts[17]. These scripts provided country code, country, county, region, city and street timestamped geo-location data in structured XML files[18].

Determining people present: The Campaignr software [Joki et al., 2007] also recorded co-present Bluetooth devices to gain further information about a user's urban environment. Use of Bluetooth networking enables us to monitor other devices (with Bluetooth enabled) present in the nearby vicinity - in today's society many people have Bluetooth technology activated on their mobile phones. This Bluetooth information was included in the structured "date-time" stamped XML files generated from the Campaignr software, as described previously. Available Bluetooth information included the 'friendly' names people had assigned to their mobile phones, e.g. Deirdre mobile, Ireland10. This Bluetooth information provides a record of the mobile phones (with Bluetooth activated) and in turn the people who were in our subject's presence at given moments in time. We converted Bluetooth 'friendly' names to the names of the individuals who owned the mobile phones (where known), e.g., 'Deirdre mobile'

[17] Thanks to Daragh Byrne, DCU for creating these scripts and managing Campaignr database.

[18] These scripts derived geo-location information by first converting the GSM/Wifi data to GPS co-ordinates using the Google Gears API (http://code.google.com/apis/gears/api_geolocation.html (September 2011)). The GPS co-ordinates were then used to obtain the geo-location data using the GeoNames API (http://www.geonames.org/ (September 2011))

```
TimeGMT,TemperatureC,Dew PointC,Humidity,Sea Level PressurehPa,VisibilityKm,Wind Direction,Wind SpeedKm/h,Gust SpeedKm/h,PrecipitationCm,Events,Conditions,WindDirDegrees
12:00 AM,8.0,5.0,81,1026,10.0,SW,11.5,-,N/A,,Mostly Cloudy,230
12:30 AM,7.0,5.0,87,1026,10.0,SW,11.5,-,N/A,,Mostly Cloudy,230
1:00 AM,7.0,5.0,87,1026,10.0,SW,12.7,-,N/A,,Mostly Cloudy,230
1:30 AM,7.0,4.0,81,1026,10.0,WSW,12.7,-,N/A,,Mostly Cloudy,240
2:00 AM,7.0,4.0,81,1026,10.0,SW,11.5,-,N/A,,Mostly Cloudy,230
2:30 AM,7.0,5.0,87,1026,10.0,WSW,11.5,-,N/A,,Mostly Cloudy,240
3:00 AM,7.0,5.0,87,1026,10.0,SW,11.5,-,N/A,,Mostly Cloudy,230
3:30 AM,7.0,5.0,87,1026,10.0,WSW,12.7,-,N/A,,Mostly Cloudy,240
4:00 AM,7.0,5.0,87,1026,10.0,WSW,13.8,-,N/A,,Mostly Cloudy,240
```

Figure 3.4: Sample weather conditions file from wunderground.com.

became 'Deirdre', 'Ireland10' became 'Sam'.

These XML files generated from Campaignr data were written to the Campaignr table of subjects' lifelog databases.

Determining light status: As described in Section 3.2 a separate light_status table was created for each country or region visited by the subject over the 20 month lifelogging period. To obtain light status (e.g. dark) information for these tables we extracted light status information for each geo-location visited by subjects from timeanddate.com[19]. The light status information for each geo-location was written to light_status tables with date and time stamps.

Determining weather conditions: As described in Section 3.2 a separate weather table was created for each country or region visited by the subject over the 20 month lifelogging period. Hourly weather conditions (e.g, raining, snowing) were downloaded from Wunderground[20] in structured CSV format. Figure 3.4 provides a sample downloaded CSV file. "Date-time" stamps and weather conditions were extracted from these CSV files and written to the weather table for each geo-location.

Problems Encountered:

For Campaignr data logging due to device crashing at various points during the 20 months collection period some periods of the lifelogs do not have Campaignr data available. During the first 8 months of the logging, the Campaignr software was undergoing iterative change and improvement. While these changes were being

[19] http://www.timeanddate.com/worldclock/sunrise.html (September 2011)

[20] http://www.wunderground.com/ (September 2011)

made software crashes occurred intermittently, but went undetected until the subject checked if the software was still running. Following the first 8 months a much more stable platform was available resulting in less frequent crashing of devices. Further periods of the 20 month lifelogs are missing Campaignr data due to subjects' need to conserve battery life, subjects forgetting to turn on the device and subjects' occasional need for privacy.

Examining the geo-location information it was found that country and region provided the most accurate geo-location information. Much noise, missing data and inaccurate data was present for the other geo-location fields. Hence country and region were the only geo-location data used in our experiments. Region information refers to the town a subject was in, as opposed to the part of town, building or room for example.

It should be noted that the utility of Bluetooth data in assigning people present context tags to item accesses is limited. Bluetooth tracking is not always enabled on all people's mobile phones, indeed some users opt to never enable Bluetooth connectivity on their phone. This is further complicated by the 'friendly' names used by people on their mobile phones. For example, use of innocuous 'friendly' names such as 'xx99now', 'U2forever' or 'My phone'. While mobile phone 'friendly' names can be obtained from people regularly encountered and matched to the person's 'real' name, this is not practical for people encountered only once or twice.

It should also be noted that the accuracy of weather conditions context data is limited by the location of weather stations. Weather conditions available in the weather tables reflect the prevailing weather conditions at the nearest weather station to a subject's given location at a given moment in time. Since the weather station may be located many kilometres from the subject's current location, weather conditions at the nearest weather station may not always represent the weather conditions observed by the subject, e.g. raining in the subject's current location, but sunny at the weather station.

3.3.4.4 Context Data Derived from Biometric Response Recording

The one month SenseCam image collection and one month of subject's textual lifelog collection were also annotated with biometric data from the subject. It was not possible to capture this data for the entire 20 month lifelogging period due to the cum-

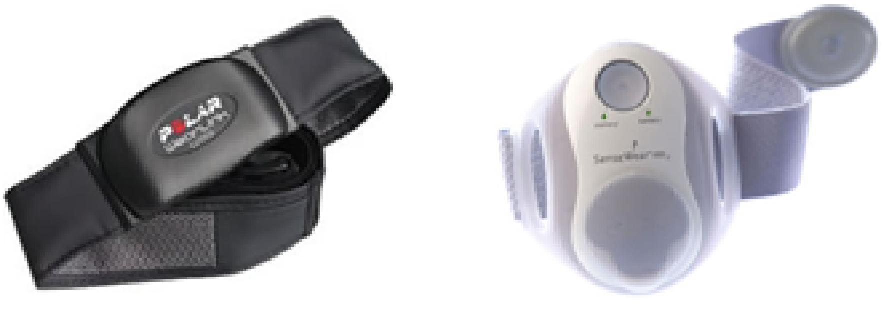


Figure 3.5: From left to right: the Polar heart rate monitor, and the BodyMedia SenseWear Pro2 armband.

bersome nature of biometric recording devices and psychological burden placed on subjects wearing these devices and recording biometric data. The following biometric measures were recorded: heart rate data (HR), galvanic skin response (GSR), heat flux (HF), skin temperature (ST) and energy expenditure. These biometric data measures are described in Chapter 2.3.1. In this section we describe how biometric data was captured and added to the biometrics table (see Table 3.7) of each subject's lifelog database.

Heart rate data was collected using a Polar heart rate monitor[21], as shown in Figure 3.5. Heart rate was sampled once every 5 seconds, this is the maximum sample rate afforded by the heart rate monitoring device. The heart rate monitor is worn around the chest, and heart rate readings are transmitted to a watch worn on the subject's wrist. Data is transferred from the watch to a PC using an infra-red sensor. Software provided with the device generates reports, graphs and text files of the heart rate readings for data analysis.

All other biometric data was collected using a BodyMedia SenseWear Pro2 armband[22] [Andre et al., 2006], as shown in Figure 3.5. The BodyMedia armband is worn on the upper arm and measures a range of psychological data. Data captured includes GSR along with transverse acceleration, longitudinal acceleration, heat flux and skin temperature. Energy expenditure is calculated by the device every minute using inbuilt software and stored onboard with the biometric readings. Data is transferred via a USB cable from the device to a PC. PC based software provided with the device gen-

[21] http://www.polarusa.com/ (September 2011)

[22] http://www.bodymedia.com/ (September 2011)

Captured Heart Rate Readings:
Timestamp: 15-09-2009:13:55:00 HR Reading: 77
Timestamp: 15-09-2009:13:55:05 HR Reading: 82

Heart Rate Readings with Missing Seconds Filled:
Timestamp: 15-09-2009:13:55:00 HR Reading: 77
Timestamp: 15-09-2009:13:55:01 HR Reading: 78
Timestamp: 15-09-2009:13:55:02 HR Reading: 79
Timestamp: 15-09-2009:13:55:03 HR Reading: 80
Timestamp: 15-09-2009:13:55:04 HR Reading: 81
Timestamp: 15-09-2009:13:55:05 HR Reading: 82

Figure 3.6: Approximating heart rate (HR) readings for seconds without HR readings.

erates graphs, reports and a CSV file of all the sensor output for data analysis. The BodyMedia device samples the values from its inbuilt sensors at settable predefined intervals. To allow for continuous recording over the course of a day without the device running out of memory, GSR was sampled once per second, heat flux once every 10 seconds and skin temperature once every 10 seconds[23]. Each time the device is placed on an individual's upper arm a calibration period is required in order for the device to produce accurate readings. Hence the biometric data captured during the first 45 minutes of each BodyMedia armband wearing period was removed from the collection. Readings were also found to be skewed during the last 2 minutes of each period of wearing the BodyMedia armband. This was the period when the device was being removed from the arm. Hence these readings were also removed from the collection.

Points in time for which biometric readings were absent due to the differing sample rates for HR, ST and HF were populated by examining the difference between two consecutive readings and filling in missing seconds readings with equal biometric intervals between the two captured readings. An example of this approach is provided in Figure 3.6. In this example two consecutively captured HR readings with timestamps of '13:55:00' and '13:55:05' are shown. In this example the difference between the two consecutive known HR readings of '77' and '82' is '5'. This means that to populate seconds between these two timestamps with equal biometric intervals we increase the HR readings by an interval of '1'. The resulting HR levels are shown in Figure 3.6. Seconds missing energy expenditure readings were similarly populated.

[23] Due to an error in recording ST was sampled once per minute for Subject 3.

The timestamped biometric data was stored in the biometrics table of the subject's lifelog database. Various approaches to annotating item accesses with biometric data based on "date-time" stamps were explored, these approaches are described in Chapter 7.

Problems Encountered:
The heart rate recording device was subject to frequent crashes, resulting in missing periods of heart rate data. In addition, clearly erroneous readings were often recorded by the device on random occasions following which it automatically recovered to correct capture. Analysis of the heart rate readings showed that when subjects' heart rate readings were higher than 140 erroneous readings were being recorded. Hence, to eliminate these erroneous readings, heart rate values greater than 140 and heart rate values of 0 (which are impossible) were deleted from the collection.

Periods of the biometric data recording period are missing biometric data for our subjects due to the psychological burden placed on subjects from continuous biometric data recording and their resulting need for breaks from the recording process. This need for breaks was due to the cumbersome nature of the biometric devices, physical irritation and discomfort experienced by subjects from prolonged contact of the devices with the skin, visibility (of the BodyMedia armband in particular) through one's clothing and knowledge that one's biometric response (while only available to the test subject) was being recorded.

Due to the storage constraints of the BodyMedia armband, which allowed for storage of one days worth of biometric data (with the biometric data sampling rate used in our studies), it was necessary for subjects to download biometric data from the armband on a nightly basis. Should the subject forget this nightly download, no further data could be captured until download was performed.

3.4 Test Set Contents Analysis

In this section a detailed breakdown of the lifelogs generated for experimentation purposes is provided. We break down this analysis into two parts. The first part, Section 3.4.1, examines the 20 months worth of textual data and associated context data contained in subjects' lifelog databases (these collections are referred to as the '20 month

	Subject 1		Subject 2		Subject 3	
Type	**Total**	**Content**	**Total**	**Content**	**Total**	**Content**
Code file	590	423	183	101	2,220	1,400
Excel file	455	266	66	40	141	69
Email	3,760	3,441	2,509	2,180	10,243	9,799
PDF	182	123	381	124	69	53
Presentation	92	82	147	124	95	23
SMS message	3,558	3,547	654	648	3,274	3,268
Webpage	3,895	1,248	15,642	12,697	44,457	39,545
Word file	311	162	310	209	373	277
Text file	381	273	81	46	308	238
Other	7	6	32	20	40	7
TOTAL	13,231	9,571	20,005	16,189	61,220	54,679

Table 3.9: Number of distinct items in each subject's collection. Code file = java, c, h, etc; Excel files = CSV, XLS and XLSX files; Presentation = Keynotes and PowerPoint files; Text file = txt, dat, tex, etc, files; and Other = class files, bib's, logs, etc.

	Subject 1		Subject 2		Subject 3	
Type	**Total**	**Content**	**Total**	**Content**	**Total**	**Content**
Code file	4,476	2,684	1,362	485	20,550	16,110
Excel file	831	566	240	175	1,158	959
Email	1,644	1,433	3,756	2,823	16,533	15,708
PDF	290	180	1,245	380	236	170
Presentation	363	259	384	246	1170	337
Webpage	3,022	829	27,209	21,823	96,833	91,544
Word file	400	263	1,558	869	4,392	2,897
Text file	542	491	175	95	1,491	1,365
Other	34	34	338	244	77	12
TOTAL	11,602	6,739	36,267	27,140	136,569	123,231

Table 3.10: Number of PC item accesses in each subject's collection. Code file = java, c, h, etc; Excel files = CSV, XLS and XLSX files; Presentation = Keynotes and PowerPoint files; Text file = txt, dat, tex, etc, files; and Other = class files, bib's, logs, etc.

textual lifelogs'). The second part, Section 3.4.2, examines the one month of subjects' lifelogs which contain SenseCam images and biometric data in addition to textual data and associated context data (referred to as the 'biometric month lifelogs').

3.4.1 20 Month Test Set

A detailed breakdown of the contents of the 3 subjects' 20 month lifelog test sets is presented in Tables 3.9 - 3.12.

As shown in Table 3.9, 13,231 distinct items were accessed by Subject 1 during the lifelogging period, of these items content data was obtained for 9,571 items. This equates to 72% of lifelog items having content data. Subject 2 accessed 20,005 distinct items during the lifelogging period, of which content data was obtained for 16,189 items.

	Subject 1		Subject 2		Subject 3	
Type	**Total**	**Content**	**Total**	**Content**	**Total**	**Content**
Code file	5,739	4,626	200	190	2,183	2,065
Excel file	2,823	1,497	23	23	153	84
Email	3,709	3,443	544	422	198	100
PDF	623	584	0	0	4	3
Presentation	536	531	50	45	659	7
Webpage	9,559	3,164	4,888	4,457	31,098	26,239
Word file	1,858	815	356	126	330	124
Text file	1,683	1,373	1	1	233	218
Other	128	127	0	0	1	0
TOTAL	26,658	16,160	6,062	5,264	34,859	28,840

Table 3.11: Number of laptop item accesses in each subject's collection. Code file = java, c, h, etc; Excel files = CSV, XLS and XLSX files; Presentation = Keynotes and PowerPoint files; Text file = txt, dat, tex, etc, files; and Other = class files, bib's, logs, etc.

Device	Subject 1	Subject 2	Subject 3
Mobile phone	2,184 (61%)	142 (22%)	2,660 (81%)
Laptop	18,893 (71%)	767 (13%)	26,091 (75%)
PC	9,836 (85%)	12,997 (36%)	134,215 (93%)
TOTAL	30,913 (74%)	13,906 (32%)	162,966 (93%)

Table 3.12: Total number of item accesses with geo-location tags in each subject's collection. The percentages in brackets provide the percentage of the total number of item accesses on each device that these figures correspond to. The number of item accesses on the mobile phone with geo-location tags corresponds to the number of SMS messages with geo-location tags.

This equates to 80% of lifelog items having content data. Subject 3 accessed 61,220 distinct items, of which 54,679 had content data. This equates to 89% of lifelog items having content data. Content data missing from the subjects collections are mostly attributed to the inability to locate the accessed files on the subjects' computers due to file deletion, file renaming or access to items located on some external device. The remaining files missing content also arise due to the inability of the Java libraries used to extract file content in all cases. Missing webpage content is attributed to dynamic webpages for which content could not retrospectively be extracted. The large volume of webpages missing content data for Subject 1 is attributed to the fact that this subject used tabbed IExplorer browsing thus preventing content extraction in all cases, as described in Section 3.3.1.3.

The substantial difference in the number of distinct items accessed by subjects is largely due to the significant difference in distinct webpages accessed by each subject. Subject 1 accessed 3,895 distinct webpages during the lifelogging period, while Subject 2 accessed 15,642 and Subject 3 accessed 44,457. Subject 3 also accessed a larger number of emails than the other two subjects. Subject 3 accessed 10,243 distinct emails (of these 5,870 were IMPORTED_EMAILS (described in Section 3.3.1.1), while Subject 1 accessed 3,760 distinct emails (of these 2,122 were IMPORTED_EMAILS) and Subject 2 2,509 distinct emails (of these 1,189 were IMPORTED_EMAILS). Other notable differences in subjects collections were the far greater number of distinct code files accessed by Subject 3, relative to Subjects 1 and 2; the greater use of excel by Subject 1 relative to the other subjects; Subject 2's greater access to distinct PDF files; and Subject 2's lesser access to distinct text files relative to the other two subjects.

Recall that Slife records the act of a window being brought to the foreground, this means that every time a user switches between two windows (or brings a window to the foreground) a new event (or item access) will be recorded by Slife. Tables 3.10 and 3.11 present a breakdown of item accesses by each subject on PC and laptop. As can be seen, the majority of Subjects 2 and 3's computer activity is carried out on their PC's, whereas Subject 1 makes greater use of their laptop.

The make up of subjects' collections will have implications on the content and content+context retrieval algorithms. Items missing content data will not be retrievable using content only retrieval; here the title of items will be the only content available to retrieve items. Further differences observed in subjects' collections will have implications on retrieval. For example, subjects recalling that an item was accessed on a laptop will narrow down the search space to a greater extent for Subjects 2 and 3 who made less use of their laptop relative to their PC; and the ability of recalled extension type to narrow down the search space will vary depending on the make up of individual subjects collections (e.g. given the relatively huge volume of webpages accessed by Subject 3 the web extension will not be particularly useful in narrowing down the search space for this subject, whereas use of the presentation extension type would be useful in narrowing down their search space).

Table 3.12 shows the number of item accesses on each device annotated with geo-location data for each subject and the percentage of their total accesses that these figures correspond to. The percentages of item accesses annotated with geo-location tags also correspond to the number of item accesses annotated with weather and light status information, for which geo-location information is required. Since the presence of geo-location tags means that the Campaignr software (as described in Section 3.3.4.3 the Campaignr software logs subjects GPS location from which geo-location was inferred) was running on subjects mobile phones at those moments in time, and since the Campaignr software also logs co-present Bluetooth devices from which people present can be inferred (described in Section 3.3.4.3), the percentage of geo-location tagged items also corresponds to the number of items which can have people present annotations. Of the items accessed by subjects, 74% were annotated with geo-location data for Subject 1, 32% for Subject 2 and 93% for Subject 3. As can be seen there are wide differences between the percentages of lifelog item accesses annotated with geo-location data across the three subjects. These differences are accounted for by subjects' different levels of need for privacy and break from geo-location logging, subjects' varying levels of forgetfulness in turning the Campaignr software on, varying speed at which subjects realised that the Campaignr software had crashed (as described in Section 3.3.4.3) and restarted the Campaignr software, variations in the volume of corrupt Campaignr databases across subjects (as described in Section 3.3.4.3), and subjects' varying level of need to conserve battery life. Missing weather, light status, geo-location and people present tags will have implications on retrieval in so far as recall of these context types only has the potential to aid retrieval of items which have these context tags, and indeed could potentially have negative impact on required items missing these tags, in so far as irrelevant items containing the tags will receive an extra boost in retrieval scores or be selected based on some filtering criteria. This potential problem could be particularly prevalent for Subject 2 given the particularly low number of items in their collection containing these context tags.

3.4.2 Biometric One Month Test Set

A complete breakdown of the one month (September 2008) of the subjects' lifelogs which contain biometric data and SenseCam images is provided in Tables 3.13 - 3.16. Table 3.13 shows the number of distinct items accessed during this period. Similar to the statistics observed for the entire 20 month collection in the previous section, over the biometric month the vast majority of the distinct items accessed by Subject 3 were webpages. This subject also accessed a larger number of distinct emails and code files than the other subjects. Also similar to the 20 month collections, Subject 2 accessed a much higher volume of distinct PDF files and presentations than the other subjects during the one month biometric period. However, while Subject 2 was also shown to access a large number of distinct webpages over the entire 20 month lifelogging

	Subject 1		Subject 2		Subject 3	
Type	**Total**	**Content**	**Total**	**Content**	**Total**	**Content**
Code file	117	92	3	1	264	117
Excel file	11	11	4	0	1	0
Email	201	181	152	125	669	634
PDF	27	21	71	16	16	12
Presentation	11	10	64	58	6	6
SMS message	139	139	24	24	135	135
Webpage	305	96	77	75	3,554	3,207
Word file	27	23	39	32	39	27
Text file	42	37	39	27	12	9
Other	2	2	11	9	10	2
TOTAL:	882	612	484	367	4,706	4,149

Table 3.13: Number of distinct items in each subject's 'biometric month' collection. Code file = java, c, h, etc; Excel files = CSV, XLS and XLSX files; Presentation = Keynotes and PowerPoint files; Text file = txt, dat, tex, etc, files; and Other = class files, bib's, logs, etc.

	Subject 1		Subject 2		Subject 3	
Type	**Total**	**Content**	**Total**	**Content**	**Total**	**Content**
Code file	654	622	9	4	3,658	1,369
Excel file	18	18	5	-	1	-
Email	141	119	207	110	1,393	1,329
PDF	48	37	176	42	35	28
Presentation	19	17	74	1	81	81
Webpage	630	179	125	129	10,349	9,724
Word file	50	48	120	19	664	191
Text file	49	43	46	31	27	12
Other	1	1	25	16	23	7
TOTAL:	1,610	1,084	787	352	16,231	12,741

Table 3.14: Number of PC item accesses in each subject's 'biometric month' collection. Code file = java, c, h, etc; Excel files = CSV, XLS and XLSX files; Presentation = Keynotes and PowerPoint files; Text file = txt, dat, tex, etc, files; and Other = class files, bib's, logs, etc.

period, this was not the case for the biometric month. Subjects' volume of access to different media types during a one month period will naturally not always reflect their general patterns of behaviour depending on their activities and information needs during a given one month period.

Tables 3.14 and 3.15 present the number of item accesses on PC and laptop recorded for each subject. No computer item accesses were recorded on the laptops of Subjects 2 and 3 during this month. This was caused by corrupt data, leading to impossibility in recovering the Slife data from these subjects laptops. Contrary to the entire 20 month lifelogging period, in the biometric month test set Subject 1 made greater use of their PC than laptop.

In Table 3.16 we see that there was no geo-location data available for Subject 3 during

	Subject 1		Subject 2		Subject 3	
Type	**Total**	**Content**	**Total**	**Content**	**Total**	**Content**
Code file	549	508	-		-	
Excel file	3	3	-		-	
Email	99	97	-		-	
PDF	27	27	-		-	
Presentation	-	-	-		-	
Webpage	329	178	-		-	
Word file	66	56	-		-	
Text file	57	56	-		-	
Other	9	9	-		-	
TOTAL:	1,139	934	-		-	

Table 3.15: Number of laptop item accesses in each subject's 'biometric month' collection. Code file = java, c, h, etc; Excel files = CSV, XLS and XLSX files; Presentation = Keynotes and PowerPoint files; Text file = txt, dat, tex, etc, files; and Other = class files, bib's, logs, etc.

Device	Subject 1	Subject 2	Subject 3
Mobile phone	82 (59%)	0 (0%)	0(0%)
Laptop	836 (73%)	-	-
PC	1,528 (95%)	396 (50%)	0 (0%)
TOTAL	2,446 (85%)	396 (50%)	0 (0%)

Table 3.16: Total number of item accesses with geo-location tags in each subject's 'biometric month' collection. The percentages in brackets provide the percentage of the total number of item accesses on each device that these figures correspond to. The number of item accesses on the mobile phone with geo-location tags corresponds to the number of SMS messages with geo-location tags.

the biometric period. This was caused by a corrupt Campaignr database, from which data could not be extracted. This means that recall of geo-location, people present, weather conditions or light status during the biometric month cannot offer utility in retrieval for this subject during this period. 85% of Subject 1's and 50% of Subject 2's item access during the biometric month had geo-location tags.

With regard to SenseCam images, during the biometric one month period Subject 1 generated 28,929 SenseCam images, Subject 2 had 17,377 images and Subject 3 had 47,301 images. Some periods of subjects' lives during the biometric month are not captured in SenseCam images due to issues observed with using the SenseCam, as described in Section 3.3.3. Further due to SenseCam data corruption (described in Section 3.3.3), one week of Subject 2's and Subject 3's SenseCam images are missing.

3.5 Conclusions

The lifelogs described in this chapter are used for the retrieval investigations described in the remainder of this thesis. Specifically, in Chapter 5 the 20 month lifelogs generated for each subject are used to investigate content+context-based retrieval algorithms for the personal lifelogging domain, and in Chapter 6 the biometric month textual lifelogs, SenseCam images generated during the biometric month and biometric data are used to explore extraction of important items from personal lifelogs based on biometric response associated with past experience of the items. Following on from Chapters 5 and 6, in Chapter 7 the biometric month textual lifelogs and biometric data are used to investigate integrating query independent biometric scores into the ranked retrieval algorithms developed in Chapter 5.

Before moving on to our retrieval investigations in Chapters 5 - 7, the next Chapter provides some background on information retrieval approaches of interest to our research and presents the approaches we used to index the subjects lifelog database content for retrieval experiments and to create search test cases for these investigations.

CHAPTER

FOUR

Towards Information Retrieval in the Lifelog Domain

Chapter Overview: This chapter serves as a precursor to Chapters 5 and 7, which describe our investigations into query driven retrieval algorithms for lifelog search, and integrating query independent biometric scores into these algorithms, for the textual data in the lifelogs. Following the chapter introduction, we review existing query driven retrieval approaches (Section 4.2). We then review approaches for integrating static query independent scores into query driven retrieval algorithms (Section 4.2.3). Finally, in Section 4.3, we describe how we created test set indexes from the lifelog databases of the subjects described in Chapter 3 and test cases (i.e. queries and target result sets) for these indexes based on subjects' perception of items they would like to retrieve from their lifelogs, and their recalled content and context associated with these items, which will be used for our retrieval algorithm development investigations in Chapters 5 and 7.

4.1 Introduction

Traditional content-based retrieval takes a number of words provided by a user as a search request and seeks to retrieve documents or more generally items which are relevant to the searcher's underlying information need by finding those which best match the query based on the IR content. However, other fields containing information about documents are often available to query on in many collections (e.g. XML document collections). Using these fields can offer potential for improving the retrieval process. Examples here include search of library records where searches can be performed on the multiple fields associated with library holdings, such as year of publication, author, title, etc, and, most similar to our work on PL search, search of desktop collections where document content can be associated with authors, date of creation or access, and file location (e.g. Microsoft's Windows Desktop Search (WDS)[1]). Such systems use *structured search* where query terms are entered in separate fields for the various context types associated with required documents. While multi-field structured search offers potential for improving the retrieval process, in so far as it can help narrow down a search space, it places extra cognitive burden on subjects by requiring them to enter query terms into several fields, and can also limit the scope for the use of the context of the query in so far as it requires explicit association of query terms with specific fields. *Flat querying* a queries elements represents a more flexible alternative to this problem by enabling a searcher to enter all content and context associated with required information in a single simple flat query. Search techniques for flat queries over multi-field documents are important to a number of IR search tasks. A key example of this is web data where the content of the webpage itself can be augmented with the title, the page location and additional content derived from linked anchor text for example [Craswell and Hawking, 2003]. Other important examples are search of XML document structures of varying complexity [Carmel et al., 2003, Lalmas, 2009], search of email collection with to, from, etc, fields [Craswell and Vries, 2005], and search of movie collections containing such information as genre of movie, actors, etc, [Kim et al., 2009]. Algorithmically, developing retrieval techniques to match simple flat queries to structured multi-field documents for successful retrieval is a significant challenge. We discuss this topic further in Section 4.2.2.2.

[1] http://www.microsoft.com/windows/desktopsearch/default.mspx (September 2011)

In this thesis we are interested in developing effective retrieval algorithms for the lifelogging domain which retrieve information relevant to the owner's information need, based on their memory of this information expressed as a query (we examine this in Chapter 5). Such queries might potentially consist of individuals' memories of the content of required information (e.g., content of email, content of webpage) or both the content and associated context (e.g., extension type of required item, prevailing weather conditions when required item was previously accessed) of required information. In the next section we overview existing query driven retrieval techniques. Specifically, Section 4.2.1 examines content only retrieval and Section 4.2.2 examines both structured and flat multi-field retrieval techniques.

Finding important relevant items from within collections in response to user queries poses significant challenges. That is, given a number of items which match a user's query, how do we infer which of these items are likely to be the item or items which an individual requires or which will be most useful to an individual. Any additional information which can assist query driven retrieval approaches in identifying important items is thus potentially very important. This query independent evidence of items importance can be referred to as *static scores*. There are many examples of the use of various types of static scores to boost user query driven scores in different domains. Examples include the well known PageRank algorithm which uses the structure of the web to compute static scores indicating the expected significance of web pages [Page et al., 1998], using webpage features such as document length and anchor text as static scores [Richardson et al., 2006], and using links created between computer files to infer static file importance scores [Soules, 2006]. Since we are interested in examining the potential utility of integrating biometric response, associated with past experience of items, into retrieval algorithms for the lifelogging domain (we examine this in Chapter 7), in Section 4.2.3 we review methods of integrating static scores into retrieval algorithms.

In order to perform either content or content+context query driven retrieval an index of the data to be queried is required. In Section 4.3.1 we describe the indexes of subjects' textual lifelog data which we created to allow us explore different retrieval techniques in Chapters 5 and 7. To evaluate the performance of retrieval techniques a set of user queries and target result sets are also required. In Section 4.3.2 we describe the means by which we generated user queries and target result sets from subjects' lifelog indexes to facilitate evaluation in our investigations of the development of retrieval techniques, and analyse the created test cases. Finally, in Section 4.4 we conclude the Chapter.

4.2 Query Driven Retrieval Approaches

In this section we review existing query driven retrieval approaches. Specifically, in Section 4.2.1 we review retrieval approaches which retrieve items based on the match between the users' query and content of the items available in the archive (i.e. content only based retrieval). In Section 4.2.2 we review multi-field based retrieval approaches. This is divided into two parts: 1) retrieval across multi-fields where the query is provided in a structured manner; and 2) retrieval across multi-fields where a flat query is provided. Finally, in Section 4.2.3 we review the re-ranking of retrieved result lists using static query independent scores.

4.2.1 Content Only Based Retrieval

Research in development of IR approaches which index the terms in documents to facilitate content-based retrieval has been ongoing since its first proposal in the 1950's [Luhn, 1957]. Luhn proposed that keywords could be either manually or automatically extracted from documents in a collection to create a representation of documents. Each document's representation (D) can be represented as a vector, as shown in Equation 4.1 where t_k is a term in the document representation[2]. Queries (Q) can similarly be represented as vectors, as shown in Equation 4.2 where q_k is a term in the query representation.

$$D = (t_1, t_2, ..., t_n) \tag{4.1}$$

$$Q = (q_1, q_2, ..., q_m) \tag{4.2}$$

[2] In modern retrieval systems generating automatic document representations for indexing purposes (to create vector representations of documents in a collection) can consist of automatic tokenization, stop word removal, stemming process, etc. The reader is referred to [Manning et al., 2009] for a good introduction to this topic.

In its simplest form retrieval of relevant textual documents which match a user's text query operates by searching the document collection for documents containing the query terms, and returning these documents to the user as candidate relevant documents. In other words, documents either contain the query terms (in which case they are retrieved for the user) or they don't contain the query terms (in which case they are not retrieved). This type of search uses Boolean algebra, and is hence referred to as Boolean search. Naturally, using Boolean algebra individuals can create more complex queries, using the Boolean operators, e.g., AND, OR or NOT (incidentally, in this case a query vector might be something like Q = ((q_1 AND q_2) OR q_3...). Boolean search is described more fully in [Manning et al., 2009, van Rijsbergen, 1979].

A problem with simple Boolean search is that only queries which exactly match the constraints of the Boolean query are returned as candidate relevant documents to the user. That is, a simple Boolean matching approach performs 'exact matches' where it checks for the presence or absence of query terms in combination with the constraints of the specified Boolean query in documents to determine documents which may be relevant to the query. Another problem with this approach is that no weighting of the likely relevance of documents or ranking of the output based on level of relevance occurs.

When determining if a document is relevant to a given query, it is useful to establish a degree of likely relevance of each document and to rank result lists for users based on these degrees of relevance. This is useful since the documents which are perceived to be most relevant to a query will appear at the top of a result list, making them more accessible to the individual performing the query. This led to the development of best (or partial) match retrieval approaches which produce result lists ordered according to the similarity of documents to the user query, where this similarity is some function of the number of search terms a query and document have in common. Using best match approaches documents and queries can be represented as vectors of weighted terms in a t-dimensional space, where *t* is the number of terms in the document collection representation (i.e. number of indexed terms). Equations 4.3 and 4.4 show the term vectors for a document (*D*) and query (*Q*) using this approach, where w_{ti} is the weight assigned to term t_i in the document representation and w_{qi} is the weight assigned to term t_i in the query. In these vector representations w_{ti} (or w_{qi}) is set to 0 when the term does not occur in D's (or Q's) representation.

$$D = (w_{t1}, w_{t2}, ..., w_{tt}) \tag{4.3}$$

$$Q = (w_{q1}, w_{q2}, ..., w_{qt}) \tag{4.4}$$

To calculate the similarity between a query vector and document vector, the vectors are compared using for example the dot product, shown in Equation 4.5 [Salton and Buckley, 1988]. In the simplest best match approach terms occurring in a document (or query) are assigned a weight of 1. This results in a query document similarity weight consisting of a count of the number of query terms present in a document in Equation 4.5. Documents are then ranked according to decreasing matching score. This is referred to as coordination-level matching.

$$Sim(Q, D) = \sum_{i=1}^{t} w_{ti} \cdot w_{qi} \tag{4.5}$$

Coordination-level matching assumes that all terms are equally discriminating in a document collection, and hence equally useful in determining the likely relevance of a document to a query, which is not the case. Hence, best match retrieval approaches which weight terms according to their discriminating power were developed. Term weighting allows for *selectivity*, where a good query term is one which has a high chance of selecting relevant documents from the many which will be non-relevant [S. E. Robertson and S. Jones, 1994]. The three commonly used characteristics of terms and document collections which are used to weight the occurrence of a query term in a document are: term frequency, inverse document frequency and document length.

Term frequency (tf) counts the number of occurrences of a term in a document. The rationale being that the more times a term occurs in a document the more representative of a document it is. The term frequency can be normalised, for example by dividing by the maximum term frequency (*maxtf*) as shown in Equation 4.6 [Salton and Buckley, 1988], where $tf_{d,t}$ is the term frequency of term *t* in a document *d*. Various variations of this weighting function are possible, such as that shown in Equation 4.7 for example.

$$tf_{d,t} = \frac{tf_{d,t}}{maxtf} \tag{4.6}$$

$$tf_{d,t} = 0.5 + 0.5 \cdot \frac{tf_{d,t}}{maxtf} \tag{4.7}$$

Inverse document frequency (idf): The concept underlying idf is that query terms which occur in few collection documents are more selective, more useful, than those which occur in many [Sparck-Jones, 1972]. An idf weight considers the number of documents a query term *t* occurs in (*df(t)*) relative to the number of documents in the collection (*N*), as shown in Equation 4.8. This idf weighting function always gives positive weights, with terms which occur in all documents receiving a weight of 0, which can be desirable given that the term offers no distinguishing power. Other approaches, for example as shown in Equation 4.9, do not reduce the idf score to zero for terms which occur in all documents. Some idf approaches, for example as shown in Equation 4.10, give negative weights to terms occurring in more than half the documents in the collection. Depending on the collection format this may or may not be desirable. The logs in idf scoring functions can be taken to any convenient base.

$$idf(t) = log\frac{N}{df(t)} \tag{4.8}$$

$$idf(t) = log(\frac{N}{df(t)+1}) + 1 \tag{4.9}$$

$$idf(t) = log\frac{N - df(t) + 0.5}{df(t) + 0.5} \tag{4.10}$$

tf×idf: Idf scores are commonly multiplied by term frequency in term scoring, this is referred to as *tf×idf* (term frequency times inverse document frequency) [Salton and Yang, 1973, Salton et al., 1975b]. tf×idf weights allow for term discrimination, the idea being that terms which occur frequently in a document, but infrequently in the collection as a whole, allow for the identification of individual documents from within a collection, and hence are the best terms for identifying the content of a document [Salton and Buckley, 1988].

Document length normalisation: Documents which are longer may have more occurrences of a query term simply because they are long relative to much shorter documents. This has the potential to result in a long document receiving a higher term score than a shorter document simply because it is longer, as opposed to as a result of it holding greater potential relevance to an individual's querying need. Several techniques to account for this have been proposed, such as normalising using a vector length normalisation factor (as shown next).

The vector space model [Salton et al., 1975a] uses the vector dot product function, shown earlier in Equation 4.5. It performs document length normalisation on the term weight by dividing the dot product by the moduli of the two vectors, as shown in Equation 4.11.

$$Sim(Q, D) = |Q||D|cos\theta = \frac{Q \cdot D}{\|Q\| \times \|D\|} = \frac{\sum_{i=1}^{t} w_{ti} \cdot w_{qi}}{\sqrt{\sum_{i=1}^{t}(w_{ti})^2}\sqrt{\sum_{i=1}^{t}(w_{qi})^2}} \tag{4.11}$$

4.2.1.1 Probabilistic Information Retrieval

An alternative to the Vector Space Model in IR is the probabilistic retrieval model. This seeks to measure the probability of a document (item) being relevant to a query given that the document possesses certain attributes (typically words or phrases) occurring in the user's request. Full details on the theory underlying probabilistic model in IR is contained in [Robertson and Sparck-Jones, 1976]. A well proven implementation of probabilistic IR is the Okapi BM25 model [Robertson et al., 1992, Robertson et al., 1993, S. E. Robertson and S. Jones, 1994]. Various term frequency and length normalisation approaches have been explored in the Okapi model. Equation 4.13 shows the term weighting approach used in Okapi BM25 [S. E. Robertson and S. Jones, 1994]. For a given document *d* and a given query term *t* the BM25 weighting function (shown in Equation 4.13) calculates a term weight ($w_{t,d}$). The overall probability of relevance (matching score *ms(q,d)*) for a document *d* is the sum of the weights of the query terms present in the document, shown in Equation 4.12.

$$ms(q, d) = \sum_{t \in q \cap d} w_{t,d} \tag{4.12}$$

$$w_{t,d} = idf(t) \cdot \frac{tf_{d,t} * (k_1 + 1)}{k_1 * ((1\text{-}b) + (b * \frac{l_d}{avl_d})) + tf_{d,t}} \tag{4.13}$$

where,

idf(t) = $\log \frac{N}{df(t)}$ in the implementation of BM25 presented in [S. E. Robertson and S. Jones, 1994], where N = number of documents in the collection, *df(t)* = number of documents term *t* occurs in. Of course other approaches can be used to calculate the idf, as discussed earlier in this section.

$\text{tf}_{d,t}$ is the number of occurrences of term *t* in document *d*.

l_d is the length of *d*.

avl_d is the average length of all documents in the collection.

k_1 = Tunable parameter which modifies the extent of the influence of term frequency. The higher values of k_1 increase the influence of *tf*; k_1 = 0 eliminates the influence altogether.

b = Tunable parameter which ranges between 0 and 1. Modifies the effect of document length normalisation. *b*=1, the assumption that documents are long simply because they are repetitive. *b*=0, assumption that documents are long because they are multi-topic.

In Chapter 5.3 we explore the use of BM25 for content only based retrieval of items in PL collections. Justification for use of BM25 in these investigations is provided in Chapter 5.2.

4.2.2 Multi-field Based Retrieval

When allowing users to query over multi-field documents two approaches can be taken. The first approach sees a user specify the query terms to be used for each field. A retrieval system then performs a document field by field search, using the query terms provided for each field. This is referred to as 'structured search', described in Section 4.2.2.1. In the second approach the user specifies all the query terms to be used in the search, without providing any indication as to the field each query term should be targeted to. A retrieval system then performs a search across all document fields

EXAMPLE 1 - structured query:

Content field query terms: *content context lifelogs*
Title field query terms: *PhD thesis*
Extension field query terms: *pdf*

EXAMPLE 2 - flat query:

Query terms: *content context lifelogs PhD thesis pdf*

Figure 4.1: Sample structured and flat queries.

using the provided query terms. We refer to this as 'flat search', described in Section 4.2.2.2. Figure 4.1 shows the difference between a structured query and flat query to a retrieval system.

4.2.2.1 Structured Content+Context Search

A simple approach to scoring in structured search is using simple data fusion [Belkin et al., 1995], whereby each field of a document is queried separately using the query terms for that field, and a simple linear combination of the individual field scores taken, as shown in Equation 4.14, where *f* is a field in document *d*, w_f is a weight assigned to each field *f*, and $ms(q_f, d_f)$ is a matching score approach which computes a matching score for field *f* of document *d*, given the query terms q_f which were used for this field. The $ms(q_f, d_f)$ score can be calculated using any query-document matching algorithm. For example, Equation 4.15 shows the use of BM25 (Equation 4.12 described in Section 4.2.1.1) in this process, where $w_{t_f,f}$ is the application of Equation 4.12 to the field *f* of document *d* using a query term t_f in q_f. Indeed in Equation 4.15, any matching score function can be used to calculate $w_{t_f,f}$.

$$ms(q,d) = \sum_{f \in d} w_f \cdot ms(q_f, d_f) \tag{4.14}$$

$$ms(q,d) = \sum_{f \in d} (w_f \cdot \sum_{t_f \in q_f \cap f} w_{t_f,f}) \tag{4.15}$$

In Chapter 5.3 we investigate the performance of this simple data fusion approach for structured content+context-based search on PL collections. This allows us to examine the effect of allowing the PL owner form structured queries based on their recalled content and context associated with required items, relative to content only querying. Since we use BM25 for our content only based retrieval investigations in Chapter 5, in exploring structured content+context-based retrieval the BM25 term weighting function is used in this simple data fusion approach shown in Equation 4.15.

4.2.2.2 Flat Content+Context Search

As highlighted in Section 4.1 structured search can limit the scope for the use of the context of the query in so far as it requires explicit association of query terms with specific fields. Flat search represents a more flexible alternative by enabling a searcher to enter all content and context associated with required information in a single simple flat query. This suggests that in the development of IR methods for PLs, we should focus on the need to support effective search of multi-field items using simple flat queries.

Given a simple flat query for search, the key research challenge is how to score the individual fields individually or in combination to generate the most effective overall score for retrieval. This issue is analyzed in detail in [Ogilvie and Callan, 2003, Robertson et al., 2004, Wilkinson, 1994]. The simplest approach to this is to index all item fields as one field, hence reducing the collection to a single field collection on which queries can be processed using a content only retrieval approach, such as the VSM or BM25, described in Section 4.2.1. This approach acts as a flat query based retrieval baseline for the PL retrieval investigations in Chapter 5. However, by reducing multi-field documents to single field documents the rich information in structured documents is lost. For example, consider a news articles archive with fields such as content, title, author and a query looking for articles on a given topic by a given author. The presence of the queried for author field of a structured document would help narrow the search space, however in a flat document this author information is essentially lost, or not deemed as significant, amongst the many terms that are in the flat representation of the document. Similarly, the significance of terms in a document's title field will be lost when a structured document is converted to a flat representation. Hence, it is desirable to maintain the structure of documents.

Much of the research exploring the challenge of retrieval from structured documents using flat queries has focused on using content only retrieval algorithms (described in Section 4.2.1) to determine field weights (where each field is queried against all the terms in the flat query) and then using various approaches to combine the individual field scores [Chowdhury et al., 2003, Savoy and Rasolofo, 2003, Xu et al., 2003], the simplest of these being the linear sum of individual field scores, described in the context of structured queries in Section 4.2.2.1. However, as highlighted in [Robertson et al., 2004] for flat queries simple data fusion of multi-field document approaches suffers a number of weaknesses[3]. In particular, linearly combining scores across fields can lead to a great over-estimation of the importance of terms occurring in multiple fields at the expense of a term occurring in only one field. They also highlight that careful exploitation of field structure is important for optimal retrieval performance. Their proposed solution, BM25F, is to weight terms at the field level and linearly combine these weights. This overall term weight is then applied to the BM25 saturating function. In subsequent work, they refined their BM25F term scoring approach to also consider field length at the term scoring level [Zaragoza et al., 2004], to account for fields of extremely different length, e.g. title and content fields. The BM25F term scoring approach [Zaragoza et al., 2004] for calculating the weight $\bar{w}_{t,d}$ of term *t* in document *d* is presented in Equation 4.16. The term weight $\bar{w}_{t,d}$ is applied to the BM25 saturating function, presented in Equation 4.17.

$$\bar{w}_{t,d} = \sum_{f \in d} \frac{tf_{d,f,t} \cdot w_f}{((1 - b_f) + b_f \cdot \frac{l_f}{avl_f})} \tag{4.16}$$

where,

tf$_{d,f,t}$ is the frequency of term *t* in field *f* of document *d*.

l$_f$ is the length of *f* in *d*.

avl$_f$ is the average length of field *f*.

w$_f$ is the field weight assigned to *f*.

b_f is a length normalising parameter for *f*.

[3] This makes no claim as to the utility, or lack there of, of using a simple data fusion approach for structured queries.

$$ms(q,d) = \sum_{t \in q \cap d} \frac{\bar{w}_{t,d}}{k_1 + \bar{w}_{t,d}} \cdot idf(t) \qquad (4.17)$$

where,

k_1 is a saturating parameter.

idf(t) = $log\frac{N-df(t)+0.5}{df(t)+0.5}$ in the BM25F implementation [Robertson et al., 2004, Zaragoza et al., 2004], where N is the number of documents in the collection, *df(t)* is the number of documents containing term *t*. Of course other approaches can be used to calculate the idf, as discussed in Section 4.2.1.

BM25F provides a simple but effective state-of-the-art solution to flat querying on multi-field documents, and we investigate its utility in PL retrieval in Chapter 5.4. However, it should be noted that BM25F was developed for flat querying on collections where query terms may match any number of fields, e.g. abstract, title and main body of text fields, and that its utility was shown for web retrieval [Zaragoza et al., 2004] and email content retrieval, combined with other query independent measures [Craswell et al., 2005b]. Our multi-field lifelog collections are different in that the fields from a querying point of view are independent of each other in that different fields do not have common meaning (with the exception of the title and content fields). Mapping of query terms to their target field may therefore be important in flat content+context retrieval approaches for the lifelogging domain. In related work Kim et al [Kim et al., 2009, Kim and Croft, 2009] argue that the importance of individual terms to individual fields should be captured in the term's weight when searching semi-structured documents. Clearly an approach using structured queries where the user enters search terms for fields separately represents one extreme where the term is only searched for in this field and its presence in other fields makes no contribution to the score. However, as noted previously users generally prefer to enter simple flat queries, which while they know the expected importance of terms to fields, e.g. the name of a place in the location field, this is not captured in a flat query. In their work Kim et al [Kim et al., 2009, Kim and Croft, 2009] explore the mapping of flat query terms to semi-structured movie database and desktop collections, where fields from a querying view point have separate meaning. Their retrieval technique for known-item desktop search uses the content field, context fields related to the content of the item (specifically title and abstract), and item specific context data (specifically to, from, URL and modified date). The focus of their investigations was how to combine fields for scoring within the language modelling IR framework. They found beneficial, a term scoring adaptation which weights the term score for each field according to the frequency of the term in the given field relative to its frequency across all fields in the document, with the expectation that this maps query terms to their target field. In so doing, they form a type of structured query where the presence of a query term is treated individually for each field. This process is referred to in the literature as *query transformation* [Croft, 2009], and is attractive since it introduces some of the strengths of structured queries, by mapping query terms to their target fields, without requiring users to understand or engage in the process of creating such queries. In Chapter 5.5 we examine the application of Kim et al's [Kim et al., 2009, Kim and Croft, 2009] query transformation approach to PL retrieval, and based on the findings of this study develop a novel approach to query transformation for PL retrieval.

4.2.3 Integrating Query Independent Evidence

To assist with item scoring in query driven retrieval, query independent evidence of item importance is sometimes considered. Query independent evidence (static scores) provides an indicator of the importance of a document in the collection as a whole (discussed in Chapter 2.3). Various approaches can be used for integrating static scores with query dependent scores.

Past research has investigated using static scores to re-rank the output of query driven result lists, whereby the static scores contribute to or determine the ordering of items in a result list. Examples here include, the desktop space where query driven result lists for desktop search are re-ranked using static link scores [Soules and Ganger, 2005, Soules, 2006], and web search where static features of web pages are used to contribute to the re-ranking of content-based search [Cai et al., 2004, Upstill et al., 2003].

In other work static scores are combined with query dependent scores. Chirita et al [Chirita, 2007] multiply various static scores in the desktop search space by query driven retrieval scores calculated using a tf×idf approach. Kraaij et al [Kraaij et al., 2002] also multiply static scores by query driven retrieval scores (calculated using language modelling approach) in the web space. Multiplying static scores by query driven scores results in items with high importance in the collection (detected through static scores) and high match with the user's query (detected through query driven scores) being moved to the top of result lists.

Others use a linear combination of query dependent scores and static scores, which results in boosting items query dependent scores by a factor which reflects their importance in the collection as a whole. In this linear approach either raw static scores can be used (e.g. [Kang and Kim, 2003]) or non-linear transformations of the static scores, e.g. the log of static scores [Upstill, 2004]. A particularly promising approach for calculating non-linear transformations of static scores is presented in [Craswell et al., 2005a] where a sigmoid functional form is used to transform PageRank scores, link indegree, ClickDistance and URL length into static scores.

$$w \cdot \frac{S^a}{k^a + S^a} \tag{4.18}$$

$$w \cdot \frac{k^a}{k^a + S^a} \tag{4.19}$$

This transformation is shown in Equations 4.18 and 4.19, where S is the raw static score and w, k and a are tuneable parameters. Equation 4.18 is used to transform static relevance scores for features where higher values indicate greater importance (e.g. PageRank). Equation 4.19 is used to transform static relevance scores for features where lower values indicate greater importance (e.g. ClickDistance on the web). In Chapter 7 we investigate the use of these and other transformations to convert biometric response into a static score for linear combination with a query dependent score in PL retrieval.

4.3 Created Test Sets and Test Cases for Ranked Retrieval Technique Development

In order to explore ranked retrieval techniques for the textual data in PL collections we indexed the textual data and associated context data in our three subjects lifelog databases and each subject created test cases (i.e. queries and target result items) for their collection. In the next section we describe the indexing process used to create our test sets. In Section 4.3.2 we describe the process used to generate the test cases and provide an analysis of the generated test cases. These test sets and test cases are used in our ranked retrieval technique investigations presented in Chapters 5 and 7.

4.3.1 Textual Test Set Indexing

In order to facilitate investigation of ranked retrieval techniques, presented in Chapter 5, on the textual items in our 3 subjects lifelog test sets, described in Chapter 3, we indexed the SMS messages, computer file, email and webpage accesses and associated context data (i.e., extension type; path to file; URL (for web pages only); to/from (for SMS messages and emails only); year; season; month; day of week; weekday or weekend; beginning of week, mid-week or end week; part of day (i.e., morning, afternoon, evening and night); begin date and time; end date and time; device (e.g., laptop, mobile phone); light status (i.e. daylight and dark); weather; geo-location; and people present) in each subjects collection. These we call the 20 month indexes. To facilitate the ranked retrieval approaches with static biometric scores investigated in Chapter 7, we also created indexes consisting of the SMS message, computer file, email and webpage accesses and associated context data[4] from the biometric month (described in Chapter 3.4) in subjects' collections. These we refer to as the biometric month indexes. In this section we describe how the 20 month and biometric month indexes were generated.

4.3.1.1 Indexing the Collections

Lucene[5], an open source search engine, was used to index SMS messages and computer item accesses and their associated context data into different fields (e.g. day of week field). We also included in these indexes an additional field consisting of all content+context associated with the item access[6]. Table 4.1 shows the complete set of fields included for each item access in our indexes.

Prior to indexing, the StandardAnalyzer built into Lucene was used to parse the con-

[4] Note: biometric data is not included in the indexes. As we will see in Chapter 7, biometric data is used as a query independent measure and is read directly from subjects lifelog database biometrics table (this database table was described in Chapter 3.2).

[5] http://lucene.apache.org/java/docs/ (September 2011)

[6] As we will see in Chapter 5, this field is required to investigate flat querying using BM25.The reader is referred back to Section 4.2.2.2 for full details on flat querying using BM25.

Item ID	Item content
Title (i.e., computer filename, email subject, webpage title)	Extension Type
Path to File	URL
To (for emails & SMS messages only)	From (for emails & SMS messages only)
Begin Date	Begin Time
End Date	End Time
Year	Season
Month	Day of Week
Weekday or Weekend	Part of Week (i.e., begin week, midweek, end week)
Part of Day	device
Geo-Location	Light Status
Weather	People Present
Item content + all above context fields	

Table 4.1: List of fields indexed for each textual lifelog item access.

tent and context fields. The StandardAnalyzer tokenizes text based on a sophisticated grammar that recognises email address, acronyms, alphanumeric and more; converts to lowercase, and removes stopwords using it's inbuilt stopword list. It was necessary for us to parse all fields using this approach, since in the flat ranked retrieval investigations presented in Chapter 5 we have no way of knowing which field each query term corresponds to and hence we parse all query terms using Lucene's StandardAnalyzer. That is, since all terms/fields in a query are tokenized, converted to lowercase and stop words removed, for consistency we also performed the same operations on all fields in the indexes. The terms in many collections are also stemmed as part of the indexing process. However, we found stemming our collections resulted in overall inferior retrieval performance; hence we did not stem the terms in our collections. We examined retrieval performance on the stemmed and unstemmed collections using the biometric month indexes and corresponding test sets (described later in this Chapter in Section 4.3.2.3). The PorterStemmerAnalyzer built into Lucene, which stems the text that has passed through Lucene's StandardAnalyzer was used to stem the content and context fields of the queries and indexes. In this examination we compared average precision (AveP), precision after 5 documents retrieved (P@5) and precision after 10 documents retrieved (P@10) on the stemmed and unstemmed collections using each of the retrieval approaches that will be investigated in Chapter 5. The resulting decrease in performance observed using the stemmed collections is likely due to the loss of the context of some searches. For example, taking 'computers' as a content field query term, in a stemmed collection this would also match the query terms ’compute’ and ’computing’. Stemming would also decrease the distinguishing power, in flat queries, of context query terms. For example, given a flat query containing the term ’web’ intended for the extension type context field, in a stemmed collection this query term would incorrectly match occurrences of the term ’web’ and ’webs’ in the content field of items. Whereas in an unstemmed collection it would only incorrectly match occurrences of the term ’web’.

Since we wished to explore the use of BM25 and BM25F in lifelog retrieval, our indexes needed to store the average length of fields (see Sections 4.2.1.1 and 4.2.2.2 for the use of average field length in BM25 and BM25F). However, standard Lucene does not support the BM25 and BM25F retrieval algorithms, and therefore does not store the average length of fields. Hence during the indexing process we stored the average length of each field in a structured text document using the CollectionSimilarityIndexer provided with the open source BM25&BM25F Lucene library [Pérez-Iglesias et al., 2009]. We then also used this open source library when investigating the use of BM25 and BM25F in content+context-based lifelog retrieval, described in Chapter 5.

The BM25 implementation in this library did not include the multiplication of the term weight by the *(k_1+1)* parameter. This parameter is used to normalize the weight of query terms with frequency of 1 in average length fields, hence we edited this BM25 implementation to include this *(k_1+1)* parameter. The resulting BM25 implementation is shown in Equation 4.12. The BM25F implementation provided with the library is that shown in Equation 4.17.

4.3.2 Textual Test Case Creation

As highlighted in Chapter 2.4 test case creation in the lifelogging domain is challenging due to the personal nature of lifelogs, and the need for test case creation to attempt to mimic the ‘real’ re-finding requirements of individuals, and details that they are likely to recall about required items as closely as possible. Hence, in generating test cases for our experiments we first asked our lifelog owners to list search tasks they would be likely to perform on their collections and the content+context they recalled for the required items. This process is described in the next section. We then asked subjects to indicate the items from their collections which were relevant for these search tasks, described in Section 4.3.2.2. Using this approach we believe we are obtaining a realistic approximation to 'real' re-finding requirements of lifelog owners, a good spread of the type of search tasks they may perform on their collections and the content and context they recalled about these items at a given moment in time. These test cases are used in our ranked retrieval investigations presented in Chapter 5. We also used a subset of these test cases, described in Section 4.3.2.3, to examine the utility of our biometric static scoring functions, presented in Chapter 7. A detailed breakdown of the generated test cases is provided in Section 4.3.2.4.

4.3.2.1 Query Generation

The following procedure was used to generate 100 search tasks and content+context-based queries for these tasks for each of our 3 test subjects:

1. Half way through the lifelog collection build up process subjects were provided with an excel form containing a field to enter retrieval tasks they might want to perform in the future from their lifelogs and several fields to enter query terms associated with the retrieval tasks. Subjects were instructed to complete the provided form, in their own time, by listing 50 retrieval tasks they might want to perform in the future from their lifelogs and keywords and recalled context associated with the tasks. Further details on the provided form and instructions given to subjects follow.

    (a) Listing retrieval tasks: Subjects were aware of the reason for listing retrieval tasks and informed that there was no restriction on the type of retrieval tasks that they could list, but that they should be tasks they imagined they might have a need or desire to perform in the future. Subjects were also informed that the item types available for retrieval were computer files, web pages viewed, emails sent or received, and SMS messages sent or received during the lifelogging period, and that where possible examples of each of these item types should be included in the list of 50 tasks. Sample retrieval tasks were provided and a list of some possible general topics for task formation. This list of topics were for our research student subjects, details related to a conference attended or trip taken, papers written or read, details related to development work. Subjects managed to freely recall 35-40 retrieval tasks. To complete the list of 50 retrieval tasks subjects were then free to browse their computer folder structure to 'jog their memories' on the type of activities they had engaged in during the lifelogging period, following which they were able to complete the list of 50 retrieval tasks. Typical retrieval tasks generated by subjects were: 'show me documents I created associated with conference X', 'show me SMS messages on topic Y'.

(b) Entering query terms: The provided Excel form contained, in addition to a field to list task descriptions in, the following fields: content keywords; title keywords; extension type(s); file location; date(s); time(s); year(s); season(s); month(s); day(s) of week; part(s) of week; part(s) of day; light status; people present; device; weather; geo-location. A sample completed entry was provided in the form along with details on the context types associated with each context field. Subjects were also free to ask the investigator questions about the form. Subjects were instructed to enter keywords and remembered context, e.g. extension type, associated with the retrieval scenario into the provided form.

2. Subjects returned the completed Excel form (minus the task descriptions for privacy reasons) to the investigator, who manually transformed the query terms (i.e. the recalled content and context) into the format required for querying, for example extension type 'word documents' and 'doc' were changed to the extension type 'word'. In the future, such mappings of written query text could be automatically transformed to the required format, however automatic query transformation extended beyond the scope of our current work. Transformed queries were assigned unique IDs and stored in a structured table in subjects' lifelog databases, for use by the ranked retrieval experiments described in Chapters 5 and 7.

3. At the end of the lifelog collection build up process an additional 50 retrieval tasks and queries were generated for each subject's lifelog using the process described in steps 1-3 above.

Queries generated by subjects included:

- Content: java regular expressions, Title: java regular expressions, Year: 2009, Extension: web, Month: October, Season: Autumn, Part of Week: Weekday, Part of Day: morning afternoon, Light Status: daylight, Device: laptop, Geo-location: Dublin.

- Content: transfer report, Extension: word, Month: July August October, Season: Summer Autumn, Location: Dublin, Device: PC, Year: 2008.
- Content: elsweiler, Extension: pdf, Location: Dublin, Year: 2009.
- Content: virtual memory eclipse java IDE, Extension: web, Month: November December, Season: Winter, Location: Dublin, Device: laptop, Year: 2008.
- Content: examples csvreader java, Extension: text web, Month: August September December, Year: 2008.
- Content: lifelogging, Extension: web, Year: 2008.

4.3.2.2 Result Set Generation

In this section we describe the means by which we created lists of relevant items for the user queries created in the previous section.

Using the generated queries described in the previous section, pooled result lists were created by entering content only, context only, content+extension type, and content+context query types into each subject's 20 month test set indexes (these indexes were described in Section 4.3.1) using two standard retrieval systems: the vector space model (VSM) and BM25 (see Section 4.2.1 for full details on VSM and BM25), to retrieve as many relevant items from subjects' collections as possible. As is standard in IR experiments in the TREC tradition, the top 1000 results were taken in each case. The Lucene implementation of the VSM[7] and an open source implementation of BM25 for Lucene[8] were used to process these queries. Queries were parsed using Lucene's StandardAnalyzer, described in the previous section's index creation process. Queries combining content and context were straightforward concatenations of the content data score with the individual context types scores (see Section 4.2.2.1 for full details on this retrieval approach). The results from each of the BM25 and VSM content and/or context retrieval techniques were pooled and the top 500 pooled results for

[7] See http://lucene.apache.org/java/2_3_0/api/core/index.html (September 2011) for full details on the Lucene implementation of VSM.

[8] See [Pérez-Iglesias et al., 2009] for full details on this implementation of BM25

each query presented to the relevant subject for relevance judgment. A binary relevance judgement was used, i.e., 0 = irrelevant; 1 = relevant. These were recorded in an Excel form which contained an entry for each query, the retrieval scenario followed by the title, contents, file path/URL, to/from information (for emails and SMS messages) and extension type of the items to be judged as relevant or irrelevant to the retrieval scenario/query. The items rated as relevant for each query by the subject formed the result set for their queries. As will be seen in the ranked retrieval experiment in Chapters 5 and 7, TrecEval for Windows OS[9] was used to examine retrieval performance. TrecEval requires the subject's result set to be stored in a structured text file. Details stored include the query IDs and the unique IDs of relevant items for the queries. Our subjects result sets were written to files structured in the manner required by TrecEval.

4.3.2.3 Biometric One Month Test Cases

In order to examine the utility of our biometric static scoring functions, presented in Chapter 7, it was necessary to generate test cases for the biometric month test set. The result set for the queries used needed to be items which were accessed during the biometric response capture month, in order to allow us examine the impact of adding static scores to these items. Hence, we used the subsets of each subject's 100 test cases which contained relevant items occurring during the biometric capture month. That is, our biometric one month test cases consisted of the queries which contained relevant items occurring during the biometric capture month, and the result sets for these queries were the relevant items for the queries which were accessed during the biometric capture month. Subject 1 had 22 such test cases, Subject 2 had 8 and Subject 3 had 36. These test cases and the biometric month test set are used for retrieval algorithm parameter tuning, described in Chapter 5.2.

4.3.2.4 Test Case Contents Analysis

Table 4.2 shows the total number of relevant items across the 100 queries for each subject's test cases and the average number of relevant items per query. As can be seen the total number of relevant items across the 100 queries in these test cases, was: for Subject 1 942 items; for Subject 2 244 items; and for Subject 3 3,067 items. This

[9] http://www2.sims.berkeley.edu/academics/courses/is240/s05/trec_eval.zip (September 2011)

	Subject 1	Subject 2	Subject 3
Rel Items	942	244	3,067
Ave	9.42	2.44	30.67

Table 4.2: Total number of relevant items (Rel Items) per subject across the 100 test case queries and average number of relevant items per query (Ave).

	Subject 1	Subject 2	Subject 3
Rel Items	154	69	560
Ave	7	9.86	15.56

Table 4.3: Total number of relevant items (Rel Items) per subject across the biometric month test case queries and average number of relevant items per biometric month query (Ave).

corresponds to an average of 9.42 relevant items per query for Subject 1, an average of 2.44 relevant items per query for Subject 2 and an average of 30.67 relevant items per query for Subject 3. The particularly high average of relevant items for Subject 3 is explained by the fact that this subject performed very broad searches, for example 'I'm looking for all code files on topic X', 'I'm looking for all details related to a conference I attended'. Comparatively Subjects 1 and 2 performed much narrower searches, hence the lower average observed for these subjects. Of the queries generated by subjects, 10 of Subject 1's, 21 of Subject 2's and 3 of Subject 3's were targeted at retrieving only one relevant item. Subjects' remaining queries were targeted at retrieving multiple relevant items.

Table 4.3 shows the total number of relevant items across the queries in the subjects' biometric month test cases and the average number of relevant items for these queries. As can be seen the total number of relevant items across the biometric month queries was: for Subject 1 154 items; for Subject 2 69 items; and for Subject 3 560 items. The average number of relevant items occurring across these tasks during the biometric month was 7 for Subject 1, 9.86 for Subject 2 and 15.56 for Subject 3. The lower averages observed here for Subjects 1 and 3 relative to those observed on the 100 test cases is explained by the fact that not all relevant items associated with queries occurred during the biometric month (recall, we are only using the relevant items for queries which were accessed in the biometric month in our biometric month result sets, as described in the previous section). The higher average observed for Subject 2 is caused by one query which had a relatively large number of relevant items, a lot of which were accessed in the biometric month. While these biometric month test cases are small, given the difficulties in generating test cases in the lifelogging do-

	Subject 1		Subject 2		Subject 3	
	Num	Ave	Num	Ave	Num	Ave
content keywords	99	7.59	96	3.08	99	5.27
title	61	2.08	15	1.67	23	3.52
extension	98	1.47	99	1.16	98	2.14
date	0	0.00	22	1.00	37	1.00
month	76	2.68	83	1.31	86	1.73
season	83	1.55	95	1.09	57	1.30
dayOfWeek	9	2.56	6	1.17	13	2.15
partOfWeek	25	1.00	54	1.00	6	1.33
partOfWeek1	4	1.00	3	1.00	10	1.30
partOfDay	31	1.68	31	1.03	15	1.73
timeRange	4	1.00	7	1.00	1	1.00
lightstatus	35	1.17	6	1.00	8	1.25
peoplePresent	16	1.63	17	1.35	34	1.44
region	76	1.41	94	1.09	93	1.13
weather	3	1.00	9	1.33	1	1.00
from	9	1.00	9	1.00	19	1.84
to	0	0.00	2	1.00	19	1.84
device	77	1.32	71	1.00	94	1.41
year	100	1.00	100	1.00	100	1.00
country	0	0.00	1	1.00	12	1.00

Table 4.4: Number of fields containing query terms (Num) and average number of terms per query field (Ave) for each subject across their 100 queries.

Extension	*code*	*dat*	*email*	*excel*	*pdf*	*powerpoint*	*SMS*	*tex*	*text*	*web*	*word*
Subject 1	6	2	10	6	16	10	2	13	14	38	26
Subject 2	6	0	15	0	25	8	9	0	0	31	21
Subject 3	15	0	61	2	40	5	11	0	3	43	30

Table 4.5: Number of uses of extension types in queries per subject. Note the *code* extension type is an aggregate of the different code extensions searched for (e.g., java, mxml).

main, described in Chapter 2.4, they provide means for us to explore the use of static biometric scores in lifelog retrieval and means for us to compare different static scoring approaches, in Chapter 7, using 'real' user target items and recalled content and context associated with these items.

Tables 4.4 and 4.5 show the breakdown of the subjects' 100 queries. Specifically, Table 4.4 presents the number of queries that each context type was used in and the average number of query terms used in each field by each of our 3 subjects, and Table 4.5 presents the number of queries that each extension type was used in. These tables provide an insight into the use of context data in querying by subjects, which will be important when we move on to investigate retrieval algorithms for the lifelogging domain in Chapter 5. We return to these tables in Chapter 5 when we are analysing retrieval performance using different context types. For now we provide an overview of the makeup of users queries exhibited through the statistics in these tables.

As can be seen in Table 4.4 different recall processes are exhibited by the subjects with differing importance levels being placed on various context types. Subject 1 exhibits longer text searches relative to the other subjects. This subject also makes far greater use of the facility to search based on title (recall that title is the filename of computer files, the subject of emails, and the title of web pages). In contrast Subject 2 uses the least number of words to form queries and makes the least use of titles in the querying process. Subject 3 performs the broadest searches, evidence for this is shown in the larger number of extension types used per query. This is also shown in the query types performed by this subject, queries for all travel information associated with attending a conference for example, as opposed to a query for hotel information for a location visited in the past. This subject also makes greater use of date ranges and of the device field in their queries than the other subjects. Light status is most used by Subject 1. Season, part of week, time range and weather are recalled more by Subject 2. Full path to files and URLs were not recalled by subjects and hence are not included here. For the remainder of this thesis the use of the term 'context data' refers to all the context types listed in Table 4.4. These findings show that different types of context data will be more or less important for different subjects in the retrieval process, but also show the important role context data plays in subjects' recall of items required from their lifelogs. However, it should be noted that the volumes of recalled context data was not distributed evenly across queries. Some queries had very little associated context data (e.g., extension type and year only, extension type, year and month range of previous access only), while others had several rich sources of recalled context. Further, it should be noted that the nature of the query generation process described in Section 4.3.2.1 prompts subjects to enter the context types used in our investigations. In a flat querying approach, where individuals are not provided with context type prompts the same volume of context data might not be recalled.

As can be seen in Table 4.5, Subject 3 has a greater number of multi-type queries than the other subjects. Overall this subject queried for emails and PDFs to a far greater extent than the other subjects. Subject 1 performed the greatest number of queries for textual items authored by themself relative to the other subjects. Subject 3 carried out the greatest number of searches for code. Items which show interaction with others, namely SMS messages and emails are least searched for by Subject 1, while Subject

3 shows greatest amount of search for these types of items particularly in the case of emails. These observations give us insight into the item types queried for by subjects. Given the varying volumes of different information types in subjects' collections (shown in Chapter 3.4), different item types will be more or less useful in narrowing the search space in the retrieval process, for example we imagine that use of the extension type pdf will provide greater utility in narrowing the search space than use of extension type web (given the volumes of these item types which we saw to be present in subjects collections in Chapter 3.4).

4.4 Conclusions

Search of semi-structured lifelogs has similarities to both web search and standard desktop search. However, PLs have a greater number of more diverse fields available than seen in either of these scenarios. Also, as discussed in Chapter 3, real PL collections are likely to have large amounts of missing data arising from various sources including equipment and software failure during data gathering. In this chapter we overviewed existing retrieval techniques of interest to this research. In the next chapter we investigate the use of both structured and flat content+context retrieval algorithms in the lifelogging domain using our created 20 month textual test sets and test cases. Following this, in Chapter 6, we explore the potential utility of biometric response as an indicator of lifelog item importance, before moving on to investigate the potential utility of integrating biometric response as a static score into ranked retrieval algorithms using our biometric month test collections.

Part III

Experimentation

CHAPTER

FIVE

Queried Content-and-Context-Based Retrieval Algorithms for the Lifelogging Domain

Chapter Overview: In this chapter we investigate the utility of recalled content and context types in lifelog retrieval and various approaches to integrate recalled context with recalled content in retrieval algorithms. We first describe the setup of our experiments to investigate the role of recalled context data and various content+context retrieval algorithms in lifelog retrieval. Following this we present and analyse the results of structured content only and content+context retrieval algorithms using BM25, flat content+context retrieval algorithms using both BM25 and BM25F. We then explore developing novel extensions to BM25F for flat content+context querying in the lifelogging domain and present the results of this investigation.

5.1 Introduction

As highlighted in Chapter 1.1, items in PLs are personal to the individual PL owner, in that they have been created or obtained by the individual or represent one of their life experiences. Since this is the case, the individual will often have personal experiences and memories associated with the items in their archive. This means that items for retrieval from a PL consist of the item itself, e.g. a document or SMS message, and possibly available associated context information, e.g. the time and date when a document was viewed or the location from which an SMS message was sent. We seek to explore if there is benefit in allowing individuals query based on recalled context data associated with required lifelog items and the types of context data which might prove most beneficial in this process. However, content only search is a valid possibility to support PL search, and adding context to search (beyond filtering) might not improve retrieval performance. Content only based retrieval therefore forms a baseline for PL search. We hypothesise that including recalled context with content in the retrieval process will improve content-based retrieval. In this chapter we explore this hypothesis.

In integrating recalled context into the retrieval process it is not clear how best these multi-field semistructured documents should most effectively be scored for retrieval. Additionally, users of existing search engines show a strong preference for entering simple single field queries. This suggests that in the development of IR methods for PLs, we should examine methods to support effective search using simple flat queries. PL data and the preference for simple queries has strong similarities to the type of semi-structured data considered for retrieval in [Kim et al., 2009, Kim and Croft, 2009], as discussed in Chapter 4.2.2.2.

In this chapter we investigate the use of the textual content of items combined with multiple context fields in PL retrieval. BM25 content only retrieval (see Chapter 4.2.1.1 for details) is used as a baseline for this investigation. For field combination we apply BM25 [S. E. Robertson and S. Jones, 1994] (see Chapter 4.2.1.1 for details) to structured queries and also explore its utility across flat collections (see Chapter 4.2.2.2 for details). We also apply BM25F [Robertson et al., 2004] (see Chapter 4.2.2.2 for details) to flat content+context queries for field combination, and various novel extensions of BM25F based on the characteristics of the fields of PL items and the findings of

[Kim and Croft, 2009] (see Chapter 4.2.2.2 for details). The experimental procedure is presented in the next section. BM25 in PL fielded retrieval is investigated in Section 5.3. The retrieval effectiveness of BM25 and BM25F for flat querying is described in Section 5.4. Section 5.5 focuses on our novel extensions to BM25F for PL retrieval. Finally, in Section 5.6 we conclude the chapter.

5.2 Experiment Procedure

In this section we describe the setup of our experimental studies to examine performance of content only retrieval on PL collections. Content only retrieval then acts as our baseline to explore: (1) the role of recalled context data in lifelog retrieval; (2) the performance of existing structured and flat content+context retrieval approaches on PL collections; and (3) novel flat content+context retrieval approaches which we developed for the lifelogging domain. For simplicity we describe the setup of all these experiments in this section. Further details on the rational for conducting each individual investigation are provided in the description of each investigation later in this chapter.

Existing Querying Approach Selection

As we saw in Chapter 4.3.2.2 queries consisting of various combinations of content and context were used to obtain pooled result lists for subjects to rate, using both the vector space model (VSM) and BM25. Looking at the result lists generated using the VSM and those generated using BM25, we found that using BM25 retrieved more relevant results at top rank than when using the VSM. Given this observation and since BM25 forms a solid well proven retrieval approach, BM25 is used as the content only querying baseline (described in Chapter 4.2.1.1 and shown in Chapter 4 Equation 4.12) and for structured content+context querying (described in Chapter 4.2.2.1 and shown in Chapter 4 Equation 4.15). We also use BM25 on flat queries where all item fields are indexed as one field, hence reducing the collection to a single field collection (described in Chapter 4.2.2.2). For simplicity, in this chapter we refer to this approach as *flatBM25*. This acts as a flat query based retrieval baseline. The performance of BM25F (described in Chapter 4.2.2.2 and shown in Equation 4.17) on flat queries is also examined, since it is a state-of-the-art flat query retrieval approach.

Code Base

Standard Lucene[1], an open source search engine, and an open source library containing BM25 and BM25F implementations for Lucene [Pérez-Iglesias et al., 2009] were used in our investigations, and are described in Chapter 4.3.1.1. We edited the libraries BM25F implementation to include our novel extensions to BM25F described in this chapter (Section 5.5).

Test Collection

Our generated indexes of the textual data in our 3 subjects 20 month lifelogs, described in Chapter 4.3.1, and the 100 queries and result sets generated by each subject, described in Chapter 4.3.2, were used in these experiments. Since the textual test set was parsed using the StandardAnalyzer built into Lucene, user content+context queries were also parsed using this StandardAnalyzer. Details of this process and Analyzer are available in the description of the test set indexing process in Chapter 4.3.1.

Parameter Tuning

FlatBM25 parameters and both BM25 and BM25F field weights and parameters were manually tuned using a one month subset (the biometric month test set index, described in Chapter 4.3.1) of the subjects' 20 month test sets and the test cases for which relevant items occurred during this one month period (the biometric month test cases, described in Chapter 4.3.2.3). To determine the best settings in the tunings we used average precision, precision after 5 documents retrieved (P@5) and precision after 10 documents retrieved (P@10). The absolute parameter values may vary depending on the order they are optimized in, but variation on retrieval performance should be minimal. Exhaustive testing of all possible combinations was not feasible. We adopted the following parameter tuning approach.

The BM25 b and k_1 parameters were tuned for flatBM25 based retrieval by setting the b parameter to 1 and tuning the k_1 parameter to give overall best retrieval performance across the 3 subjects. With the k_1 parameter set to it's tuned value, the b parameter was then tuned to give overall best retrieval performance across the 3 subjects. This process resulted in the b parameter being tuned to a value of 0.8 and the k_1 parameter being tuned to 4.0.

In the case of BM25 parameter tuning for fielded retrieval, we first tuned the content field's parameters by not considering the other fields in the retrieval process. To do

[1] http://lucene.apache.org/java/docs/ (September 2011)

this content field parameter tuning, the content field weight and b parameter were set to 1 and the k_1 parameter tuned to give overall best retrieval performance across the 3 subjects. The b parameter was then tuned to give overall best retrieval performance across the 3 subjects by setting the weight to 1 and the k_1 parameter to its tuned value. Note, the content fields weight parameter (w) was not tuned at this point as more than one field (i.e. more than the content field) needs to be considered in the retrieval process to tune a field's weight.

We then considered the title field combined with the content field in retrieval (i.e. content+title retrieval), to allow us tune the title fields weight and parameters. In tuning the title fields weight and parameters, we set the content field weight to 1 and the b and k_1 parameters to their tuned values. The procedure for tuning the title fields parameters began by setting the title field weight and b parameter to 1 and tuning the k_1 parameter to give overall best retrieval performance across the 3 subjects. The b parameter was then similarly tuned by setting the title fields weight to 1 and the title fields k_1 parameter to its tuned value. Finally, the weight was similarly tuned by setting the title fields k_1 and b parameters to their tuned values.

We then added the extension field to the content and title fields in retrieval (i.e. content+title+extension retrieval) and tuned the extension fields weight and parameters using the same technique as that used for title field weight and parameter tuning.

All remaining fields weights and parameters were tuned using the same approach as that used for the extension field, by iteratively adding these fields to the retrieval process and tuning their weights and parameters. Having tuned all context fields weights and parameters, we then tuned the content fields weight with it's k_1 and b parameters set to their previously tuned values, and using all context fields and their tuned parameters in the retrieval process, to give overall best retrieval performance.

The weight (w) assigned to each field when using BM25 with the corresponding k_1 and b parameter tunings for each field is shown in Table 5.1. As can be seen in Table 5.1, with the exception of the content and title fields, all fields k_1 and b parameters were tuned to 1. This means that for fields of length 1 and average length of 1, at the term scoring level, these fields' term weights *w(i,j)* reduce to tf×idf with no document length compensation.

To tune BM25F field weights and parameters we used a similar approach to that

Field	**w**	**k**	**b**
content	1	3	0.9
title	5	2	0.75
extension	0.5	1	1
year	0.01	1	1
month	0.1	1	1
date	0.2	1	1
time range	0.1	1	1
season	0.05	1	1
day of week	0.08	1	1
part of week	0.05	1	1
part of week 1	0.1	1	1
part of day	0.05	1	1
light status	0.01	1	1
from	0.3	1	1
to	0.3	1	1
device	0.05	1	1
people present	0.5	1	1
weather	0.3	1	1
region	0.1	1	1
country	0.1	1	1

Table 5.1: BM25 k, b and weight (w) parameter tuning for fields used in retrieval.

used for BM25 fields weights and parameters tuning. That is, in parameter tuning, we first tuned using content only retrieval, then content+title retrieval, then content+title+extension retrieval, etc. Given that the BM25F k_1 parameter was field independent, we re-tuned this parameter after each field tuning (i.e. k_1 was tuned for content only retrieval, and then retuned for content+title retrieval, content+title+extension retrieval, etc). The BM25F k_1 parameter was tuned to 3.0. The weight (w_i) and b_i parameter assigned to each field (i) when using BM25F are shown in Table 5.2. As can be seen with the exception of the content field, each field's b parameter was tuned to 1. This means that for the fields of length 1 and average length 1, the term score for the fields reduces to *tf*.

Choosing the Optimal IDF approach

As described in Chapter 4.2.1 various approaches can be used to calculate the idf score used to weight query terms appearing in documents/fields. Having tuned parameters using the 'default' BM25 and BM25F idf calculating approaches (i.e. $idf = log \frac{N}{df(t)}$ in the case of BM25 [S. E. Robertson and S. Jones, 1994] and $idf = log \frac{N-df(t)+0.5}{df(t)+0.5}$ in the case of BM25F [Zaragoza et al., 2004], presented in Chapter 4 Equations 4.8 and 4.10), as described above, we then examined the retrieval performance obtained using each of the idf approaches presented in Chapter 4 Equations 4.8 to 4.10 in the BM25 content,

Field	$\mathbf{w}_i$	$\mathbf{b}_i$
content	0.2	0.9
title	4.0	1.0
extension	1.5	1.0
year	8.0	1.0
month	0.5	1.0
date	1.0	1.0
time range	0.2	1.0
season	1.0	1.0
day of week	0.01	1.0
part of week	3.0	1.0
part of week 1	1.0	1.0
part of day	1.0	1.0
light status	0.1	1.0
from	1.0	1.0
to	2.0	1.0
device	0.001	1.0
people present	1.0	1.0
weather	1.0	1.0
region	0.001	1.0
country	1.0	1.0

Table 5.2: BM25F b_i and weight (w_i) parameter tuning for fields used in retrieval. k_1 parameter was tuned to 3.0.

BM25 content+context, flatBM25, BM25F and our extensions to BM25F retrieval algorithms on our test collection. Use of the idf approach presented in Equation 4.8 ($idf = log\ \frac{N}{df(t)}$) in each of these retrieval algorithms gave overall best performance. Hence, in this chapter we present the results obtained using this idf approach.

Evaluation Metrics

As is standard in IR experiments, the top 1000 results were taken in each case. Recall that during the indexing process (described in Chapter 4.3.1), we index all accesses to items to allow for retrieval based on details recalled about an access to an item. In examining the utility of our retrieval approaches we then merge the result set. That is, we only maintain the highest ranked occurrence of each item ID in the result set. The rank of the relevant items in the (merged) result sets was noted. Average precision (AveP), precision after 5 documents retrieved (P@5) and precision after 10 documents retrieved (P@10) were investigated for BM25 content only retrieval, structured BM25 content+context retrieval, flatBM25 content+context retrieval, BM25F content+context retrieval and retrieval performance on our modified versions of BM25F. These evaluation metrics were chosen since the majority of subjects' queries were targeted at retrieving multiple relevant items, as discussed in Chapter 4.3.2.4. TrecEval for Win-

Technique	Subject 1			Subject 2			Subject 3		
	AveP	P@5	P@10	AveP	P@5	P@10	AveP	P@5	P@10
Content	0.230	0.228	0.193	0.149	0.098	0.084	0.404	0.508	0.457
C+C	0.314 (37%)	0.281 (24%)	0.235 (22%)	0.177 (19%)	0.102 (5%)	0.091 (8%)	0.459 (14%)	0.575 (13%)	0.521 (14%)

Table 5.3: BM25 content and content+context (C+C) retrieval results, using subjects' 20 month collection and full set of 100 queries, for average precision (AveP), P@5 and P@10. Percentage improvement of BM25 C+C over the BM25 content only baseline shown in brackets.

dows OS[2] was used to calculate AveP, P@5 and P@10. Sections 5.3 - 5.5 present the results of these experiments.

5.3 Investigation 1: Structured BM25 Content+Context-Based Retrieval - The Utility of Recalled Context in Lifelog Retrieval

In this section the utility of using recalled context data in lifelog retrieval is examined on our 3 subjects' PL collections. We take content only retrieval using BM25 as our baseline and analyse the impact of using recalled context data for retrieval from each subject's collection by adding the queried context scores to the base content score (i.e. each of the item's context types were given appropriate weights and then summed). Here a structured approach is used, whereby each context field is queried separately using the recalled term(s) for that field. This retrieval algorithm is shown in Chapter 4 Equation 4.15. This approach acts as our content+context retrieval baseline for the other content+context retrieval approaches which will be investigated in this chapter. While the results observed on our 3 subjects collections cannot be generalised to the wider populous, they do provide unique insights into querying using recalled context data.

5.3.1 Results and Analysis

Table 5.3 presents the AveP, P@5 and P@10 results obtained using BM25 content only retrieval and using BM25 content+context retrieval averaged across the 3 subjects. P@5 and P@10 show how effective our techniques were at moving relevant items to-

[2] http://www2.sims.berkeley.edu/academics/courses/is240/s05/trec_eval.zip (September 2011)

	Subject 1			Subject 2			Subject 3		
content+	**AveP**	**P@5**	**P@10**	**AveP**	**P@5**	**P@10**	**AveP**	**P@5**	**P@10**
content only	0.230	0.228	0.193	0.149	0.098	0.084	0.404	0.508	0.457
+context	0.314	0.281	0.235	0.177	0.102	0.091	0.459	0.575	0.521
+title	0.302	0.267	0.228	0.154	0.093	0.080	0.441	0.556	0.502
+extension	0.246	0.251	0.205	0.158	0.098	0.091	0.411	0.521	0.463
+date	0.230	0.228	0.193	0.161	0.098	0.084	0.408	0.515	0.460
+year	0.232	0.228	0.193	0.150	0.098	0.084	0.405	0.510	0.458
+season	0.230	0.228	0.193	0.150	0.098	0.087	0.404	0.508	0.459
+month	0.231	0.228	0.193	0.139	0.098	0.084	0.406	0.508	0.458
+part_of_week	0.231	0.228	0.193	0.150	0.098	0.087	0.404	0.508	0.457
+part_of_week1	0.231	0.228	0.193	0.150	0.098	0.087	0.404	0.508	0.457
+part_of_day	0.230	0.226	0.193	0.150	0.098	0.087	0.404	0.508	0.457
+lightstatus	0.231	0.228	0.193	0.150	0.098	0.087	0.404	0.508	0.457
+peoplepresent	0.231	0.228	0.193	0.150	0.098	0.087	0.405	0.515	0.458
+region	0.231	0.228	0.193	0.149	0.098	0.084	0.404	0.508	0.458
+country	0.230	0.228	0.193	0.150	0.098	0.087	0.404	0.508	0.457
+to	0.230	0.228	0.193	0.152	0.102	0.087	0.406	0.513	0.459
+from	0.231	0.228	0.193	0.155	0.102	0.089	0.406	0.513	0.458
+device	0.231	0.228	0.193	0.150	0.098	0.087	0.404	0.508	0.456
+timerange	0.231	0.228	0.193	0.150	0.098	0.087	0.404	0.508	0.457
+weather	0.231	0.228	0.193	0.150	0.098	0.087	0.404	0.508	0.457
+dayofweek	0.231	0.228	0.193	0.150	0.098	0.087	0.404	0.508	0.456
+title+extension	0.314	0.279	0.236	0.162	0.093	0.087	0.448	0.567	0.506
+all “date-time” information	0.231	0.228	0.193	0.149	0.098	0.082	0.411	0.519	0.462
+region+country	0.231	0.228	0.193	0.149	0.098	0.084	0.404	0.508	0.458
+to+from	0.231	0.228	0.193	0.155	0.102	0.089	0.407	0.515	0.462

Table 5.4: Individual subjects’ results, on their 20 month collections using full set of 100 queries, for average precision (AveP), P@5 and P@10 by adding context types to the content only retrieval baseline.

wards the top of the result lists. The percentage improvement over the content only baseline for content+context retrieval averaged across the 3 subjects is also provided in Table 5.3. As was expected, the addition of queried context data to content-based queries improves retrieval performance in our subjects’ collections, since it helps disambiguate it with respect to the content field. Indeed it greatly improves overall performance across the 3 subjects with 37%, 24% and 22% improvement in AveP, P@5 and P@10 respectively being observed for Subject 1, 19%, 5% and 8% improvement in AveP, P@5 and P@10 for Subject 2 and 14%, 13% and 14% improvement in AveP, P@5 and P@10 for Subject 3, as shown in Table 5.3. This is a positive result for the use of recalled context data in lifelog retrieval. While we cannot generalise to the entire populous based on 3 subjects collections it does provide initial support for our research hypothesis that allowing individuals to query based on their memories of the context

	Subject 1			Subject 2			Subject 3		
content+	**AveP**	**P@5**	**P@10**	**AveP**	**P@5**	**P@10**	**AveP**	**P@5**	**P@10**
+context	37%	24%	22%	19%	5%	8%	14%	13%	14%
+title	32%	17%	18%	3%	-5%	-5%	9%	9%	10%
+extension	7%	10%	6%	6%	0%	8%	2%	3%	1%
+date	0%	0%	0%	8%	0%	0%	1%	1%	1%
+year	1%	0%	0%	1%	0%	0%	0%	0%	0%
+season	0%	0%	0%	1%	0%	3%	0%	0%	1%
+month	1%	0%	0%	-6%	0%	0%	0%	0%	0%
+part_of_week	1%	0%	0%	1%	0%	3%	0%	0%	0%
+part_of_week1	1%	0%	0%	1%	0%	3%	0%	0%	0%
+part_of_day	0%	-1%	0%	1%	0%	3%	0%	0%	0%
+lightstatus	1%	0%	0%	1%	0%	3%	0%	0%	0%
+peoplepresent	1%	0%	0%	1%	0%	3%	0%	1%	0%
+region	1%	0%	0%	0%	0%	0%	0%	0%	0%
+country	0%	0%	0%	1%	0%	3%	0%	0%	0%
+to	0%	0%	0%	2%	5%	3%	1%	1%	1%
+from	1%	0%	0%	4%	5%	5%	1%	1%	0%
+device	1%	0%	0%	1%	0%	3%	0%	0%	0%
+timerange	1%	0%	0%	1%	0%	3%	0%	0%	0%
+weather	1%	0%	0%	1%	0%	3%	0%	0%	0%
+dayofweek	1%	0%	0%	1%	0%	3%	0%	0%	0%
+title+extension	37%	23%	22%	9%	-5%	3%	11%	12%	11%
+all "date-time" information	1%	0%	0%	0%	0%	-3%	2%	2%	1%
+region+country	1%	0%	0%	0%	0%	0%	0%	0%	0%
+to+from	1%	0%	0%	4%	5%	5%	1%	1%	1%

Table 5.5: Subjects' percentage improvement, rounded to nearest whole number, on their 20 month collections using full set of 100 queries, for average precision (AveP), P@5 and P@10 by adding context types to the content only retrieval baseline.

associated with required information is beneficial in the lifelogging domain. While these results are promising, the improvement of content+context search over content only search on the subjects collections are not statistically significant for AveP, P@5 or P@10 (100 samples, Welch two sample t-test, p>0.05), with the exception of the improvement in AveP for Subject 1 (100 samples, Welch two sample t-test, p<0.05).

While overall the addition of recalled context data greatly improves over content-based retrieval for our 3 subjects, certain context types had different impacts on performance. Table 5.4 presents a breakdown of results observed across each subject's 100 queries, providing details of the impact of the addition of each context type in isolation to content only retrieval. Table 5.5 provides the percentage improvement over content only retrieval. Table 5.6 (which is a copy of Table 4.4 presented in Chapter 4) shows the number of queries each context type was used in by the subjects. Overall,

	Subject 1		Subject 2		Subject 3	
	Num	**Ave**	**Num**	**Ave**	**Num**	**Ave**
content keywords	99	7.59	96	3.08	99	5.27
title	61	2.08	15	1.67	23	3.52
extension	98	1.47	99	1.16	98	2.14
date	0	0.00	22	1.00	37	1.00
month	76	2.68	83	1.31	86	1.73
season	83	1.55	95	1.09	57	1.30
dayOfWeek	9	2.56	6	1.17	13	2.15
part_of_week	25	1.00	54	1.00	6	1.33
part_of_week1	4	1.00	3	1.00	10	1.30
part_of_day	31	1.68	31	1.03	15	1.73
timeRange	4	1.00	7	1.00	1	1.00
lightstatus	35	1.17	6	1.00	8	1.25
peoplePresent	16	1.63	17	1.35	34	1.44
region	76	1.41	94	1.09	93	1.13
weather	3	1.00	9	1.33	1	1.00
from	9	1.00	9	1.00	19	1.84
to	0	0.00	2	1.00	19	1.84
device	77	1.32	71	1.00	94	1.41
year	100	1.00	100	1.00	100	1.00
country	0	0.00	1	1.00	12	1.00

Table 5.6: Copy of Table 4.4 which was presented in Chapter 4: Number of fields containing query terms (Num) and average number of terms per query field (Ave) for each subject across their 100 queries.

Subject 2 who had the lowest default AveP, P@5 and P@10 for content only retrieval, and who also made the least use of the title field (see Table 5.6) for querying benefited the most from the use of the other types of context data. These differences in results, and greater dependency on context data to improve retrieval performance for Subject 2 can be explained by the fact that Subject 2 had shorter content (keyword) query lengths than the other subjects (average query length of 3.08 terms versus average query length 7.59 terms and 5.27 terms for Subjects 1 and 3 respectively, as shown in Table 5.6). This coupled with their relatively lower use of the title field in querying resulted in a higher dependency by Subject 2 on the other types of context data to improve retrieval performance. This is a promising result for the use of context data in retrieval, as people's content queries often just contain a few terms. In the next section we take a more detailed look at the impact of using queried context data with queried content terms in PL retrieval.

5.3.1.1 Impact of Different Context Types on Retrieval

Analysing the individual context types we see that content + title field retrieval had the greatest impact on improving performance for Subjects 1 and 3. Use of title field had relatively little positive impact for Subject 2 (indeed, for P@5 and P@10 performance decreased by 5%, possibly caused by the use of title query terms which did not occur in required items), who did not make as much use of this field in retrieval. As shown in Table 5.6 only 15 of Subject 2's queries used the title field, with an average title field query length of 1.67 terms. Use of extension field had least impact for Subject 3 who entered the broadest queries (i.e. queries for multiple items and extension types), with an average extension field query length of 2.14. Use of extension field yielded 2% improvement in AveP and 3% and 1% improvement in P@5 and P@10 respectively for this subject. This compares to 7%, 10% and 6% improvement in AveP, P@5 and P@10 respectively for Subject 1 and 6%, 0% and 8% improvement in AveP, P@5 and P@10 for Subject 2. Using both title and extension field with content retrieval near full content+context retrieval performance was observed for Subjects 1 and 3 relative to Subject 2 (as shown in Table 5.5). That is the addition of further context types did not substantially improve performance for Subjects 1 and 3, whereas Subject 2 who used shorter keyword queries and who made less use of the title field relied more on richer context types to improve retrieval performance. This is reflected in the impact these context types taken in isolation had on content only retrieval in Subject 2's queries, relative to Subjects 1 and 3. This impact can be seen in Table 5.5 where for Subject 2 all context types taken in isolation, with the exception of region and month, yielded great improvement over content only retrieval relative to Subjects 1 and 3. We next analyse these results in greater detail.

To and From Fields:

The 'to' and 'from' fields were available to query on for the SMS messages and emails in the subjects' collections. In Chapter 3, Table 3.9 provided details of the numbers of these item types in each subjects collection. As can be seen (from Table 3.9), emails were one of the most prevalent item types in each subject's collection. In the case of Subjects 1 and 3, SMS messages were also one of the highest occurring item types in their collection.

Subject 3 performed the greatest number of queries using the 'to' and 'from' fields.

This subject executed 19 queries in total using the ’to’ field and 19 queries using the ’from’ field. Subject 2 used the ’to’ field in 2 of their queries and the ’from’ field in 9 of them. Subject 1 did not use the ’to’ field in their queries, but used the ’from’ field in 9 of them. These figures are shown in Table 5.6.

Use of the ‘to’ and ‘from’ fields with the content field proved beneficial, particularly in the case of Subject 2 where 2%, 5% and 3% improvement was observed for AveP, P@5 and P@10 respectively using the ‘to’ field and 4%, 5% and 5% improvement for AveP, P@5 and P@10 respectively using the ‘from’ field. Using both the ‘to’ and ’from’ fields in this subject’s queries did not improve over the results observed for the ’from’ field. Subject 1 did not use the ‘to’ field in their queries (see Table 5.6), however using the ‘from’ field yielded a 1% improvement in performance for AveP. For Subject 3, a 1% improvement in AveP, P@5 and P@10 was observed for the ‘to’ field and 1% improvement in AveP and P@5 for the ’from’ field; using both the ‘to’ and ‘from’ fields, yielded the same results as using the ’to’ field in isolation.

“Date-time” Fields:

As can be seen in Table 5.6 subjects had far greater recall of vague “date-time” information such as month, season and year, than more precise “date-time” information such as exact date and time. The use of individual “date-time” information types in isolation had little impact on improving retrieval performance for Subjects 1 and 3 who performed rich term based queries. No improvement in performance was observed for Subject 3 by including any of the “date-time” information context types in isolation in the queries, with the exception of use of the season and the ‘date’ context types. Use of the season context type here resulted in 1% improvement in P@10. Use of the ’date’ context type yielded 1% improvement in each of AveP, P@5 and P@10. However, using all “date-time” context types in the querying process provided greater utility in assisting to locate relevant items for this subject, with 2% improvement in AveP, 2% improvement in P@5 and 2% improvement in P@10.

Subject 1 did not use the exact ’date’ field in their queries (as shown in Table 5.6), hence content+date retrieval could not improve over content only retrieval for this subject. A 1% decrease in P@5 was yielded for Subject 1 using the part_of_day field, possibly caused by incorrect recall on the part of the subject. Minimal improvement in performance was observed for Subject 1 using the year, month, part_of_week, part_of_week1, timerange and day_of_week fields where a 1% improvement in AveP was noted using each of these context types in isolation with the content field. Using all “date-time” information in queries provided no further improvement in retrieval performance for this subject.

Use of the content field with the ’season’, ’part_of_week’, ’part_of_week1’, ’part_of_day’, timerange or ’day_of_week’ fields proved beneficial in moving relevant items towards the top of the result list for Subject 2, who relied more on recalled context data in the retrieval process due to the vaguer nature of their content queries. This subject also performed more precise “date-time” queries than the other two subjects. For example, using ’morning’ as query term for the ’part_of_day’ field, as opposed to ’morning or afternoon’ (see Table 5.6). 3% improvement in performance was observed for P@10 for each of the aforementioned “date-time” related context types. Using each of these context types a 1% improvement in AveP was also noted. 1% improvement in AveP was also observed for this subject using the ’year’ field. Use of the ’date’ field yielded an 8% improvement in AveP, but did not help move items to the top of the result list. Use of the month field resulted in a 6% decrease in performance, possibly caused by incorrect recall on the part of the subject. While overall, taken in isolation with content only querying, “date-time” context types improved retrieval performance for Subject 2, using all “date-time” information in queries did not prove useful for this subject, with 3% decrease in P@10 being observed.

Device Field:

Overall the ’device’ field had a low discriminating power within subjects collections. Given that the majority of their queries were for items on their laptops or PCs, and relatively few for items on their mobile phones which was more discriminating within their collections. Despite this, some improvement in performance over the content only baseline was noted using the device field in queries. Use of the ‘device’ field with the content field in retrieval proved most useful for Subject 2 who only ever used one device query term in their queries (e.g. using the query term ’laptop’, or query term ’PC’, but not both the query terms ’laptop’ and ’PC’ in a query), where 3% improvement in P@10 and 1% improvement in AveP was observed. A minimal 1% improvement in AveP was gained here for Subject 1 who sometimes used both the query terms ’laptop’ and ’PC’, for example, in their queries. No improvement over the content only baseline was yielded for Subject 3, who performed the largest number of queries using more than one device type as query terms.

Light Status Field:

Use of recalled ’light status’ with content in retrieval proved most useful for Subject 2, who used a single query term in the ’light status’ field for all queries in which they used this field (i.e. either ’daylight’ or ’darkness’). 3% improvement in P@10 and 1% improvement in AveP over content only retrieval was observed for this subject. A minimal 1% improvement in AveP was obtained for Subject 1 using the ’light status’ field in retrieval. On examining the make up of this subjects’ queries (see Table 5.6), we observed that they sometimes removed the discriminating power of ’light status’ in retrieval by entering both ‘daylight’ and ‘darkness’ as query terms in this field. Subject 3 had a higher rate of removing the discriminating power of ’light status’ in retrieval (also shown in Table 5.6). For this subject no improvement over content only retrieval was obtained for content+light_status-based retrieval.

Weather Field:

Use of weather information, which was only recalled for 3 queries by Subject 1, yielded a minimal 1% improvement in AveP for this subject. Weather conditions were only used in 1 of Subject 3’s queries. Here no improvement over content only retrieval was observed. These results can be partly explained by the large percentage of items missing weather data in the subjects’ collections and the imprecise nature of weather data, as discussed in Chapters 3.3.4.3 and 3.4. Subject 2 made a little more use of the ’weather’ field in their queries, recalling weather information for 9 queries. For this subject 1% improvement in AveP and 3% improvement in P@10 was observed. This is a surprising but promising result given the exceptionally low number of items annotated with weather information in Subject 2’s collection (as described in Chapter 3.4.1).

People Present Field:

Recall of people present when interacting with an item improved retrieval performance to a certain extent for all three subjects. 1% improvement in AveP was observed for Subject 1, 1% in AveP and 3% in P@10 for Subject 2, and 1% improvement in P@5 for Subject 3. Similar to the use of the ’light status’ field, retrieval performance might have been affected by the number of items in the subjects’ collections which were missing ’people present’ annotations (as discussed in Chapter 3.4.1). Another factor which could potentially have negatively impacted on retrieval performance here, is that the recalled people present may not have had Bluetooth activated on their mobile phones (which is required to detect that the people were present) at the points in time the subjects are recalling.

Geo-location Fields:

Use of context type ‘country’ with content in retrieval provided 1% improvement in AveP and 3% improvement in P@10 for Subject 2. This subject only used the ’country’ field in 1 query (see Table 5.6). The ’country’ recalled was a country in which the subject spent very little time. This fact means that the ’country’ query term had good discriminating power in the retrieval process. No improvement in performance was noted for Subject 3 using the ’country’ field. Subject 1 did not use the country field in their queries (shown in Table 5.6). The only improvement in retrieval performance observed using the ’region’ field was 1% in AveP for Subject 1.

Missing geo-location data tags, described in Chapter 3.3.4.3, most likely also impacted on the utility of the ’region’ and ’country’ fields observed here. As would the low granularity of these location features (described in Chapter 3.3.4.3)

5.3.2 Concluding Remarks

While we cannot read too much into the results obtained using our relatively small data and query sets, the use of context data appears to be beneficial in lifelog retrieval. Using all our context fields in the retrieval process provided a measurable improvement over content only retrieval in the results presented in this section.

As highlighted in this section, different subjects have different recall and perform different types of queries on their unique collections with differing make up. For example, searching based on title (i.e., filename for computer files, subject of email, title of webpage) was important for Subject 1, and Subjects 1 and 2 had a tendency to perform narrower searches than Subject 3. Different context types prove more useful for different subjects, showing the need to allow subjects to make use of the different types of context that they recall in their queries. In other words, a retrieval approach which caters for the different querying habits and recall of different individuals is required. We believe that our retrieval approach, which allows for querying based on many context types accommodates this.

In a 'real' use scenario, we believe it unlikely that subjects will enter query terms reflecting their recall for the many context types used in our study in all their querying tasks. Rather it is more likely that a few key query terms will be used for different content and/or context fields. Which and how many fields this may result in being used will most probably vary from query to query, subject to subject and the difficulty of the retrieval task (difficulty in obtaining the required information).

As discussed in Chapter 3.1, 'real' personal collections are likely to have large amounts of missing data arising from various sources including equipment and software failure during data gathering. For example, the problems observed in our test collections with geo-location and co-present Bluetooth device data logging which yielded geo-location, people present, lightstatus and weather context types, as discussed in Chapter 3.3.4.3, meant that a large percentage of items in subjects collections were not tagged with these context types. This can negatively impact on retrieval performance in two ways. Firstly, items for which a queried context type is missing would result in relevant item(s) not receiving a contribution to their matching score from the context field, and secondly, non-relevant items tagged with the queried context type would receive score contribution which the relevant item(s) did not receive. Coupled with this, the nature of the data sets does not exercise the potential of all contexts, e.g. geo-location where the subjects are not moving around much and our low granularity of location features (described in Chapter 3.3.4.3) means that their potential utility in longer-term collections or collections for more mobile individuals is not properly explored here.

Further we also acknowledge that in our experiments there is the potential for incorrectly recalled context on the subjects' part and poor choice of content query terms may also impact on retrieval behaviour here further complicating our attempts to understand the potential of context in PL search techniques.

Exploration of all the issues discussed in this conclusions section forms a significant body of research in its own right and hence is beyond the potential of our current research, and indeed not possible using the small number of data sets and queries available to us. What we are particularly interested in exploring, having established that recalled context data may be useful in retrieval, is its utility in simple flat queries. The next section begins our exploration of this topic.

5.4 Investigation 2: Flat BM25 and BM25F Content+Context-Based Retrieval

Given the utility of allowing our 3 test subjects to enter queries based on content and context associated with required PL items in a structured manner (as shown in Investigation 1), we wished to establish if similar benefit over content only retrieval could be obtained if users were allowed to enter their recalled content and context associated with required items in a simple flat query. As highlighted in [Robertson et al., 2004] and discussed in Chapter 4.2.2.2, for flat queries, linear combination of separate field scores obtained using a content only retrieval algorithm can lead to over estimation of the importance of documents where a term occurs across multiple fields. The content field and the title context field in our lifelog collections share the same vocabulary, and hence would be susceptible to such an over estimation were a linear combination of field scores to be used in extracting relevant documents for flat queries. The use of scoring at term level employed by BM25F (described in Chapter 4.2.2.2) overcomes this problem. While BM25F can be used effectively for flat querying on collections where terms across all fields are drawn from the same vocabulary, given that it is a state-of-the-art field combination algorithm and will not lead to over estimation on our lifelog collections where the title and content fields share the same vocabulary, we wished to explore how it would perform on our multi-field PL collections in which fields from a querying point of view are independent (with the exception of the title and content fields). As a baseline for our flat experiments, we also examine the use of the simplest approach to flat querying. In this approach all item fields are indexed as one field, hence reducing the collection to a single field collection on which queries can be processed using the BM25 content only retrieval approach (also described in Chapter 4.2.2.2). We refer to this approach as flatBM25.

The purpose of the investigation presented in this section then is to examine the performance of BM25 on flat lifelog collections containing multi-field information (i.e. to examine the performance of flatBM25) and of BM25F in fielded lifelog search. The performance of these approaches is compared to the BM25 content only retrieval and the BM25 structured query results described in the previous section.

Technique	Subject 1			Subject 2			Subject 3		
	AveP	**P@5**	**P@10**	**AveP**	**P@5**	**P@10**	**AveP**	**P@5**	**P@10**
BM25 (C)	0.230	0.228	0.193	0.149	0.098	0.084	0.404	0.508	0.457
BM25 (C+C)	0.314	0.281	0.235	0.177	0.102	0.091	0.459	0.575	0.521
flatBM25 (C+C)	0.247	0.177	0.184	0.145	0.089	0.087	0.357	0.494	0.453
BM25F (C+C)	0.326	0.279	0.235	0.179	0.133	0.124	0.421	0.558	0.505

Table 5.7: Comparison of results for each subject for structured content (C) and content+context (C+C) retrieval using BM25 and flat content+context (C+C) based retrieval using flatBM25 and BM25F, on the 20 month collections using the full sets of 100 queries, for average precision (AveP), P@5 and P@10.

Technique	Subject 1			Subject 2			Subject 3		
	AveP	**P@5**	**P@10**	**AveP**	**P@5**	**P@10**	**AveP**	**P@5**	**P@10**
BM25 (C)	0.230	0.228	0.193	0.149	0.098	0.084	0.404	0.508	0.457
flatBM25	8%	-22%	-5%	-2%	-9%	3%	-12%	-3%	-1%
BM25F	42%	22%	22%	20%	36%	47%	4%	10%	10%
BM25 (C+C)	0.314	0.281	0.235	0.177	0.102	0.091	0.459	0.575	0.521
flatBM25	-21%	-37%	-22%	-18%	-13%	-5%	-22%	-14%	-13%
BM25F	4%	-1%	0%	1%	30%	37%	-8%	-3%	-3%

Table 5.8: Subjects' percentage improvement over BM25 content (C) and BM25 content+context (C+C) retrieval, rounded to nearest whole number, on their 20 month collections using full set of 100 queries, for average precision (AveP), P@5 and P@10 using flatBM25 and BM25F.

5.4.1 Results and Analysis

Tables 5.7 and 5.8 show a comparison of the results obtained by using flatBM25 and BM25F as opposed to the structured BM25 content and BM25 content+context approaches presented in Section 5.3 for the 3 subjects. These results are examined in this section.

5.4.1.1 Comparison of flatBM25 and BM25 Based Retrieval

Using the flatBM25 approach the only improvement over BM25 content only retrieval observed was an 8% improvement in AveP for Subject 1 and a 3% improvement in P@10 for Subject 2. The fact that flatBM25 content+context retrieval outperformed BM25 content only retrieval in these two instances might be explained by the make up of these subjects queries shown in Table 5.6. Subject 1, relative to the other two subjects, makes much greater use of the title field in querying. When we convert this subject's structured queries to flat queries, their title query terms can be queried across all fields in the collection, including importantly the content field. This may have factored in improving the average percision observed for this subject using flatBM25. Subject 2 had short content-based queries and did not make much use of the title field in retrieval, relative to the other two subjects. As we saw in the previous section, this subject therefore relied more on the other rich context types to improve on content only retrieval performance. These factors, may partly explain why we observed an improvement over content only retrieval using flatBM25 retrieval for this subject.

In all other cases substantial decreases in performance relative to using BM25 content only retrieval were observed. These decreases were: for Subject 1, 22% and 5% in P@5 and P@10 respectively; for Subject 2, 2%, 9% for AveP and P@5 respectively; and for Subject 3, 12%, 3% and 1% for AveP, P@5 and P@10 respectively. These results suggest that use of BM25 content only retrieval would be a better alternative than allowing subjects to query based on recalled context using the flatBM25 approach, certainly in the case of our subjects' collections and their queries. In the flatBM25 results we also note that performance against BM25 structured content+context retrieval was unsurprisingly even worse (given that BM25 structured content+context retrieval outperforms BM25 content only retrieval). Here decreases in performance were: for Subject 1, 21%, 37% and 22% for AveP, P@5 and P@10 respectively; for Subject 2, 18%, 13% and 5% for AveP, P@5 and P@10 respectively; and for Subject 3, 22%, 14% and 13% for AveP, P@5 and P@10 respectively. These results are unsurprising given that in using the flatBM25 approach we have pooled all item fields together and hence are losing the rich information we had in the structured documents, as discussed in Chapter 4.2.2.2.

5.4.1.2 Comparison of BM25F and BM25 Based Retrieval

BM25F content+context retrieval V BM25 content only retrieval:

In contrast to the use of flatBM25 in retrieval using the BM25F flat content+context querying approach provides dramatic improvement over content only retrieval for Subjects 1 and 2, with increases in AveP, P@5 and P@10 of 42%, 22% and 22% for Subject 1 and 20%, 36% and 47% for Subject 2. For Subject 3, for whom high default content only scores were obtained, the improvement was not as dramatic, but still substantial, with improvement of 4%, 10% and 10% being observed for AveP, P@5 and P@10 respectively. These improvements over content only retrieval by allowing subjects to also provide recalled context data associated with required items in a simple flat query are promising.

BM25F content+context retrieval V BM25 content+context retrieval:

Using BM25F the field level score for a query term is calculated based on the term frequency in the field and the length of the field only. Whereas our structured querying approach presented in Investigation 1 (Section 5.3) calculates the field score based on, in addition to the term frequency in the field and the length of the field, the idf of the field. That is, using BM25 in structured queries the idf of a query term is calculated at the field level. The idf score using BM25F is calculated at the document level. Further in calculating weights using BM25F each query term, regardless of the target field intended by the subject, is tested for relevance against each field in the document. Whereas using BM25 query terms are only tested for relevance against the target field intended for the terms by the subject. Based on the different content of subjects' collections and their query construction approaches, BM25F gave optimal performance for some subjects while the structured BM25 content+context approach gave superior performance for others (see Tables 5.7 and 5.8).

Subject 2 benefited the most from addition of further context types to content+title retrieval using the structured BM25 content+context retrieval approach, as we saw in the previous section. For this subject, great improvement over structured BM25 content+context retrieval was observed using BM25F, with increases of 1%, 30% and 37% in AveP, P@5 and P@10 respectively. For Subject 1, 1% improvement in AveP over BM25 content+context-based retrieval was observed using BM25F. However, no improvement for P@10 was yielded and for P@5 a decrease in performance of 1% was observed. For Subject 3, who benefited the least from addition of context types to BM25 content+title retrieval (as we saw in the previous section), BM25F gave inferior performance to structured BM25 content+context retrieval. Decreases in performance of 8%, 3% and 3% in AveP, P@5 and P@10 respectively were observed.

Across the three subjects, we note that as the BM25 content only retrieval baseline and in turn the BM25 content+context baseline increases the percentage improvement gained by using BM25F decreases. Leading from greatest utility being observed for Subject 2 with the lowest BM25 content+context baseline, to no utility for Subject 3 with the highest baseline.

We further note that for Subject 2, who made the least use of the title field in retrieval and who had the shortest content-based queries (see Table 5.6), BM25F yielded greatest improvement over structured BM25 content+context retrieval. Given that the title and content query terms share a common vocabulary, using BM25F their content field query terms are matched against both the content and title field. Similarly their title field query terms are matched against both fields. One would assume that matching the subjects query terms from these two fields against both the content and title field is helping negate the fact that they did not have as many query terms for these two fields as the other subjects.

The inferior performance observed for Subject 3 using the BM25F retrieval technique, may partially be explained by the nature of this subject's queries. The query terms used in the content field of their queries often contain terms which occur in context fields. For example, use of geo-location, year, month, etc, type terms as search terms for the content field of their queries. While this does not create a problem for structured content+context querying where each field and its associated terms are weighted separately, it has the potential to create a problem when we convert the subject's structured queries to flat queries. The result being that terms which relate to the content of the item but also occur as context types in other fields, are now potentially contributing greater weight to non-relevant items, at the expense of relevant items. To highlight this with an example, consider a query where a term in the content field is 'Paris', as the subject wishes to look up web pages regarding sight seeing in 'Paris' which they previously looked at from the geo-location 'Dublin' before making a trip to 'Paris'. In this example, high weight will be incorrectly given to items accessed in the geo-location 'Paris'. Using flat queries, these words which do not relate to memory of context associated with the required information, will be matched to these context fields, negatively impacting retrieval performance where these words do not match the context associated with the required items. This issue was not observed for Subjects 1 and 2 due to their choice of query terms for the content field.

These are just some of the factors which may have contributed to the differences observed in this section's results across our three subjects. A detailed in-depth analysis of the make up of subjects collections, queries and relevant items and how they impact on each other in retrieval is beyond the scope of this research.

5.4.2 Concluding Remarks

Despite the mixed results observed in this section using BM25F relative to using BM25 in content+context-based retrieval, BM25F did greatly out perform content only retrieval across the three subjects. Given that BM25F allows subjects to search across multiple fields using simple flat queries, which as discussed in Section 5.1 is desirable, we wish to explore the use of BM25F in lifelog retrieval further to establish whether we can improve its performance by considering some of the attributes of our subjects' collections and queries and modifying BM25F based on this analysis. We explore this topic in the next section.

5.5 Investigation 3: Flat BM25F_mod Content+Context-Based Retrieval - Moving Beyond the State of the Art

The retrieval challenge for lifelog collections is fundamentally different to that for which BM25F was originally created, in that in our multi-field lifelog collections fields from a querying point of view are independent of each other in that different fields do not have common meaning (with the exception of the title and content fields). That is, in entering flat queries for these collections, subjects have particular types of context in mind. For example, when a subject recalls that they were in 'Rome' when they previously interacted with a required lifelog item, they are specifically thinking of the geo-location context field per se and not looking for items which contain the term 'Rome' in the main body of text. Such mappings of query terms to their intended target fields is considered in [Kim et al., 2009, Kim and Croft, 2009] and discussed in Chapter 4.2.2.2.

When considering multi-field documents for retrieval using flat queries, the standard BM25F strategy extends the traditional BM25 model in enabling the importance of individual terms retrieval effectiveness to be captured in the field weights. Terms which have higher term frequency in high weighted fields then have overall greater influence on document scores. This however assumes that a single scalar weighting on each field is sufficient to capture the optimal contribution each term can make to retrieval. To accommodate lifelog retrieval and better map query terms to their intended field, we modify BM25F at the term scoring level. Our proposed modified

BM25F term scoring approach (BM25F_mod) for calculating the weight $\bar{w}_{t,d}$ of term *t* in document *d* is shown in Equation 5.1.

$$\bar{w}_{t,d} = \sum_{f\ in\ d} \frac{tf_{d,f,t} \cdot W_f}{((1 - b_f) + b_f \cdot \frac{l_f}{avl_f})} \cdot m_{d,f,t} \tag{5.1}$$

where,

tf$_{d,f,t}$ is the frequency of term *t* in field *f* of document *d*.

l$_f$ is the length of *f* in *d*.

avl$_f$ is the average length of field *f*.

W$_f$ is the weight assigned to *f*.

b_f is a length normalising parameter for *f*.

$m_{d,f,t}$ is an importance calculation for the score assigned to term *t* in field *f* of document *d*.

$m_{d,f,t}$ is our proposed modification to BM25F term scoring. Setting $m_{d,f,t}$ to 1, reduces the equation to BM25F's standard term scoring approach described in Chapter 4.2.2.2. In Sections 5.5.1 - 5.5.3 we motivate and investigate techniques for calculating $m_{d,f,t}$ to map query terms to their intended target field. As with standard BM25F, the term weight $\bar{w}_{t,d}$ is applied to the BM25 saturating function shown in Equation 5.2.

$$BM25F\_mod(q,d) = \sum_{t\ in\ d} \frac{\bar{w}_{t,d}}{k_1 + \bar{w}_{t,d}} \cdot log\frac{N}{df(t)} \tag{5.2}$$

where,

N is the number of documents in the collection.

df(t) the number of documents containing term *t*.

k_1 is a saturating parameter.

5.5.1 BM25F_mod1

To begin our investigation of modifying BM25F at the term scoring level to map query terms to their intended target field, we investigate the term scoring adaptation put forward by Kim et al [Kim et al., 2009, Kim and Croft, 2009] and validated in their language modelling approach to IR. This term scoring adaptation weights the term score for each field according to the frequency of the term in the given field relative to its frequency in the entire document, with the expectation that this maps query terms to their target field. Application of this approach to our BM25F term scoring approach presented in Equation 5.1 results in the m$_{d,f,t}$ function shown in Equation 5.3.

$$m^1_{d,f,t} = (1 - \alpha) + \alpha \frac{tf_{d,f,t}}{tf_{d,t}} \tag{5.3}$$

where,

tf$_{d,f,t}$ is the frequency of term *t* in field *f* of document *d*.

tf$_{d,t}$ is the frequency of *t* in *d*.

α is a tuneable parameter. Setting α = 1 reduces the equation to $\frac{tf_{d,f,t}}{tf_{d,t}}$, setting α to 0 results in no term field to document weighting. For our experiments α was manually tuned using the biometric month test set and associated queries (described in Chapters 4.3.1 and 4.3.2.3) to give overall best retrieval performance across the three subjects. This resulted in α being set to 0.2.

For the remainder of this Chapter we use *BM25F_mod1* to refer to the use of $m^1_{d,f,t}$ in Equation 5.2.

5.5.1.1 Results and Analysis

Table 5.9 shows the performance of BM25F_mod1 for the 3 subjects. For comparison purposes we also include the performance of BM25 content, BM25 content+context and BM25F. The percentage improvement over structured BM25 content and over BM25 content+context retrieval for BM25F_mod1 for the 3 subjects is presented in Tables 5.10 and 5.11. For comparison purposes we also include the percentage improvement for BM25F over structured BM25 retrieval in these tables. Table 5.12 provides the percentage improvement over BM25F observed using BM25F_mod1.

Technique	Subject 1			Subject 2			Subject 3		
	AveP	P@5	P@10	AveP	P@5	P@10	AveP	P@5	P@10
BM25 (C)	0.230	0.228	0.193	0.149	0.098	0.084	0.404	0.508	0.457
BM25 (C+C)	0.314	0.281	0.235	0.177	0.102	0.091	0.459	0.575	0.521
BM25F	0.326	0.279	0.235	0.179	0.133	0.124	0.421	0.558	0.505
BM2525F_mod1	0.322	0.274	0.234	0.189	0.129	0.118	0.419	0.558	0.506
BM25F_mod2	0.339	0.291	0.240	0.174	0.133	0.129	0.421	0.563	0.503
BM25F_mod3	0.334	0.281	0.240	0.177	0.129	0.124	0.420	0.567	0.505

Table 5.9: BM25F_mod1, BM25F_mod2 and BM25F_mod3 results, using subjects' 20 month collection and full set of 100 queries, for average precision (AveP), P@5 and P@10. BM25 content only (C), BM25 content+context (C+C) and BM25F results shown for comparison purposes.

Technique	Subject 1			Subject 2			Subject 3		
	AveP	P@5	P@10	AveP	P@5	P@10	AveP	P@5	P@10
BM25 (C)	0.230	0.228	0.193	0.149	0.098	0.084	0.404	0.508	0.457
BM25F	42%	22%	22%	20%	36%	47%	4%	10%	10%
BM25F_mod1	40%	20%	21%	27%	32%	40%	4%	10%	11%
BM25F_mod2	48%	28%	24%	17%	36%	53%	4%	11%	10%
BM25F_mod3	46%	23%	24%	19%	32%	47%	4%	11%	10%

Table 5.10: BM25F, BM25F_mod1, BM25F_mod2 and BM25F_mod3 percentage improvement over BM25 content only (C) retrieval, rounded to nearest whole number, on subjects' 20 month collection using full set of 100 queries, for average precision (AveP), P@5 and P@10. For comparison purposes the percentage improvement of BM25F over BM25 content only retrieval is also presented.

Similar to the results observed using BM25F in the previous section, BM25F_mod1 provides improvement over content only retrieval for all 3 subjects, as shown in Tables 5.9 and 5.10. However, BM25F_mod1 did not overall fare as well as standard BM25F in comparison to BM25 content+context retrieval, shown in Tables 5.9 and 5.11. Subsequently, little overall improvement was observed over BM25F using BM25F_mod1, as shown in Tables 5.9 and 5.12. Specifically, as can be inferred from Table 5.12, reductions of 1%, 2% and 1% were observed for BM25F_mod1 relative to BM25F in Subject 1's case for AveP, P@5 and P@10 respectively, 3% for P@5 and 5% for P@10 reduction was observed for Subject 2 and in the case of Subject 3 no improvement over BM25F was observed using BM25F_mod1. The only improvement over BM25F observed was a 6% improvement in AveP for Subject 2 using BM25F_mod1.

On analysis, the reason for this overall under performance becomes apparent. In cases where a query term occurs multiple times in the content field as well as in a context field (e.g. the term 'web' occurring in extension field and multiple times in the content field), the term score of the context field will be downgraded relative to standard

Technique	Subject 1			Subject 2			Subject 3		
	AveP	P@5	P@10	AveP	P@5	P@10	AveP	P@5	P@10
BM25 (C+C)	0.314	0.281	0.235	0.177	0.102	0.091	0.459	0.575	0.521
BM25F	4%	-1%	0%	1%	30%	37%	-8%	-3%	-3%
BM25F_mod1	2%	-2%	-1%	6%	26%	29%	-9%	-3%	-3%
BM25F_mod2	8%	3%	2%	-2%	30%	41%	-8%	-2%	-3%
BM25F_mod3	6%	0%	2%	0%	26%	37%	-8%	-1%	-3%

Table 5.11: BM25F_mod1, BM25F_mod2 and BM25F_mod3 percentage improvement over BM25 content+context (C+C) retrieval, rounded to nearest whole number, on subjects' 20 month collection using full set of 100 queries, for average precision (AveP), P@5 and P@10. For comparison purposes BM25F percentage improvement over BM25 content+context retrieval is also presented.

Technique	Subject 1			Subject 2			Subject 3		
	AveP	P@5	P@10	AveP	P@5	P@10	AveP	P@5	P@10
BM25F	0.326	0.279	0.235	0.179	0.133	0.124	0.421	0.558	0.505
BM25F_mod1	-1%	-2%	-1%	6%	-3%	-5%	0%	0%	0%
BM25F_mod2	4%	4%	2%	-3%	0%	4%	0%	1%	0%
BM25F_mod3	3%	1%	2%	-1%	-3%	0%	0%	2%	0%

Table 5.12: BM25F_mod1, BM25F_mod2 and BM25F_mod3 percentage improvement over BM25F retrieval, rounded to nearest whole number, on subjects' 20 month collection using full set of 100 queries, for average precision (AveP), P@5 and P@10.

BM25F. This occurs because BM25F_mod1 multiplies the field's term score by the frequency of a term in a field divided by the frequency of the term in the document as a whole. To illustrate this, consider the first sample term scoring scenario presented in Figure 5.1, in which a subject used the query term 'February' because they recalled accessing the required item in the month of 'February'. Applying this query term to the sample lifelog item presented in Figure 5.1, the term score for the content field obtains a boost of 4/5 while the month fields term score is reduced. Similarly in the second sample scenario presented in Figure 5.1, the term score of the title field for the query term 'retrieval' gets reduced due to the more frequent occurrence of the query term in the content field. Further, a rich context query term which does not occur in the target context field of an item, but occurs one or several times in the content of a non-relevant item, will give the same incorrect term scoring boost to this non-relevant item as the standard BM25F term scoring approach. To highlight this with an example, consider a non-relevant item which contains 3 occurrences of a queried context term (e.g. 'Summer') in the content field of the item and 0 occurrences of this query term in the target context field. In this example, the $m^1_{d,f,t}$ score for this context query term in the content field will be 1, hence reducing the term scoring approach to standard

BM25F term scoring.

With the exception of the improvement in AveP observed for Subject 2, this approach is clearly not appropriately mapping queried context terms to their intended target field in our lifelog collections. We suppose that this problem would not manifest itself in the desktop and movie database collections considered for retrieval in [Kim et al., 2009, Kim and Croft, 2009], discussed in Chapter 4.2.2.2, where the query independent context fields were unlikely to contain terms in common with the content field. However, in our collections where terms have separate meaning but share the same vocabulary, instead of boosting the occurrence of rich context query terms in their intended fields, it potentially has the opposite effect. This problem could be overcome by using a different vocabulary for rich context terms, e.g., using the suffix 'ext' with extension types (e.g., web_ext), using the extension 'season' for season field items (e.g., summer_season). However, this would essentially reduce to structured querying, place extra burden on individuals, and does not fit with the spirit of allowing individuals to enter simple flat open vocabulary queries.

5.5.2 BM25F_mod2

Given that a problem with term scoring is the mapping of rich context sources to appropriate fields (as shown in the preceding BM25F_mod1 results analysis) and that these rich context sources have short fields, we propose that boosting the occurrence of query terms in short fields should prove an effective way to map queried context terms to the fields intended by the individual performing the query. To achieve this we propose considering the length of fields (the field length is obtained by counting the number of terms in the field) relative to the length of the documents (the document length is obtained by counting the number of terms in all fields of the document) in our field term score boosting approach. This led to the modified approach to calculating m$_{d,f,t}$ in Equation 5.1, shown in Equation 5.4.

$$m^2_{d,f,t} = (1-\beta) + \beta \frac{1}{avl_f/avl_d} \tag{5.4}$$

where,

avl_f is the average field length.

EXAMPLE 1 - query term 'February'

Content field contents:
...February...February...February...

Month field contents:
February

Term score boosts:
Content field term score boost = 3/4
Month field term score boost = 1/4

EXAMPLE 2 - query term 'retrieval'

Content field contents:
...retrieval...retrieval...retrieval...retrieval...

Title field contents:
...retrieval...

Term score boosts:
Content field term score boost = 4/5
Title field term score boost = 1/5

Figure 5.1: Sample term score boost scenarios using BM25F_mod1's $m^1_{d,f,t}$ term score boosting approach (i.e. $(1-\alpha)+\alpha\frac{tf_{d,f,t}}{tf_{d,t}}$) with α set to 1.

avl_d is the average document length.

β is a tuneable parameter. Setting β = 1 reduces the equation to $\frac{1}{avl_f/avl_d}$, setting β to 0 results in no field length boost being applied. For our experiments β was manually tuned using the biometric month test set and associated queries (described in Chapters 4.3.1 and 4.3.2.3) to give overall best retrieval performance across the three subjects. This resulted in β being set to 0.0004[3].

For the remainder of this Chapter we use *BM25F_mod2* to refer to the use of $m^2_{d,f,t}$ in Equation 5.2.

[3] Although this value may appear small the fraction is often very large.

5.5.2.1 Results and Analysis

Similar to BM25F_mod1, using BM25F_mod2 provides improvement over content only retrieval, as shown in Tables 5.9 and 5.10. Overall BM25F_mod2 also performs better than BM25F_mod1, as shown in Tables 5.9 - 5.12. Further, using BM25F_mod2 improvement in performance over both BM25 content+context retrieval and BM25F retrieval was observed for Subjects 1 and 2, as shown in Tables 5.11 and 5.12. For Subject 1, 8%, 3% and 2% improvement in AveP, P@5 and P@10 respectively was observed using BM25F_mod2 as opposed to BM25 content+context retrieval and 30% and 41% improvement in P@5 and P@10 observed for Subject 2 (2% reduction in AveP was noted for this subject). Using BM25F_mod2 as opposed to BM25F, 4%, 4% and 2% improvement in AveP, P@5 and P@10 respectively was observed for Subject 1 and 4% improvement in P@10 was observed for Subject 2 (no improvement in P@5, and 3% disimprovement in AveP were noted for this subject).

For Subject 3, no improvement over BM25 content+context was obtained using BM25F_mod2, with 8%, 2% and 3% decreases in AveP, P@5 and P@10 yielded here (see Table 5.11). However, for this subject, 1% improvement in performance over BM25F was observed using BM25F_mod2 for P@5 (see Table 5.12), showing this technique to be slightly better at moving relevant items to the top of result lists. Although no improvement in AveP or P@5 was gained here. The minimal improvement in performance observed for Subject 3 may be explained by the fact that this subject used content query terms which shared the same vocabulary with context field terms, e.g. using location and month as query terms to the content field which did not correspond to the location and month in which the required item was accessed but rather related to the content of the required item(s). This issue was not observed for Subjects 1 and 2 who did not use content query terms sharing the same vocabulary as context fields.

Overall, the results obtained for BM25F_mod2 show some utility for a field length normalisation term score in mapping queried context types to the intended target field.

5.5.3 BM25F_mod3

While field length normalisation term score boost (i.e. BM25F_mod2) gave overall superior performance to BM25F (as shown in Section 5.5.2), there is the potential that too much weight will be given to terms occurring in shorter context fields at the ex-

Technique	Subject 1			Subject 2			Subject 3		
	AveP	**P@5**	**P@10**	**AveP**	**P@5**	**P@10**	**AveP**	**P@5**	**P@10**
BM25F_mod2	0.339	0.291	0.240	0.174	0.133	0.129	0.421	0.563	0.503
BM25F_mod3	0.334 (-1%)	0.281 (-3%)	0.240 (0%)	0.177 (2%)	0.129 (-3%)	0.124 (-3%)	0.420 (0%)	0.567 (1%)	0.505 (0%)

Table 5.13: BM25F_mod2 and BM25F_mod3 results comparison, using subjects' 20 month collection and full set of 100 queries, for average precision (AveP), P@5 and P@10. Percentage improvement of BM25F_mod3 over BM25F_mod2 shown in brackets.

pense of the content field. To highlight this with an example, if the average field length across all documents fields for a subject is: 494; the average length of this subjects content field is: 476; and the average length of their extension field is: 1.0. The extension fields term score will be divided by a boost of: 1/494 (weighted according to the tuned parameter in BM25F_mod2), which is a huge boost relative to the content field's boost of division by 476/494 (weighted according to the tuned parameter in BM25F_mod2). We hypothesise that using length normalisation in combination with term frequency normalisation will prove most beneficial for lifelog retrieval. This leads to *BM25F_mod3* which combines the approaches adopted in BM25F_mod1 and BM25F_mod2. BM25F_mod3 calculates m$_{d,f,t}$ in Equation 5.1 as shown in Equation 5.5.

$$m^3_{d,f,t} = ((1-\alpha) + \alpha\frac{tf_{d,f,t}}{tf_{d,t}}) \cdot ((1-\beta) + \beta\frac{avl_d}{avl_f}) \qquad (5.5)$$

where,

tf$_{d,f,t}$ is the frequency of term *t* in field *f* of document *d*.

tf$_{d,t}$ is the frequency of term *t* in document *d*.

avl$_f$ is the average field length.

avl$_d$ is the average document length.

α and β are tuneable parameters set to 0.2 and 0.0004 respectively as described in the preceding subsections.

5.5.3.1 Results and Analysis

Our hypothesis that combining field length normalisation with term frequency normalisation to boost field term score only proved true for Subject 3. Where BM25F_mod3 was a little more successful in moving relevant items to the top of the result list than the other flat querying techniques (i.e., BM25F, BM25F_mod1 and BM25F_mod2), as shown in Tables 5.9 - 5.13. For this subject BM25F_mod3 yielded 4%, 11% and 10% improvement over BM25 content only retrieval for AveP, P@5 and P@10 respectively (see Table 5.10). Similar to the other flat querying approaches it gave inferior performance to BM25 content+context retrieval (see Table 5.11). It yielded 2% improvement in P@5 over BM25F (see Table 5.12). It also yielded 1% improvement in P@5 over BM25F_mod2 (see Table 5.13).

Of the flat retrieval techniques tested on Subjects 1 and 2's collections up to this point, BM25F_mod2 yielded greatest performance. For these subjects, BM25F_mod3, while performing better than BM25F_mod1, did not outperform BM25F_mod2 (see Tables 5.9 - 5.12) in retrieving relevant items at top rank. That is, for these two subjects the term frequency normalisation boost is degrading the utility of the field length normalisation boost in retrieving relevant items at top rank. Table 5.13 shows the percentage improvement for BM25F_mod3 over BM25F_mod2. For Subject 1, BM25F_mod3 yielded 3% and 0% reduction in P@5 and P@10 respectively over BM25F_mod2 retrieval, and for Subject 2, BM25F_mod3 yielded 3% reduction in both P@5 and P@10 over BM25F_mod2 retrieval. BM25F_mod3 also yielded a 1% reduction in AveP over BM25F_mod2 for Subject 1. However, in the case of Subject 2 a 2% improvement in AveP was observed. Tables Tables 5.10 - 5.12 show that for Subject 1, BM25F_mod3 yielded 46%, 23% and 24% improvement in AveP, P@5 and P@10 respectively over BM25 content only retrieval (see Table 5.10), 6%, 0% and 2% improvement in AveP, P@5 and P@10 respectively over BM25 content+context retrieval (see Table 5.11), and 3%, 1% and 2% improvement in AveP, P@5 and P@10 respectively over BM25F retrieval (see Table 5.12). For Subject 2, BM25F_mod3 yielded 19%, 32% and 47% improvement in AveP, P@5 and P@10 respectively over BM25 content only retrieval (see Table 5.10), 0%, 26% and 37% improvement in AveP, P@5 and P@10 respectively over BM25 content+context retrieval (see Table 5.11), and 1%, 3% and 0% reduction in AveP, P@5 and P@10 respectively over BM25F retrieval (see Table 5.12).

5.5.4 Concluding Remarks

Overall we observed BM25F_mod2 to yield the greatest improvement in performance over BM25F. While not outperforming BM25 content+context retrieval across all subjects, BM25F_mod2 benefits from the fact that it allows individuals to enter simple flat queries using their recalled content and context associated with required lifelog items, resulting in greater retrieval performance than entering content only based queries (as we observed from the performance of BM25F_mod2 relative to BM25 content only retrieval).

While we state that overall BM25F_mod2 was the best performing flat querying approach examined, we note that in the case of Subject 3 better performance was obtained using BM25F_mod3. And that in the case of Subject 2, while BM25F_mod2 showed greatest utility in moving relevant items to the top of result lists, it under performed BM25F for AveP. Indeed it could be argued that Subject 1 was the only of our three subjects for which real performance gains were achieved using BM25F_mod2. What causes the differences in performance across our three subjects is unclear, as is how this technique would fare over other personal collections. That being said, based on the results observed in this section we draw the conclusion that of the tested flat querying approaches, BM25F_mod2 is the best flat querying approach on our three subjects' collection and queries. However, it should be noted that the improvements in retrieval performance using BM25F_mod1, BM25F_mod2 and BM25F_mod3 relative to BM25F on the subjects collections are not statistically significant for AveP, P@5 or P@10 (100 samples, Welch two sample t-test, $p>0.05$).

5.6 Conclusions

The use of many rich sources of context data in retrieval is a complex process and the factors influencing its utility are complex. Individuals' queries can vary in length, context types used and query term to target field mapping complexity. Target result sets for queries can vary in size and diversity of item type. Individuals underlying lifelogs show significant diversity in the amount of interaction with item types and in the volume of items in their collections. These complexities are exacerbated by missing context data and the possibility of incorrect recall on the part of the individual (analysis of whether the subjects in our studies had such incorrect recall is beyond the scope of this study). Additionally, since we are using multi-item queries (i.e. there may be several relevant target items for a subjects query), all query terms will not map, nor are they intended to map, to each of the target items of the query. Rather in querying of this nature, individuals are providing all recalled content and context associated with the set of required items in one query.

Significant manual analysis was conducted examining the make up of fields in subjects' queries and the relationship of this to individual queries retrieval performance using the techniques investigated in this chapter, however no clear trends were apparent. In this chapter we reviewed some of the issues which can potentially affect retrieval performance stemming from the make up of subjects' collections and the types of queries they enter. Full analysis of the make up of individuals' collections and queries is beyond the scope of one thesis. This chapter served to highlight some of the possible issues, and to show that despite them, there appears to be utility in using queried context data in retrieval. Importantly for the goal of this chapter, utility is demonstrated in allowing individuals to enter this recalled context with recalled content in simple flat queries. This is only the beginning of explorations in this space, and much further analysis using larger groups of subjects is required. This further work is required to validate our findings on larger groups of individuals, explore the issues raised in this chapter further, gain further understanding of the make up of individuals collections, the make up of their queries, the make up of these two things combined and the implications they have on each other in retrieval.

Despite the complexities associated with retrieval in this domain, in this chapter we have proposed retrieval algorithms which allow subjects to enter query terms associated with their recalled content and context of required items in a simple flat query. These algorithms greatly outperform content only retrieval for the lifelogging domain. Of the algorithms we proposed, overall BM25F_mod2 showed the greatest utility in PL retrieval.

Having explored the utility of recalled context in lifelog retrieval, in the next chapter we begin our exploration of the utility of an implicit context source, namely biometric response, in lifelog retrieval. That is, in the next chapter we move towards further improvement of the BM25F_mod2 retrieval approach through the exploration of obtaining implicit indicators of lifelog item importance. The subsequent chapter, Chapter 7, integrates these implicit importance indicators, based on biometric response, into the BM25F_mod2 retrieval algorithm presented in this chapter for further exploration of PL search.

CHAPTER

SIX

Extracting Important Items using Past Biometric Response

Chapter Overview: We postulate that (1) biometric response at the times of experiencing lifelog items/events will act as an indicator of the future importance of items/events, and (2) that biometric response can therefore provide utility as a static query independent significance factor in ranked content+context retrieval algorithms in the lifelogging domain. In this chapter we investigate the first part of this hypothesis. We first describe the setup of our experiments to investigate the relationship between biometric measures of Galvanic Skin Response, Heat Flux, Skin Temperature and Heart Rate associated with past experience of computer items and events recorded by the SenseCam and the current importance of these items/events to the individual. We then provide a detailed analysis of the results obtained from these investigations before concluding the chapter with the motivation to the second part of our hypothesis which is then investigated in the next chapter (Chapter 7).

6.1 Introduction

Finding important relevant items from within PLs in response to user queries, or presenting interesting data to a subject browsing through their archive, poses significant challenges. Any additional information which can assist in identifying important items is thus potentially very important. Such information could potentially be used in the re-ranking of IR result sets, or for the promotion of interesting items when browsing a lifelog collection. One potential source of useful information is the user's biometric response associated with an item. As highlighted in Chapter 2.3.2 past research has shown that biometric response can be used as an indicator of a person's current engagement with digital data. However research has not looked at the possibility of using this biometric response as a future indicator of item importance. We propose that items or events which are important to an individual at the time they occurred, may be useful to the individual again in the future, and further that such incidents are associated with responses that can be detected by measuring the individual's biometric response when experiencing these events. Thus recording biometric response as part of a lifelog may enable us to identify important items or events in the lifelog which would be most interesting for an individual to browse through in the future or which would be most important in a given information searching task. In our study, we explore four biometric responses which can be associated with lifelog items, namely galvanic skin response (GSR), heat flux (HF), heart rate (HR) and skin temperature (ST).

In this chapter we present an exploration of this hypothesis. We describe a study designed to assess the potential for use of biometric data in detecting useful lifelog items/events for future browsing or retrieval. More specifically we explore the use of biometric response at the time of SenseCam image capture to identify events which the subject may wish to view in the future, and the use of biometric response associated with previous computer item accesses to extract computer items from lifelogs which may be of future significance. The next section describes the study setup. Results and analysis are provided in Section 6.3. We then discuss some key points on the topic of this chapter before concluding the chapter with a discussion of the findings and highlight how these findings support investigating using observed biometric responses associated with past experience of lifelog items as query independent (static) scoring factors to items in computation of best-match ranked retrieval lists. An experimental investigation of the use of biometric score factors in this way for PL search is described in Chapter 7.

6.2 Experiment Setup

In this section we describe the setup of a study to examine whether biometric data can be useful in identifying important and memorable events (an event here is defined as a temporal group of SenseCam images) and computer activity (i.e., emails sent or received, web pages viewed and computer files interacted with) which the subject may wish to view again in the future.

The one month period from our three subjects' lifelog databases which contain SenseCam images and computer items annotated with biometric data were used as the test set for this experiment (subjects lifelog databases are described in Chapter 3.2[1]). For the experiments presented in this chapter, the biometric data is read from the biometrics table of these databases, and both SenseCam and computer item details are read from the items and item_access tables. These tables are also described in Chapter 3.2. How biometric data, SenseCam images and computer activity were logged to populate these tables is described in Chapter 3.3. Finally, the content of these tables is described in Chapter 3.4.2.

The remainder of this section explains our use of biometric data to extract important SenseCam events[2] and computer items, and details the user study conducted to determine the utility of our approach.

6.2.1 Extracting Important Events

Biometric response changes all the time and many factors cause these changes, e.g., internal thought processes of the individual, external temperature, a fright caused by sudden noise, eating (see Chapter 2.3.1). Using biometric response associated with past experience of lifelog items in suggesting items to individuals which they might be interested in viewing (during a random browsing session or in a specific informa-

[1] Note: The imported emails in these databases, which do not have "date-time" of access information available, were excluded from this experiment (imported emails are described in Chapter 3.3.1.1).

[2] We use the term 'SenseCam events' to refer to life events captured by the SenseCam.

tion type targeted session) is therefore a very complex process. Physiological research focuses on attempting to understand changes in physiological response caused by different factors, or conversely the changes in physiological response caused by different emotional and other situations (see Chapter 2.3.1). Researchers have however not looked at patterns of physiological response variation associated with items in individuals' personal digital collections or indeed patterns which will indicate the future importance of items to individuals. We cannot attempt in one thesis to extract all such patterns (if such patterns do indeed exist), rather in this chapter, as an initial exploration into this new space of research, we attempt to establish if there is indeed a relationship between biometric response at the time of experiencing lifelog items and an individual's desire to view such items/events in the future.

It is known that with increased arousal levels, GSR, HR and HF levels increase and ST levels decrease (described in Chapter 2.3.1). As an initial investigation into the role of biometric response associated with past experience of lifelog items and an individual's future desire to view these lifelog items, we look at the periods where maximum GSR, HR and HF biometric response and minimum ST response were observed in our subjects' lifelog collections, and investigate if the lifelog items/events experienced at these times are more important to the individual than other periods in their collections. The hypothesis then which we wish to explore in this chapter is:

> *Important events/items from a lifelog archive are coincident with maximum observed GSR, HF and HR readings and with minimum observed ST readings at the original time of event occurrence, and that these events would be most interesting for subjects during future archive browsing.*

Variations in biometric response occur all the time and as stated previously can be caused by many things, for example changes in arousal level or changes in physical activity such as walking down a corridor or running. A problem in analysis of biometric data for the purposes of this experiment is to identify variations in biometric data which are likely to be the result of variations in arousal levels, as opposed to external physical reasons, e.g., walking, eating (as described in Chapter 2.3.1). An additional source of data that can be inferred from captured biometric data using the BodyMedia armband is the energy expenditure (also described in Chapter 2.3.1) of the individual. Energy expenditure correlates well with periods of physical activity in subjects' collections and given the nature of the energy expenditure calculation (described in Chapter 2.3.1) should account for other external physical reasons. Thus we put forward that:

> *Measured energy expenditure can be used in our lifelog collections to differentiate between high GSR, HF and HR biometric data levels and low ST biometric data levels, resulting from physical reasons and those arising from events experienced from the environment.*

We explored two techniques to remove the effect of physical activity on biometric levels. The first removes periods of high energy expenditure (referred to as *removeEng*) and the second divides the GSR, HR, HF and ST readings by the associated energy expenditure level (referred to as *divEng*). These techniques are described in Sections 6.2.1.1 and 6.2.1.2.

To determine relationship between item/event importance and GSR, HR, HF or ST, we extracted periods of high GSR, HR, HF and low ST from each subject's collection (we refer to these periods as *max GSR/HR/HF/ST*) using the *removeEng* and *divEng* techniques. For comparison purposes periods of both average GSR, HR, HF, ST, (referred to as *average GSR/HR/HF/ST*), and low GSR, HR, HF and high ST (referred to as *min GSR/HR/HF/ST*) were also extracted.

In total 5 computer items and 5 SenseCam events were extracted, using the *removeEng* and *divEng* techniques, for each of max, min and average GSR, HR, HF and ST (techniques used to extract items/events are described in Sections 6.2.1.1 and 6.2.1.2). Our choice of 5 items/events for each GSR, HR, HF and ST level was a reasonable number for values not to be selected by chance. This number of items/events also did not place too high a burden on the subjects performing the experiment described later in this section. Further, as we shall explain in the next section, given the size of our SenseCam collections and our adopted SenseCam event extraction process it was not possible to extract large numbers of SenseCam events. Indeed, as we shall also explain in the next section, it was not possible to even extract 5 SenseCam events in all cases.

The procedure for extraction of the max, min and average SenseCam and computer items/events using the *removeEng* and *divEng* techniques is described next.

6.2.1.1 Extracting SenseCam Events

Before extracting SenseCam events associated with max, min and average biometric response, each subject's biometric data was processed to remove the effect of physical reasons on the recorded biometric levels. Our first technique (*removeEng*) to remove the effect of physical reasons on biometric levels deleted biometric data captured during periods of energy expenditure above the average energy level × α (α = empirically determined scalar constant, set to 1.5 in this experiment) from the data set. For example, given our set of recorded GSR levels with "date-time" stamps, GSR readings which were recorded at the time of high energy expenditure were removed from the GSR readings set, as shown in the algorithm in Figure 6.1. Our second technique (*divEng*) divided GSR, HR and HF readings by energy expenditure and multiplied ST readings by energy expenditure. To highlight the use of the *divEng* technique with an example, the algorithm in Figure 6.2 shows the use of the *divEng* technique to transform raw GSR readings to account for energy expenditure levels.

Using each of the resulting biometric datasets in turn we extracted the time frames which corresponded with periods of max GSR, HR, HF and ST response. We then located the SenseCam images which were captured at these times of max GSR, HR, HF and ST response as candidate important events within a subjects collection. For comparison purposes we also located the periods of average and min GSR, HR, HF and ST response within subjects collections and extracted the SenseCam images captured during these periods. These sets of SenseCam images were then presented to subjects for rating to establish if there was a relationship between biometric levels and event importance (as will be described in the next section). We next describe in greater detail this process by which SenseCam images which occurred during periods of max, min and average GSR, HR, HF and ST response were extracted for this experiment. This process was as follows:

1. *Determining begin and end timestamps of max GSR, HF and HR:* Begin and end timestamps for periods in a subject's GSR/HF/HR dataset where the GSR/HF/HR level was greater than a preset threshold for an empirically determined number of seconds were recorded. (Threshold = average of GSR/HF/HR data * β, β = empirically determined scalar constant). In this experiment we sought to locate 5 SenseCam events for each of max GSR, HR and HF, hence the

```
Function: Get_GSR_removeEng_values
INPUT: GSRarrayIN = ArrayList≺double biometricLevel, String datetime≻
INPUT: EnergyArray = ArrayList≺double energyLevel, String datetime≻

INPUT: $\alpha$ = double
double EnergyThreshold = (average of all energy level values)*$\alpha$

for(int i = 0; i ≺ EnergyArray.length; i++)
[
 if(¬ EnergyArray[i] ≻ EnergyThreshold)
GSRarrayOUT = GSRarrayIN[i]
]

OUTPUT: GSRarrayOUT
```

Figure 6.1: Function to remove GSR data associated with periods of high energy expenditure (i.e., the *removeEng* technique).

```
Function: Get_GSR_divEng_values
INPUT: GSRarrayIN = ArrayList≺double biometricLevel, String datetime≻
INPUT: EnergyArray = ArrayList≺double energyLevel, String datetime≻

for(int i = 0; i ≺ EnergyArray.length; i++)
[
GSRarrayOUT = GSRarrayIN[i]/EnergyArray[i]
]

OUTPUT: GSRarrayOUT
```

Figure 6.2: Function to divide GSR data by energy expenditure level (i.e., the *divEng* technique).

number of seconds and scalar constant (β) were manually tuned for each subject to obtain, where possible, 5 periods of max GSR, HR and HF response corresponding to periods of SenseCam image capture. The empirically determined number of seconds and scalar constant were similarly set for each of the other techniques described in the remainder of this section.

Determine begin and end timestamps of max ST: Timestamps were obtained by taking periods where ST levels were less than a preset threshold for an empirically determined number of seconds. (threshold = average of ST data / β, β = empirically determined scalar constant)

Determining begin and end timestamps of min GSR, HF and HR: Timestamps were obtained by taking periods where GSR/HF/HR levels were less than a preset threshold for an empirically determined number of seconds. (threshold = average of GSR/HF/HR data / χ, χ = empirically determined scalar constant)

Determining begin and end timestamps of min ST: Begin and end timestamps for periods in a subject's ST dataset where the ST level was greater than a preset threshold for an empirically determined number of seconds were recorded. (threshold = average of ST data * χ, χ = empirically determined scalar constant)

Determining begin and end timestamps of average GSR, HF, HR and ST: Timestamps were obtained by taking periods where GSR/HR/HF/ST levels were $>$ threshold1 and $<$ threshold2 for an empirically determined number of seconds. (threshold1 = average of GSR/HR/ST data - δ, δ = empirically determined scalar constant; threshold2 = average of energy expenditure data + σ, where σ = δ)

Figure 6.3 shows an example of SenseCam event extraction, for max GSR. Part (1) in this figure shows the 'determining begin and end timestamps' of max GSR process described above.

2. *Extracting events from the subject's lifelog:* The begin and end timestamps from Step 1 were used to extract SenseCam events as follows (and shown in part (2) of Figure 6.3):

   - A window of 20 seconds was taken before the begin timestamp and after the end timestamp. This window was chosen as SenseCam images were captured every 20 seconds.

   - *If* SenseCam images occurred between the begin and end timestamps with 20 second window and there was no computer activity during this period[3], these images were chosen for presentation to subject for the user study which is described in Section 6.2.2.

3. *Removing duplicates:* Obviously, the use of a 20 second window for begin and end timestamps can result in 2 separate SenseCam events in the same category (e.g.

[3] Computer activity detected by Slife item accesses occurring during the time period. However, some SenseCam images showing the subject working on their computer still remained due to Slife crashes and manual stopping of the Slife software. In future studies image recognition techniques could be used to remove remaining SenseCam events showing computer activity. Our reasons for not including SenseCam images showing computer activity in this experiment stems from an earlier pilot study in which SenseCam images depicting computer activity were not considered interesting by subjects.

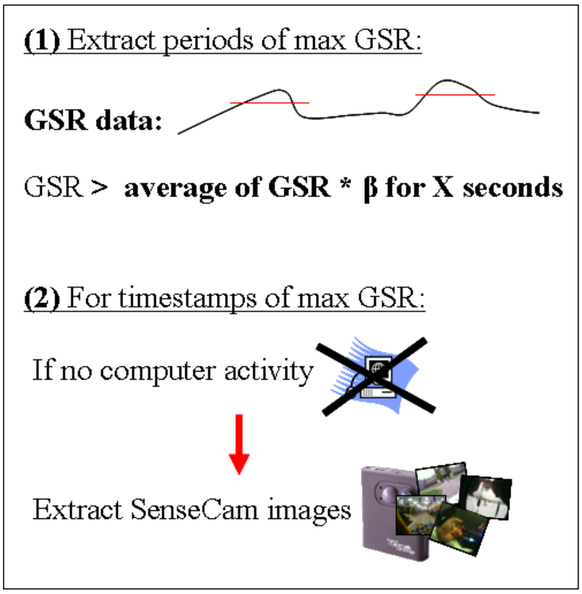

(1) Extract periods of max GSR:
GSR data:
GSR > average of GSR * β for X seconds
(2) For timestamps of max GSR:
If no computer activity
Extract SenseCam images

Figure 6.3: Extracting SenseCam events associated with periods of max GSR.

max GSR) having image(s) in common. In cases where this occurred we merged the images from the 2 events to create one event (with duplicates removed).

This 20 second window also allowed for a given biometric type (e.g. HR) images to occur across different thresholds (e.g. common image(s) in max HR and average HR events). In this instance the event generated from the higher occurring threshold was chosen as the event to present to the subject in the user study described in Section 6.2.2, and the event generated from the lower occurring threshold ignored (e.g. in the situation where there are common image(s) in a max HR event and an average HR event, only the max HR event would be selected for presentation to the subject for our experiment).

It should be noted that for max GSR/HR/HF and min ST events the ≤ 5 events containing the highest biometric reading for each biometric type were chosen. Similarly for min GSR/HR/HF and max ST the ≤ 5 events containing the lowest biometric readings for each biometric type were chosen, and events containing readings closest to the average were chosen for average GSR/HF/HR/ST. On completion of the above process, having expanded thresholds as far as possible, we had sets of ≤ 60 SenseCam events using the first technique for accounting for physical activity and ≤ 60 SenseCam events using the second technique for accounting for physical activity from each subject's lifelog[4]. The breakdown of the number of SenseCam events presented to

[4] We did not end up with 5 events for each max, min and average category as it was not possible to

	Sub1	**Sub2**	**Sub3**	**Sub1**	**Sub2**	**Sub3**	***Total***
Technique	**removeEng**			**divEng**			
GSR max	5	5	5	5	4	5	*29*
GSR ave	4	2	3	3	5	5	*22*
GSR min	3	5	5	3	2	5	*23*
HF max	3	3	5	5	5	5	*26*
HF ave	5	2	5	4	2	5	*23*
HF min	4	4	4	2	2	5	*21*
HR max	5	0	0	4	4	5	*18*
HR ave	5	0	0	3	2	5	*15*
HR min	5	0	0	2	2	5	*14*
ST max	2	5	5	2	3	5	*22*
ST ave	3	3	5	3	4	5	*23*
ST min	5	1	5	3	4	5	*23*
Total	*49*	*30*	*42*	*39*	*39*	*60*	***259***

Table 6.1: Number of SenseCam events retrieved per subject (Sub).

subjects in each category is shown in Table 6.1.

6.2.1.2 Extracting Computer Items

As for the extraction process used for SenseCam events (described at the beginning of the previous section), before extracting computer items associated with max, min and average biometric response, the subjects' biometric datasets were processed using the *removeEng* and *divEng* techniques to remove the effect of physical activity on the recorded biometric levels.

Similar to the SenseCam event extraction process described in the previous section, in extracting computer items to present to subjects in our user study (described in the next section), we wished to obtain the computer items with max GSR, HR, HF and ST levels. To do this, using each of the resulting biometric datasets in turn, we assigned computer items accessed with the highest GSR, HR, HF and 1/ST[5] values observed across all seconds of accesses to the items. We then chose the computer items with highest associated GSR, HR, HF and ST levels for presentation to subjects for rating in the user study described in the next section. For comparison purposes we also chose the computer items with average and min GSR, HR, HF and ST levels for presentation to subjects for our user study. The remainder of this section describes in greater detail this process by which computer items which occurred at points of max,

expand the thresholds any further without periods of max biometric response moving into the average biometric response range, for example.

[5] Note since ST values are inversely related to arousal level we take the inverse of ST levels to indicate arousal level. Using inverse ST values (i.e., i/ST) increases in ST indicate increases in arousal levels.

min and average GSR, HR, HF and ST response were located and extracted for this experiment. This process proceeds as follows:

1. Each computer item was assigned its highest associated GSR, HR, HF and ST readings (note here that in the case of the *divEng* technique we are referring to the biometric readings after energy has been accounted for, e.g., 1/(ST × energy), GSR/energy) across all seconds of access to the item.

2. *Extracting max GSR, HR, HF and ST items:* The 5 items with highest GSR, the 5 items with highest HR, the 5 items with highest HF and the 5 items with highest ST readings were selected for presentation to the subject for the user study described in the next section.

   Extracting average GSR, HR, HF and ST items: The 5 items closest to the subject's average GSR, the 5 items closest to the subject's average HR, the 5 items closest to the subject's average HF and the 5 items closest to the subject's average ST reading were selected for presentation to the subject for the user study described in the next section.

   Extracting min GSR, HR, HF and ST items: The 5 items with lowest GSR, the 5 items with lowest HR, the 5 items with lowest HF and the 5 items with lowest ST readings were selected for presentation to the subject for the user study described in the next section.

   As described in Chapter 3.3.1.3, it was not possible to capture the content of all computer items accessed by subjects on their PCs and laptops. In this experiment in extraction of lifelog items, items for which content data had not been captured were not considered for presentation to subjects in the above process.

6.2.2 Experiment Procedure

The goal of this research is to establish if periods associated with max GSR, HR, HF and ST are good indicators of lifelog items/events which are most useful for presentation to subjects when browsing their personal information archives. Personal lifelog items of varying GSR, HR, HF and ST were presented to subjects and a questionnaire completed to determine if GSR, HR, HF and ST corresponded with memorable-ness, significance of events and desire to view events again. Post questionnaire interviews were then conducted. This section describes the details of these procedures.

We wished to establish the relationship between biometric response at time of item/event creation/access on subjects' desire to re-view lifelog items/events over the long-term. We thus waited for 22 months after the test set collection period to present subjects with a set of events taken from their lifelogs[6]. A total of $\leq$ 240 items/events generated using the techniques described in Section 6.2.1 were presented to subjects in this set. The set included (as shown in Table 6.2): for each of GSR, HR, HF and ST 5 computer items with the max GSR/HR/HF/ST and $\leq$ 5 SenseCam image events corresponding to the max GSR/HR/HF/ST for each of the two techniques for accounting for energy expenditure (described in Section 6.2.1); and for comparison purposes similar sets of items/events with average GSR, HR, HF, ST and min GSR, HR, HF, ST. For each of average GSR, HR, HF and ST the 5 computer items and $\leq$ 5 SenseCam events closest to the subjects' average GSR, HR, HF and ST were chosen for each of the two techniques for accounting for energy expenditure (as described in Section 6.2.1). For each of min GSR, HR, HF and ST the 5 computer items and $\leq$ 5 SenseCam events closest to the subject's lowest min GSR/HR/HF/ST were chosen for each of the two techniques for accounting for energy expenditure (also described in Section 6.2.1).

Each subject was presented with their set of 120 computer items and $<$ 120 SenseCam events, and a questionnaire. Subjects were aware that the sets presented to them contained events with varying associated biometric levels and of the specific hypothesis we wished to test. However, they were not aware of the biometric response associated with each event. The questionnaire was explained to subjects and sample answers provided. Each subject completed the questionnaire for their $<$ 240 items and events (and returned the completed questionnaire to the investigator). The questions posed in the questionnaire are shown in Figure 6.4.

Questions 2 and 3 in the questionnaire specifically look at whether the subject has previously retrieved the items/events and how likely they are to retrieve them in the future. Questions 1 and 4 - 6 examine the activity depicted by the item/event to help determine the factors that might influence a subject's desire to retrieve items/events in the future. Question 1 examines how habitual the activity represented by the

[6] It should be acknowledged however that subjects had prior exposure to some of this biometric period test set from preliminary prior experiments (see [Kelly and Jones, 2009, Kelly and Jones, 2010b] for full details on these studies), which we carried out one and nine months after biometric test set buildup. Comparison between these prior experiments and the current experiment is drawn in the discussion section at the end of this chapter.

	Sub1	Sub2	Sub3	Sub1	Sub2	Sub3
Technique	**removeEng**			**divEng**		
SenseCam						
GSR max	5	5	5	5	4	5
GSR ave	4	2	3	3	5	5
GSR min	3	5	5	3	2	5
HF max	3	3	5	5	5	5
HF ave	5	2	5	4	2	5
HF min	4	4	4	2	2	5
HR max	5	0	0	4	4	5
HR ave	5	0	0	3	2	5
HR min	5	0	0	2	2	5
ST max	2	5	5	2	3	5
ST ave	3	3	5	3	4	5
ST min	5	1	5	3	4	5
Computer Item						
GSR max	5	5	5	5	5	5
GSR ave	5	5	5	5	5	5
GSR min	5	5	5	5	5	5
HF max	5	5	5	5	5	5
HF ave	5	5	5	5	5	5
HF min	5	5	5	5	5	5
HR max	5	5	5	5	5	5
HR ave	5	5	5	5	5	5
HR min	5	5	5	5	5	5
ST max	5	5	5	5	5	5
ST ave	5	5	5	5	5	5
ST min	5	5	5	5	5	5
Total	*109*	*90*	*102*	*99*	*99*	*120*

Table 6.2: Number of SenseCam events and computer items retrieved per subject (Sub).

computer item or SenseCam images is for the subject. For example, for computer items sending an email requesting information from a researcher in a different research group might be very unusual, whereas writing a conference paper might be less unusual; for events depicted by the SenseCam a regular lunch date with a group of friends would be more habitual than attending a concert. It should be noted however that for some events which subjects could not recognise at all, e.g. where blank images were presented, subjects could not rate how habitual the event was. The remaining questions, i.e. questions 4, 5 and 6, examine the moment in time depicted by the item/event. Specifically they look at the subjects recall of the activity, how important the activity was to them at the time and the current importance of the activity. For questions 5 and 6 it was necessary to provide a sixth option to accommodate situations where subjects could not properly recall or distinguish the event/item from

1. Generally, considering the types of activities you engage in, how common/distinctive is the event/item? (5-point scale: 0 = Very habitual → 4 = Very unusual).
2. Have you retrieved this event/item to date? (5-point scale: 0 = I have never retrieved this event/item, 1 = I can't recall retrieving this event/item, but it is possible that I have, 2 = I have retrieved this event/item, 3 = I have retrieved this event/item several times, 4 = I regularly retrieve this event/item).
3. Is this an event/item that you would or would like to retrieve (or be presented with) in the future? (5-point scale: 0 = I definitely won't retrieve this event/item in the future, 1 = It is unlikely that I will retrieve this event/item in the future, 2 = I might retrieve this event/item in the future, 3 = I will probably retrieve this event/item in the future, 4 = I will definitely retrieve this event/item in the future).
4. How well do you recall this event/item? (5-point scale: 0 = Could not work out what the event/item was, 1 = I could only get a gist of what the event/item was, 2 = I could work out what the event/item was but it took some time, 3 = I could quickly work out what the event/item was, 4 = I could immediately recognise what the event/item was).
5. How important/significant was the event/item to you at the time? (5-point scale: 0 = Not at all important, 1 = Of some/little importance, 2 = Of average importance, 3 = Above average importance, 4 = Very important, and extra option X = Don't recall).
6. How important/significant is the event/item to you now? (5-point scale: 0 = Not at all important, 1 = Of some/little importance, 2 = Of average importance, 3 = Above average importance, 4 = Very important, and extra option X = Can't say).
7. Comments on your ratings.

Figure 6.4: Questions posed on the questionnaire completed by subjects.

the provided information. Question 7 was a free text optional field where subjects could provide details of the type of event they rated (e.g. working on computer) and an explanation of their ratings. Post questionnaire informal interviews with subjects provided further insight into their ratings. The following sections discuss the findings of this study.

6.3 Experiment Results and Analysis

Figures 6.5 and 6.7 (at the end of the chapter) show the results of user item ratings averaged over the three subjects for questions 1-6 of the questionnaire using the *removeEng* technique (i.e., the technique which removes biometric data associated with periods of high energy expenditure, described in the previous section). The average results of user item ratings across the three subjects for the questionnaire using the *divEng* technique (i.e., the technique which divides the biometric data by energy expenditure (multiplies in case of ST), described in the previous section) are shown in Figures 6.6 and 6.8 (at the end of the chapter). For ease of viewing and analysis purposes, for each of questions 1-6 we also grouped the 5-point scale ratings; that is a score of 3 or 4 was grouped together and a score of 0 or 1 was similarly grouped together. Figures 6.5 - 6.8 also show the percentages for each score averaged across the 3 subjects for questions 1-6 using this grouping. Figures 6.9 - 6.20 (at the end of the chapter) present the individual breakdown of results for each subject. More specifically, Figures 6.9 - 6.14 show the individual breakdown of results for each subject for SenseCam events using the *removeEng* and *divEng* techniques. Figures 6.15 - 6.20 show those for computer items using the *removeEng* and *divEng* techniques.

Using the results presented in these figures, in this section we analyse the average results obtained across the three subjects, explore the differences in performance for individual subjects and suggest ways the presented techniques may be used in the future. Specifically Section 6.3.1 analyses the results obtained for the SenseCam event extraction experiment and Section 6.3.2 analyses the results obtained for the computer item extraction experiment.

6.3.1 Detecting Important SenseCam Events

Our main objective in doing this experiment is to determine whether past biometric response associated with events can be used to detect SenseCam events which individuals might like to retrieve or be presented with in the future. Naturally, perception of what one might like to view in the future is just that - a perception. The events individuals will actually want to view can change over time due to personal circumstances, information needs, desires at given points in time, etc. Thus in the questionnaire when we ask subjects whether they would like to retrieve or be presented with different SenseCam events in the future, they can only rate this from their current perspective - e.g., a subject might now rate a past once off work meeting with a person as something they would not want to view in the future, however were they in the future to get to know this person on a personal level it might then be interesting to view the images relating to their first ever meeting; or a subject might now rate images of them in their home as mundane and not something they want to view in the future, however years from now if the subject is living somewhere different it might be interesting to view images of their previous home and the 'old fashioned' devices, etc, contained within it! This of course is pure speculation. All we attempt to do in this study is take the subject's current perspective and examine whether two years after SenseCam images were captured, the biometric response associated with these images is representative of images they might wish to view in the future. Question 3 of our questionnaire looks at this. The remaining questions help us to gain some insight into the types of events subjects might want to view from their collections and whether biometric levels relate to these types of event. This may be useful for future studies in this space which seek to progress the approach presented here - e.g. perhaps subjects are likely to want to retrieve events they have retrieved to date or events depicting a more unusual activity. We next examine the overall results obtained. We then examine results across individual subjects. This section then concludes with overall conclusions on the use of biometrics as an enabling technology for applications supporting SenseCam browsing.

6.3.1.1 Overall Results

Using *removeEng* technique:

Overall observed results suggest some relationship between GSR levels associated with images using the *removeEng* technique (See Figure 6.5) and subjects' desire to view the SenseCam events in the future. Subjects would not want to view 20% of the max GSR events in the future, this compares to not wanting to view 22% of the average GSR events and 46% of the min GSR events (see GSR results with rating of 0 for 'Retrieve in Future' in Figure 6.5). Considering both SenseCam events subjects would not and were unlikely to want to view in the future shows that subjects would not or are unlikely to want to view 27% of max GSR events compared to 55% of ave GSR and 61% of min GSR events respectively (see GSR results with rating of 0 and 1 for 'Retrieve in Future' in Figure 6.5). Relationship is also observed between GSR levels and importance of the events at the original time of event occurrence, current importance of the events depicted by the images and images that have been retrieved to date. This also relates to how habitual the events were and subjects recall of the events. See GSR results in Figure 6.5.

HF levels were also a good indicator of images subjects would like to view again in the future using the *removeEng* technique, also shown in Figure 6.5. Subjects would not want to view 18% and would not or were unlikely to view 27% of max HF events, compared to not wanting to view 42% of ave HF events and 58% of min HF events and not wanting or being unlikely to want to view 50% of ave and 75% of min HF events (see HF results with ratings of 0 and 1 for 'Retrieve in Future' in Figure 6.5). HF levels also correlated with images subjects have retrieved to date. This too is related to how habitual the events were, but had no bearing on subjects' recall of the events, importance at time of event occurrence or current importance of events. See HF results in Figure 6.5.

No relationship was observed between HR and ST using the *removeEng* technique for the questions posed in our questionnaire (see HR and ST results in Figure 6.5).

Using *divEng* technique:

The *removeEng* technique removed periods of high energy expenditure from the collection, as described in Section 6.2.1.1. The remaining biometric levels are raw biometric readings which do not consider the associated energy expenditure levels. We speculated that the *divEng* technique which divides the biometric levels by energy expenditure (multiplies in the case of ST) might better get at the true biometric responses with external factors removed, as this technique factors associated energy expenditure levels into all biometric readings (described in Section 6.2.1.1). Figure 6.6 presents the average results obtained using the *divEng* technique to extract SenseCam events for the three subjects.

Averaged across the three subjects, are speculation that the *divEng* technique would perform better than the *removeEng* technique did not hold true for HF. Dividing HF values by energy expenditure no relationship between HF levels and the questions posed in our questionnaire were observed (see HF results in Figure 6.6). This is in contrast to the results observed when using HF with the *removeEng* technique described previously. However, different patterns emerge when we examine the performance for individual subjects, described later.

GSR levels still correlated to a certain extent with subjects' desire to retrieve the images in the future using the *divEng* technique. Subjects would not want to retrieve 29% of max GSR events, compared to not wanting to retrieve 46% and 30% of average and min GSR events respectively, and would not or were unlikely to want to retrieve 29%

of max GSR, 62% of average GSR and 50% of min GSR events (see GSR results with ratings of 0 and 1 for ‘Retrieve in Future’ in Figure 6.6). Some relationship was also shown between GSR levels and images current importance using the *divEng* technique (see GSR results for ‘Current Importance’ in Figure 6.6). However, in this case, this did not relate to whether subjects had retrieved the images to date, importance at time of event occurrence, recall of the events, nor how habitual the events were (see GSR results in Figure 6.6).

Similar to the use of HR levels with the *removeEng* technique, using the *divEng* technique HR levels did not provide insight into the events subjects may wish to view in the future (see HR results for ‘Retrieve in Future in Figure 6.6). Despite this HR levels did correlate with the distinctiveness of events and the importance of the events at the time of occurrence to the subjects (see HR results in Figure 6.6). This suggests that while HR levels might be useful for inferring events which are important to our subjects as they are occurring, and also for detecting distinctive events from their collections, events of this nature detected through HR levels are not necessarily the events subjects will want to view in the future.

Using the *divEng* technique with ST did not provide insight into the events subjects may wish to view in the future. Nor did ST levels show relationship with the other questions posed in the questionnaire (see ST results in Figure 6.6). This is similar to the results observed previously using ST levels in the *removeEng* technique where there was no relationship between ST levels and the six questions posed in the questionnaire.

6.3.1.2 Performance Across Individual Subjects

In this section we analyze the results obtained for each of the three subjects. A summary of these results are provided in Figure 6.3 (at the end of the chapter).

Comparison with Average Results:

Analysing the individual ratings of subjects (see Figures 6.9 - 6.14) shows that the relationship between GSR using the *divEng* technique, and both GSR and HF using the *removeEng* technique, and subjects desire to view SenseCam events in the future are present for Subject 1. Subjects 1 and 3 only show such relationship for HF using the *removeEng* technique. Additionally, further relationships which were not present in the average results (described in the previous section) are also observed, particularly in the case of Subjects 1 and 2. These relationships are discussed in this section.

Patterns of Past Interaction with SenseCam Collections (‘Retrieved to Date’):

Different patterns of past interaction with SenseCam collections were noted in these figures (see ‘Retrieved to Date’ in Figures 6.9 - 6.14). Subject 2 had never retrieved the images presented to them in this study. In fact, on questioning this subject, they have never looked back over their SenseCam images, with the exception of one episode which they decided to consult to locate information captured in the images. Subject 1 looked at very few of the SenseCam events presented to them in the study, and on consultation with this subject it was found that they have rarely browsed their SenseCam collection to date. Subject 3 on the other hand had viewed more of the SenseCam events presented to them in this study in the past, and quite regularly, relative to the other subjects, browses their SenseCam collection.

Using the *removeEng* Technique:

As mentioned at the beginning of this section, GSR levels and HF levels to a certain extent correlated with Subject 1′s desire to retrieve SenseCam events in the future using the *removeEng* technique (see ′Retrieve in Future′ in Figure 6.9). Further, GSR levels also correlated here with this subject′s desire to retrieve SenseCam events in the future (Spearmans Rank Correlation Coefficient, rho= 0.698, p<0.05). GSR levels, using this technique, also correlated with how distinctive the event captured in the SenseCam images was, recall of the event, whether the images had been retrieved to date, the importance of the event to the subject at the time of occurrence and the current importance of the event to the subject. Relationship was also observed for HF levels, using the *removeEng* technique and events retrieved to date for this subject, see GSR and HF results in Figure 6.9.

HF levels, using the *removeEng* technique, also showed some relationship with Subject 2′s desire to retrieve SenseCam events in the future, as mentioned at the beginning of this section. This too showed some relationship with the distinctiveness of the events. See HF results for ′Retrieve in Future′ and ′Common/distinctive Activity′ in Figure 6.11. While using the *removeEng* technique GSR levels also showed some relationship for Subject 2 with the distinctiveness of events, recall of events, importance of events at the time of occurrence and current importance of events, this did not correlate with the subject's desire to view events in the future (see GSR results in Figure 6.11). Another biometric measure, namely ST levels, did however show a relationship with this subject's desire to view events in the future using the *removeEng* technique. ST levels for this subject also correlated with the current importance of the events depicted in the SenseCam images to the subject, see ST results in Figure 6.11.

The only relationship between biometric levels and the subject's desire to retrieve events in the future observed for Subject 3 using the *removeEng* technique was for HF levels. HF levels also correlated with events the subject had retrieved to date and to a certain extent with how distinctive the events were, see Figure 6.13.

While certain levels of relationship between biometric response levels and desire to view events were observed using the *removeEng* technique, as highlighted at the beginning of this section, the *divEng* technique performed better. We next examine this effect.

Using the *divEng* Technique (Subjects 1 & 2):

The *divEng* technique is better than the *removeEng* at pin-pointing events Subjects 1 and 2 might view in the future (see 'Retrieve in Future' in Figures 6.9 and 6.11 for results using the *removeEng* technique, and Figures 6.10 and 6.12 for results using the *divEng* technique for Subjects 1 and 2 respectively). The *divEng* technique gives superior performance than the *removeEng* technique for Subjects 1 and 2, in particular for ST, HR and HF. Relationship was observed for both Subjects 1 and 2 for ST, HR and HF levels and the subject's desire to view events in the future using the *divEng* technique (correlation was also observed here for Subject 2, for ST and HR: Spearmans Rank Correlation Coefficient - ST: rho= 0.629, $p<0.05$; HR: rho= 0.713, $p<0.05$). In fact, neither of these subjects had any desire to retrieve SenseCam events in the future which did not have an associated max ST, HR or HF response (see Subject 1's and Subject 2's results for 'Retrieve in Future' in Figures 6.10 and 6.12 respectively).

Analysing the images with associated max biometric response for HR, HF and ST which these subjects did not wish to view in the future, we found that they were blank images, images captured when driving, images depicting working in isolation on a computer and images showing beneath a table. These types of images could simply be distinguished in the future by use of automatic image analysis techniques and not presented to subjects. Subjects could recall some of these computer events with associated max biometric response which they did not wish to view in the future, due to distinguishing factors such as working on another person's computer for example. Indeed some of these events were also important to the person at the time. While we endeavoured to remove SenseCam events depicting the subject working on their PC through the use of the computer activity we logged with Slife, some PC activity was not removed using this technique due to subjects stopping Slife tracking (during high computer processing activities), Slife crashes, and subjects working on PCs other than their own (as described in Section 6.2.1.1).

We also note in Figure 6.10 that Subject 1, who had retrieved some items to date, had only retrieved events with max biometric response for ST, HR and HF using the *divEng* technique. However this had little relationship with how common/distinctive the events were, the subject's recall of the events, the importance of the events at the time of event occurrence, and current importance of the events (see Figure 6.10).

As mentioned earlier in this section, Subject 2 has not retrieved the events presented to them in this study to date. However, relationship between biometric levels and this subject's recall of the events and perceived current importance of the events was observed using the *divEng* technique for HR, HF and ST (see Figure 6.12).

Regarding, the utility of GSR levels using the *divEng* technique in detecting events Subjects 1 and 2 might want to retrieve in the future, some relationship was noted for Subject 1 but none for Subject 2 (see GSR results for 'Retrieve in Future' in Figures 6.10 and 6.12). While GSR levels using the *divEng* technique related to Subject 1's desire to view the events in the future (correlation was also observed here: Spearmans Rank Correlation Coefficient, rho= 0.611, $p<0.05$), they did not relate to the other questions posed in our questionnaire (see Figure 6.10). However, despite not relating to Subject 2's desire to view events in the future, GSR levels using the *divEng* technique did correlate with the subject's recall of the events and their perceived current importance of the events (see Figure 6.12).

Using the *divEng* Technique (Subject 3):

The relationships between biometric levels and subjects' desire to view events in the future observed for Subjects 1 and 2 using the *divEng* technique were not present for Subject 3 (see 'Retrieve in Future' in Figure 6.14). Indeed no relationship between biometric levels and the questions posed in the questionnaire was observed for this subject.

Using the *divEng* technique this subject had previously retrieved some of the events presented to them from each of the max, min and average biometric categories, with the exception of events from the average HR category, none of which had previously been retrieved. Further, they only rated one event (a max GSR event) as being an event they would definitely not want to view in the future (see Figure 6.14). Incidentally this event, which the subject would not want to view again, was talking to someone who was regularly encountered. It should be noted though that no conclusions can be drawn from this observation, as other events showing people regularly encountered were also presented to the subject in the study, and for these other events the subject did not state that they definitely would not want to view the events again in the future. Further, no distinguishing characteristics were noted in the events the subject would be most interested in viewing again.

This subject, as distinct from the other two subjects, has often looked at their SenseCam images in the past. Also, as distinct from the other subjects, enjoyed looking at what might be considered mundane repetitive events such as images depicting cooking in a kitchen, watching TV, etc. On questioning this subject as to their SenseCam reviewing behaviour, they stated that they found it interesting to view such events again. This is similar to the findings of [Harper et al., 2008] where mundane SenseCam events depicting one's life were found to be interesting to view, albeit this study was dealing with subjects for who it was novel to be wearing a SenseCam and thus also novel to be viewing SenseCam images of their lives, while our subjects are regular SenseCamers. Further their subjects were looking at images from the recent past, while our subjects were looking at images which were captured some 22 months previously.

6.3.1.3 Concluding Remarks

Overall results showed that the *divEng* technique performed best for Subjects 1 and 2 in detecting events they might want to view in the future (with ST showing the greatest utility in detecting events the subjects might want to view in the future). All biometric measures showed some utility for Subject 1 in this regard, and all but GSR showed utility for Subject 2 in this regard. The *divEng* technique did not show utility in detecting events Subject 3 might want to view in the future. Using the *removeEng* technique, GSR and HF levels correlated with Subject 1's desire to view events in the future. ST and HF levels showed utility for Subject 2 here. While HF levels using the *removeEng* technique were the only measure which showed any relationship with desire to view events in the future for Subject 3.

As we saw in this section, while GSR performed best for Subject 1 in detecting events they might want to view in the future using the *removeEng* technique, the other measures performed better in this regard using the *divEng* technique. Showing that factoring energy expenditure is important for HF, HR and ST, but has a negative effect on GSR for this subject. This however did not hold true for Subject 2, where GSR was not shown to provide utility in this regard using either technique. We also saw in these results that biometric levels were not particularly useful in detecting events which Subject 3, who regularly browsed their SenseCam collection, might want to view in the future.

Overall from these results we conclude that for subjects who infrequently browse their collections use of biometrics in locating interesting items combined with image analysis techniques (e.g. do not present individuals with blank images or blurred images) may be a good enabling technology component in lifelog search and evaluation (as we saw in this section for Subjects 1 and 2). However, for individuals who regularly browse their collections and find everything from the mundane to the unusual interesting, biometrics may not be useful (as we saw in this section for Subject 3). We acknowledge that this study was conducted on a small sample of individuals and that further analysis is required with a much larger set of subjects to properly substantiate.

6.3.2 Detecting Important Computer Items

Similar to the SenseCam events, our main objective in doing this experiment is to determine whether past biometric response associated with computer items can be used to detect items which an individual might like to retrieve or be presented with in the future from their lifelog. The next section examines the overall results obtained, following which Section 6.3.2.2 examines performance across individual subjects. The section concludes with a summary of the overall conclusions on the use of biometrics in detecting important computer items which can be made on the basis of this study.

6.3.2.1 Overall Results

Using the *removeEng* Technique:

Overall results suggest that the *removeEng* technique is not useful for detecting computer items which subjects may wish to view in the future (see 'Retrieve in Future' in Figure 6.7). The only relationship observed between biometric levels and subjects' desire to retrieve items in the future using the *removeEng* technique is the small relationship observed using ST levels. Subjects would want to retrieve 20% of items with max ST, $>$ 10% of items with average ST and $<$ 10% of items with associated min ST using the *divEng* technique (see ST results with rating of 4 for 'Retrieve in Future' in Figure 6.7). Subjects would or probably would want to retrieve $<$ 50% of items with max ST, 20% of items with average ST and $<$ 30% of items with min ST (see ST results with rating of 3 and 4 for 'Retrieve in Future' in Figure 6.7). ST levels here also have marginal relationship with subjects recall of the items, the current importance of the items to the subjects and whether the subjects have retrieved the items to date, but no relationship with the other questions posed in the questionnaire (see Figure 6.7).

Using the *divEng* Technique:

The *divEng* technique shows greater utility in locating items subjects may wish to view in the future (see 'Retrieve in Future' in Figure 6.8) using ST levels. Here subjects would want to retrieve in the future 40% of items with max ST verses $<$ 10% of items with average ST and $<$ 10% of items with min ST (see ST results with rating of 4 for ST for 'Retrieve in Future' in Figure 6.8). They would or probably would wish to retrieve 60% of items with max ST, this compares with $<$ 10% of average ST items and $<$ 10% of min ST items (see ST results with ratings of 3 or 4 for ST for 'Retrieve in Future' in Figure 6.7). A limited relationship between ST levels and the current importance of items, whether the items have been retrieved to date and importance at the time of previous interaction with the items is also observed here for subjects (see ST results in Figure 6.8).

Some relationship is also observed, using the *divEng* technique, between HR levels and subjects' desire to retrieve items in the future. Using HR levels with the *divEng* technique subjects would or probably would want to retrieve in the future $>$ 20% of items with max HR, this compares with $<$ 10% of average HR items and 0% of min HR items (see HR results with ratings of 3 or 4 for 'Retrieve in Future' in Figure 6.8).

A relationship is also observed between HR levels, using the *divEng* technique, and both items subjects have retrieved to date and the current importance of the items to the subjects (see HR results in Figure 6.8).

HF levels, using the *divEng* technique, show some relationship with the distinctiveness of the items in subjects' collections. HF levels, using the *divEng* technique, also correlate with whether the items have been retrieved to date , subjects recall of items and the importance to the subjects of the items at the time of previous interaction. However, HF levels, using the *divEng* technique, do not correlate with the current importance of the items to the subjects and show little relationship with their desire to retrieve the items in the future, see results for HF in Figure 6.8.

GSR levels, using the *divEng* technique, while correlating with the distinctiveness of the items in subjects collections, do not relate to the subjects' desire to retrieve items in the future, nor to any of the other questions posed in the questionnaire (see results for GSR in Figure 6.8).

Overall then we find that across the 3 subjects, ST levels using the *divEng* technique, show the greatest utility in locating the computer items which the subjects may wish to view in the future. In the next section we analyse performance across each individual subject.

6.3.2.2 Performance Across Individual Subjects

In this section we analyze the results obtained for each of the three subjects. A summary of these results are provided in Figure 6.4 (at the end of the chapter).

Comparison with Average Results:

Analysing the results observed across individual subjects (Figures 6.15 - 6.20), we find that the relationship between ST and HR using the *divEng* technique and desire to view the items in the future, observed in the average results shown in the previous section, are present for Subjects 1 and 3 (see Subject 1 and 3's ratings for ST and HR for 'Retrieve in Future' in Figures 6.16 and 6.20 respectively). Beyond this relationship other biometric types levels are shown to correspond with a desire to view items in the future for our 3 subjects using the *divEng* technique (see ratings for 'Retrieve in Future' in Figures 6.16, 6.18 and 6.20).

The relationship between subjects' desire to view items in the future and ST levels using the *removeEng* technique found in the previous section, is only present for Subject 1 (see ST ratings for 'Retrieve in Future' in Figure 6.17). Indeed no other relationships between biometric response and the subjects' desire to view the items in the future using the *removeEng* technique is observed across the three subjects (see ratings for 'Retrieve in Future' in Figures 6.15, 6.17 and 6.19). This suggests that factoring of energy expenditure readings into biometric levels is important for computer items. For the remainder of this section we analyse these results in greater detail.

Subject 1 Results Analysis:

As mentioned at the beginning of this section, ST and HR levels, using the *divEng* technique, show utility for Subject 1 in locating items that the subject may wish to retrieve in the future (see ST and HR ratings for 'Retrieve in Future' in Figure 6.16). In addition to this, HF levels using the *divEng* technique, also correlate to a certain extent with the subject's desire to retrieve items in the future (see HF ratings for 'Retrieve in Future' in Figure 6.16). Further, correlation between both ST and HF levels and events this subject may wish to view in the future are also present (Spearmans Rank Correlation Coefficient - ST: rho= 0.845, p<0.05; HF: rho= 0.574, p<0.05).

Using ST, with the *divEng* technique, this subject would definitely want to retrieve 40% of items with associated max ST, 0% of items with average ST and 0% of items with min ST (see ratings of 4 for ST for 'Retrieve in Future' in Figure 6.16). ST levels here also correlate with the current importance of items for this subject, but not with the other questions posed in the questionnaire (see Figure 6.16).

Using HR, with the *divEng* technique, the subject would definitely or probably want to retrieve 60% of items with associated max HR, 20% of items with average HR and 0% of items with min HR (see ratings of 3 and 4 for HR for 'Retrieve in Future' in Figure 6.16). HR levels here, using the divEng technique also correlated with the subjects recall of items and the current importance of the items, but not with the other questions posed in the questionnaire (see Figure 6.16).

Finally, using HF, with the *divEng* technique, the subject would probably want to retrieve 40% of items with max and average HF and 0% of items with min HF (see ratings of 3 for HF for 'Retrieve in Future' in Figure 6.16). A limited relation between the subject's recall of the items and the importance of the items to the subject in the past is also observed. No relationship between HF levels, using the *divEng* technique, and the other questions posed in the questionnaire is observed, see results for HF in Figure 6.16.

No relationship was observed for Subject 1 between GSR levels, with the *divEng* technique, and the questions posed in our questionnaire (see GSR results in Figure 6.16).

As mentioned at the beginning of this section, using the *removeEng* technique, the only relationship observed for Subject 1 and desire to view items in the future was that observed using ST levels (correlation is also present here: Spearmans Rank Correlation Coefficient, rho= 0.560, $p<0.05$). Using ST levels, with the *removeEng* technique, the subject would definitely want to retrieve 40% of items with associated max ST, and 0% of items with average and min ST (see ratings of 4 for ST for 'Retrieve in Future' in Figure 6.15). ST levels did not correlate with any of the other questions in the questionnaire for this subject (see ST results in Figure 6.15).

Beyond ST levels, using the *removeEng* technique, the only other relationship observed for the questions posed in the questionnaire was between HR and HF levels and the importance of the items during past interaction for the subject, and between GSR levels and the distinctiveness of the items (see results in Figure 6.15).

Subject 3 Results Analysis:

As mentioned at the beginning of this section, for Subject 3, similar to Subject 1, ST and HR levels show a relationship with the subject's desire to view items in the future using the *divEng* technique, see ST and HR results for 'Retrieve in Future' in Figure 6.20.

Subject 3 would definitely or probably want to retrieve 60% of the items with associated max ST, using the *divEng* technique, in the future, relative to 0% of items with associated average and min ST levels (see ST results with ratings of 3 and 4 for 'Retrieve in Future' in Figure 6.20). ST levels here also showed some relationship with the subject's recall of items, whether they had retrieved the items to date and the current importance of the items to the subject (see results for ST in Figure 6.20). Relationship between ST levels, using the *divEng* technique, was not observed for the importance of the items at the time of previous access to the subject and the distinctiveness of the items (again see results for ST in Figure 6.20).

Using HR levels with the *divEng* technique, Subject 3 would definitely or probably want to retrieve 20% of the items with associated max HR levels and 0% of items with average and min HR levels (see HR results with ratings of 3 and 4 for 'Retrieve in Future' in Figure 6.20). No relationship was observed between HR levels, using the *divEng* technique, and the other questions posed in the questionnaire (see results for HR in Figure 6.20).

Relationship was not observed between Subject 3's desire to view items in the future and GSR or HF levels using the *divEng* technique (see GSR and HF results in Figure 6.20). The only relationship observed between either of these biometric measures, using the *divEng* technique for this subject and the questions posed, was the limited relationship between HF levels and the subject's recall of the items (see 'Recall' results for HF in Figure 6.20).

As can be seen in Figure 6.19, no relationship was observed between GSR, HR, HF or ST levels, using the *removeEng* technique, and any of the questions posed in the questionnaire for Subject 3.

Subject 2 Results Analysis:

Across the three subjects, Subject 2's biometric levels showed the least relationship with computer items for the questions posed in our questionnaire. Using the *removeEng* technique the only relationship observed was that between ST levels and the distinctiveness of the items (see ratings for 'Common/distinctive Activity' in Figure 6.17).

Using the *divEng* technique GSR levels were the only biometric measure which correlated somewhat with Subject 2's desire to retrieve items in the future (see results for 'Retrieve in Future' in Figures 6.17 and 6.18). Here the subject would want to retrieve 40% of items with associated max GSR, compared to wanting to retrieve 20% of items with average GSR and 20% of items with min GSR (see ratings of 4 for 'Retrieve in Future' for GSR in Figure 6.18). GSR levels, using the *divEng* technique, also showed some relationship with items the subject had retrieved to date (see GSR results for 'Retrieved to Date' in Figure 6.18).

The only other relationship observed between biometric measures using the *divEng* technique and the questions posed in our questionnaire was the relationship between the distinctiveness of items and HR levels (see HR ratings for 'Common/distinctive Activity' in Figure 6.18).

The relationship between GSR levels and desire to view items in the future observed for Subject 2 using the *divEng* technique is the opposite of what was observed for the other two subjects, where GSR levels were not useful in this regard. Further, for the other two subjects relationship between HR, HF and ST levels using the *divEng* and the subjects' desire to view the items in the future were observed - this relationship was present for Subject 1 for HR, HF and ST levels; and for Subject 3 for HR and ST levels (as described earlier in this section).

6.3.3 Concluding Remarks

Overall the *divEng* technique shows greatest utility in locating items subjects might want to view in future. However while GSR proves most useful for Subject 2, it is not useful for Subjects 1 and 3, and contrastingly HR and ST which are useful for Subjects 1 and 3 (ST being the more useful of the two) are not useful for Subject 2. While this anomaly cannot be explained here, the utility of biometric response with energy expenditure factored in provides support for exploring the use of this approach to re-rank ranked result lists. This analysis is presented in the next chapter. In this next chapter we explore in greater detail the types of biometric responses which provide greatest utility in helping locate relevant items from collections, specifically in re-ranking ranked result lists.

6.4 Discussion

Preliminary experiments exploring the utility of biometric response associated with past experience of lifelog items in detecting the importance of lifelog items to individuals were conducted on an earlier version of the test sets used in this chapter [Kelly and Jones, 2009, Kelly and Jones, 2010b]. These experiments showed a relationship between biometric response and the importance of lifelog events to individuals (albeit 1 and 9 months after the lifelog data had been captured as opposed to the 22 month time interval explored in this chapter). However, these experiments suffered a number of shortcomings in the experimental procedure adopted. In particular, the biometric data from the beginning and end of each period of wearing the biometric devices which was skewed (as discussed in Chapter 3.3.4.4) was not removed from the dataset. Also, energy expenditure values were not calculated on the given subjects personal details (i.e., age, weight and height), due to the devices not being set with the subjects' personal details (as discussed in Chapter 2.3.1 these personal details are used in the energy expenditure calculation). The experiments presented in this chapter addressed these problems.

Further, a number of lessons were learned from these earlier experiments and subsequently the experiments presented in this chapter differ in a number of key ways:

1. In this chapter's experiments for SenseCam events a 20 second window was allowed each side of the events;

2. An additional form of biometric data, heat flux, was considered in this chapter's experiments;

3. The earlier experiments showed failure to capture textual content for computer items to negatively impact on a subject's ability to recall items in some instances, hence items missing textual content were not considered in the experiments presented in this chapter;

4. Our prior experiments presented subjects with temporal groups of computer items (events) to rate, however many of the items in such events were unrelated (e.g. viewing email and then returning to a coding task). Subjects' rating such events does not provide an indication of the role of biometric response in detecting the future importance of individual computer items, hence in the study presented in this chapter subjects were presented with single items to rate;

5. In our prior experiments the only way physical activity and other external factors were accounted for was through the deletion of biometric data captured during periods of high energy expenditure (i.e. an earlier version of the *removeEng* technique). Similar to use of the *removeEng* technique in this chapter, the only computer and SenseCam events which were available for presentation to the subjects' were the ones which occurred at the times of the remaining biometric data captured. In using this approach we removed the possibility for suggestion of potentially interesting computer and SenseCam events which occurred during the periods of deleted biometric levels. As highlighted earlier, it is also possible in using this technique that the remaining biometric levels were still influenced by external factors. Hence in the experiments presented in this

chapter, we also explored factoring energy expenditure levels into the biometric readings (i.e. the *divEng* technique) as opposed to just looking at the removal of periods of high energy expenditure.

While these factors would have negatively impacted on the experiments in the prior work, the experiments nevertheless provided an insight into the utility of biometric response in locating events from lifelogs which individuals may be interested in viewing. The experiments also provide us with some insight into how the SenseCam images individuals wish to view change over time. This was particularly apparent for Subject 3, who soon after capturing SenseCam images and while still actively engaged in SenseCam image capture, did not wish to view a lot of mundane events such as cooking dinner captured by the SenseCam. However over time this changed and while no longer lifelogging and capturing such repetitive life events, the subject found pleasure in viewing these 'mundane' events (as discussed earlier in this chapter). We can only speculate as to whether further in the future the same will hold true for the other subjects. However we feel that individual difference will come into play both from the point of view of what will be interesting to view and how often subjects will consult their lifelogs. Future work should consider use of qualitative user studies to establish the different patterns of lifelog browsing exhibited by individuals. Lifelog viewing patterns will also, we believe, depend on personal circumstances at given moments in time and on the different stages in one's life.

From the findings of our studies we speculate that biometrics may prove to be a more useful tool for occasional lifelog browsers, as opposed to those who appear to browse collections on a more regular basis. We also believe that biometrics on their own are by no means the full solution, however integrated into an application which supports lifelog browsing using category facilities, e.g., show me images on topic x, images in y location, images with z person in them, it could prove a useful tool. Further investigation on this topic appears to be justified for future work.

6.5 Conclusions

In this chapter we set out to explore the role of biometric response in detecting important items within lifelogs. We investigated whether items coincident with maximum observed biometric galvanic skin response (GSR), heat flux (HF) and heart rate (HR) and with minimum observed skin temperature (ST) readings were more important to subjects, and whether this would mean they would be most useful or interesting for subjects to view in the future. From this study, relationship between biometric levels and both SenseCam event and computer item importance was observed. The SenseCam event selection results are important since ability to extract interesting events from vast SenseCam collections is challenging but important, if these archives are to have long-term use. As mentioned previously, while these results are promising, it is acknowledged that this study was conducted on a limited number of subjects over a short period of time. Investigation using more participants over a longer timeframe is required to further test our suggested conclusions.

Overall we saw that HF levels with periods of high energy expenditure removed (i.e., *removeEng* technique) is most beneficial for detecting SenseCam events subjects may wish to view in the future, but that consideration of other additional factors beyond biometric response may improve performance. ST proved most beneficial for detecting computer items individuals may wish to view in the future using the *divEng* technique.

However, different pictures began to emerge when we looked at the results of individual subjects. HF levels using the *removeEng* technique was the only one which showed relationship with subjects' desire to view SenseCam events in the future across the 3 subjects. Indeed, for Subject 3 who regularly browsed their SenseCam collection, no other relationship between biometric response (using either the *removeEng* or *divEng* technique) and the subject's desire to view SenseCam events in the future was observed. However, for Subjects 1 and 2 who rarely if ever browsed their SenseCam collections, the *divEng* technique, showed greatest utility in detecting SenseCam events these subjects might want to view in the future, with ST levels showing the greatest utility in this regard. For computer items, while ST levels using the *divEng* technique proved to be the most useful technique in detecting items Subjects 1 and 3 might want to view in the future, this technique did not prove useful for Subject 2. For Subject 2 GSR level using the *divEng* technique was the only technique which showed relationship with items that the subject might wish to view in the future.

Overall from these results we can suggest that factoring of energy expenditure into the individual biometric readings (through the use of the *divEng* technique) is important. However, it is hard to draw conclusions on a biometric measure which may prove most useful in detecting events/items subjects may wish to view in the future. Rather, we conclude that biometric levels associated with past experience of lifelog items seem to show promise in detecting important items/events in individuals personal digital collections. Future work needs to be done to explore the nature of biometric response associated with lifelog items and the role of this response in detecting important lifelog items in greater detail. This caveat notwithstanding, the results observed in this chapter support investigating the use of the biometric tags that can be assigned to lifelog items as static scores for ranked retrieval of personal items in a lifelog. The next chapter combines the work of this chapter with the techniques for ranked retrieval from lifelogs investigated in Chapter 5.

Subject 1		**Subject 2**		**Subject 3**	
Retrieve in Future	**Other Questions**	**Retrieve in Future**	**Other Questions**	**Retrieve in Future**	**Other Questions**
divEng					
< GSR			GSR: current importance; recall		
ST	ST: retrieved to date	**ST**	ST: current importance; recall		
HR	HR: retrieved to date	HR	HR: current importance; recall		
HF	HF: retrieved to date	HF	HF: current importance; recall		
removeEng					
GSR	GSR: distinctiveness; current importance; importance at time; recall; retrieved to date		GSR: distinctiveness; current importance; importance at time; recall		
		ST	ST: current importance		
HF	HF: retrieved to date	< HF	HF: distinctiveness	**HF**	HF: retrieved to date; < distinctiveness

Table 6.3: Summary of the relationship observed between subjects' ratings for SenseCam events and biometric levels. Best performing biometric measure for each subject highlighted in bold font, '<' refers to little relationship.

Subject 1		Subject 2		Subject 3	
Retrieve in Future	**Other Questions**	**Retrieve in Future**	**Other Questions**	**Retrieve in Future**	**Other Questions**
divEng					
		< **GSR**	GSR: retrieved to date		
ST	ST: current importance			**ST**	ST: current importance; recall; retrieved to date
HR	HR: current importance; recall		HR: distintiveness	< HR	
< HF	HF: importance at time; recall				< HF: recall
removeEng					
	GSR: distinctiveness				
ST			ST: distinctiveness		
	HR: importance at time				
	HF: importance at time				

Table 6.4: Summary of the relationship observed between subjects’ ratings for computer items and biometric levels. Best performing biometric measure for each subject highlighted in bold font, ’<’ refers to little relationship.

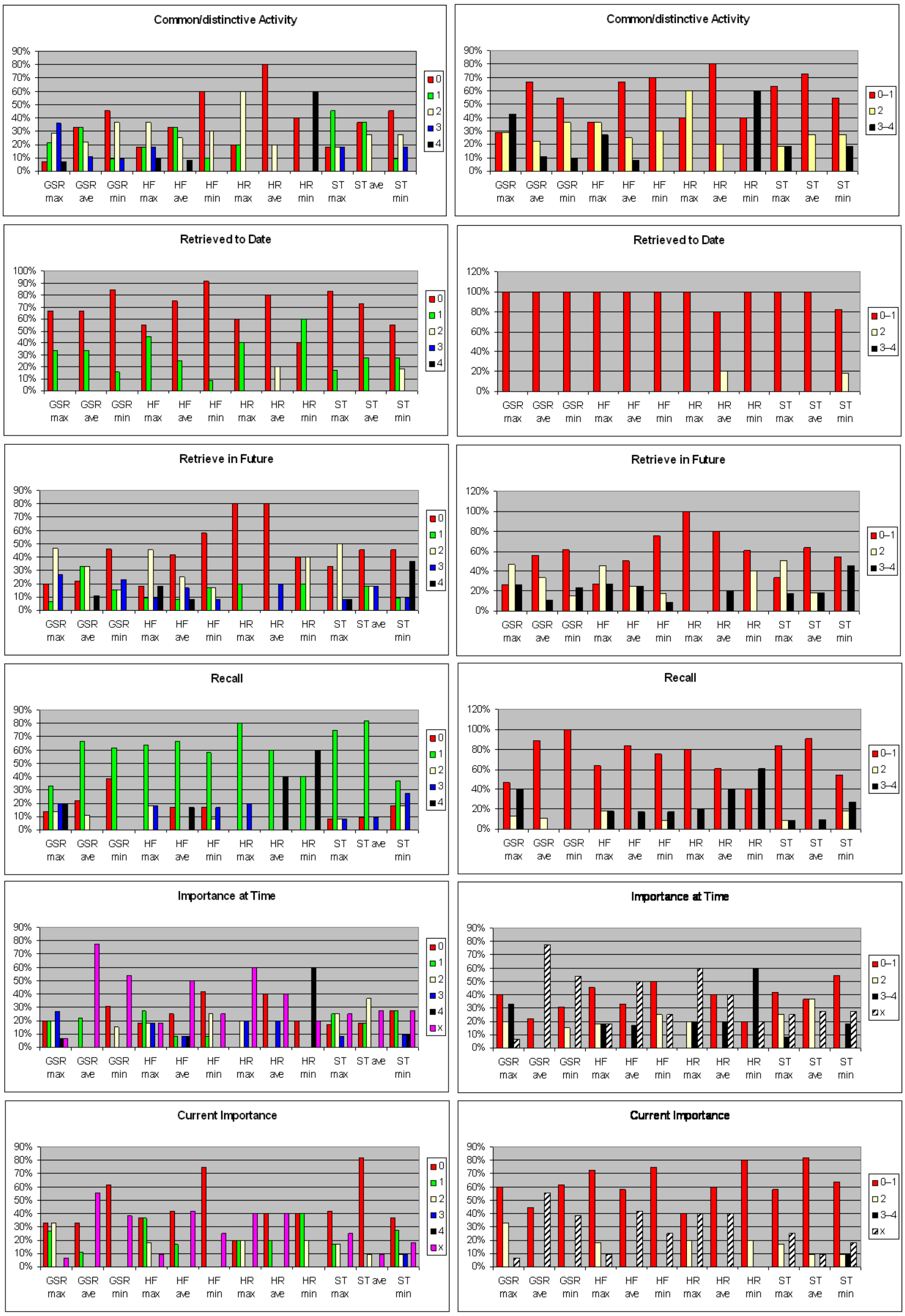


Figure 6.5: Questionnaire results - average SenseCam event ratings for the three subjects combined for max, average (ave) and min GSR, HR, HF and ST using the *removeEng* technique. Note the graphs on the left column present a break down of the results, while the right column graphs present a grouping of the questions 5-point scale.

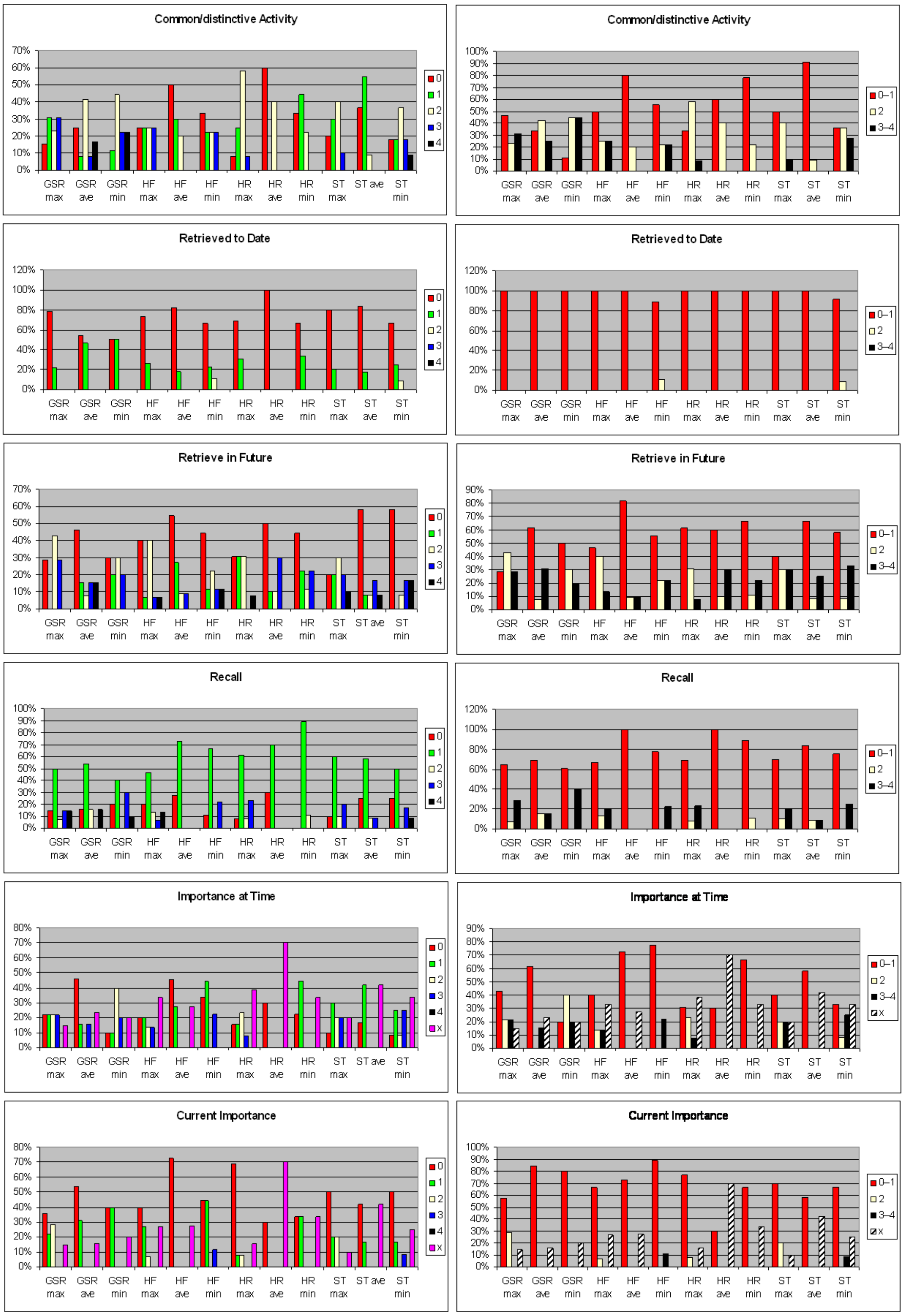


Figure 6.6: Questionnaire results - average SenseCam event ratings for the three subjects combined for max, average (ave) and min GSR, HR, HF and ST using the *divEng* technique. Note the graphs on the left column present a break down of the results, while the right column graphs present a grouping of the questions 5-point scale.

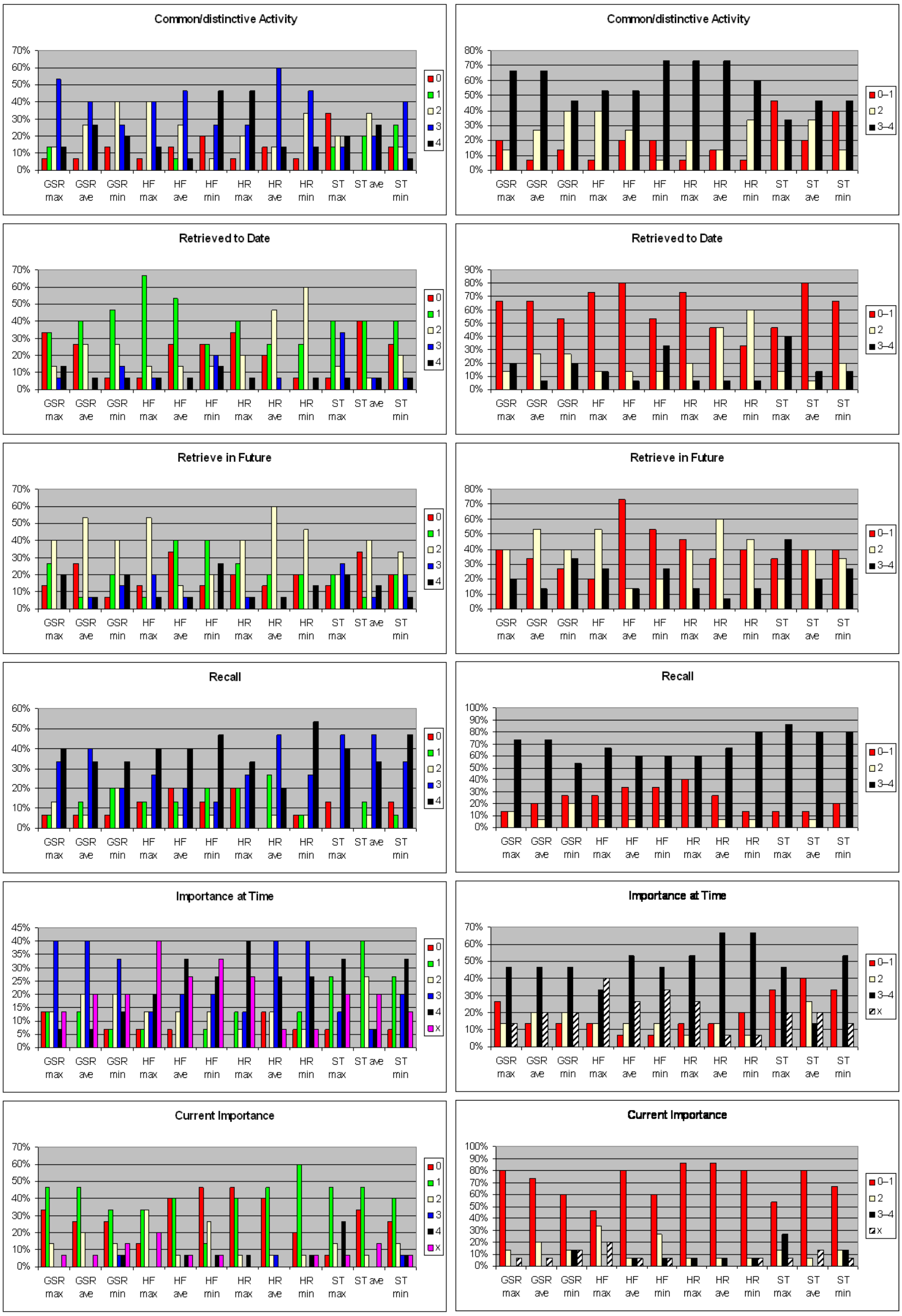


Figure 6.7: Questionnaire results - average computer item ratings for the three subjects combined for max, average (ave) and min GSR, HR, HF and ST using the *removeEng* technique. Note the graphs on the left column present a break down of the results, while the right column graphs present a grouping of the questions 5-point scale.

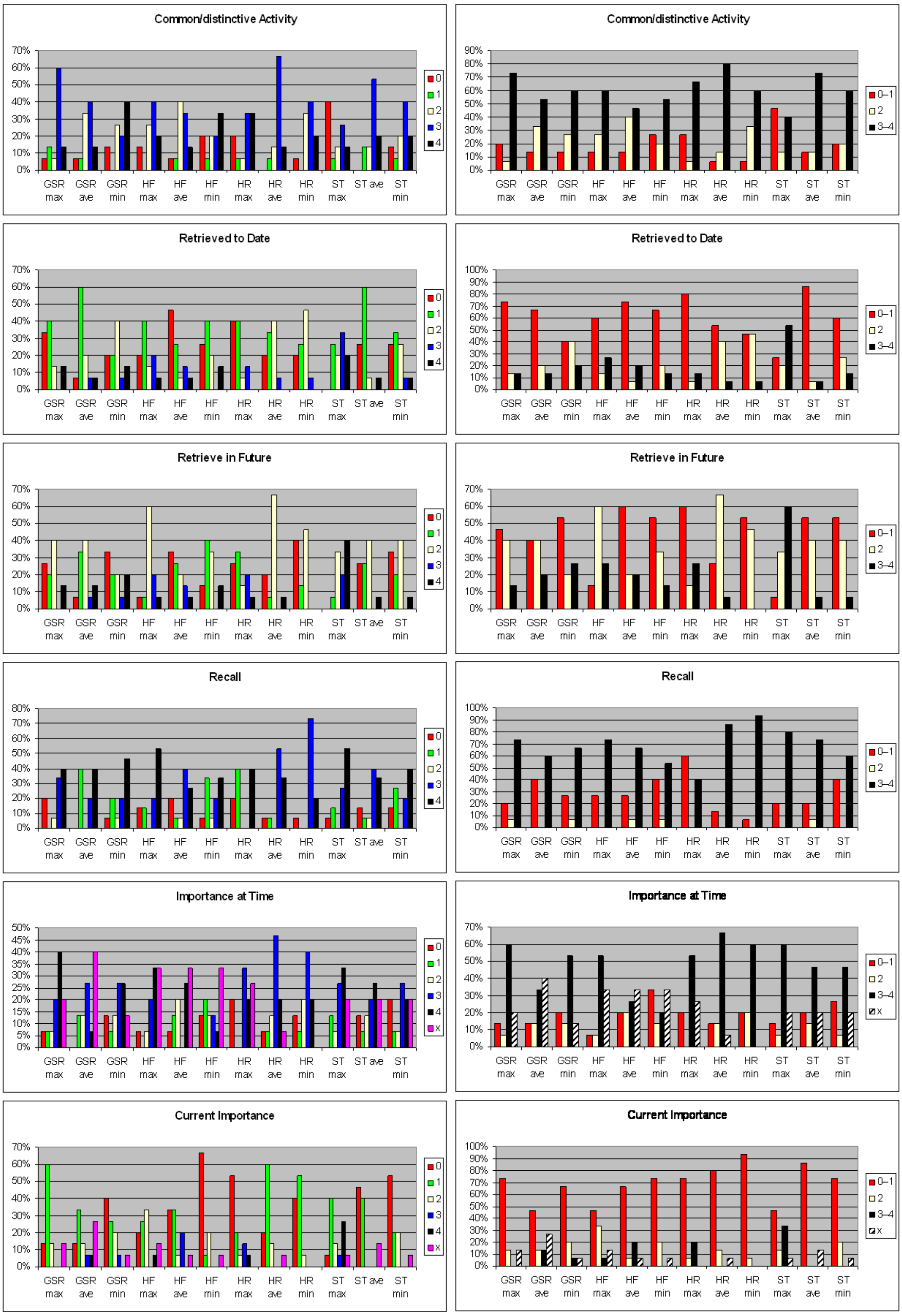


Figure 6.8: Questionnaire results - average computer item ratings for the three subjects combined for max, average (ave) and min GSR, HR, HF and ST using the *divEng* technique. Note the graphs on the left column present a break down of the results, while the right column graphs present a grouping of the questions 5-point scale.

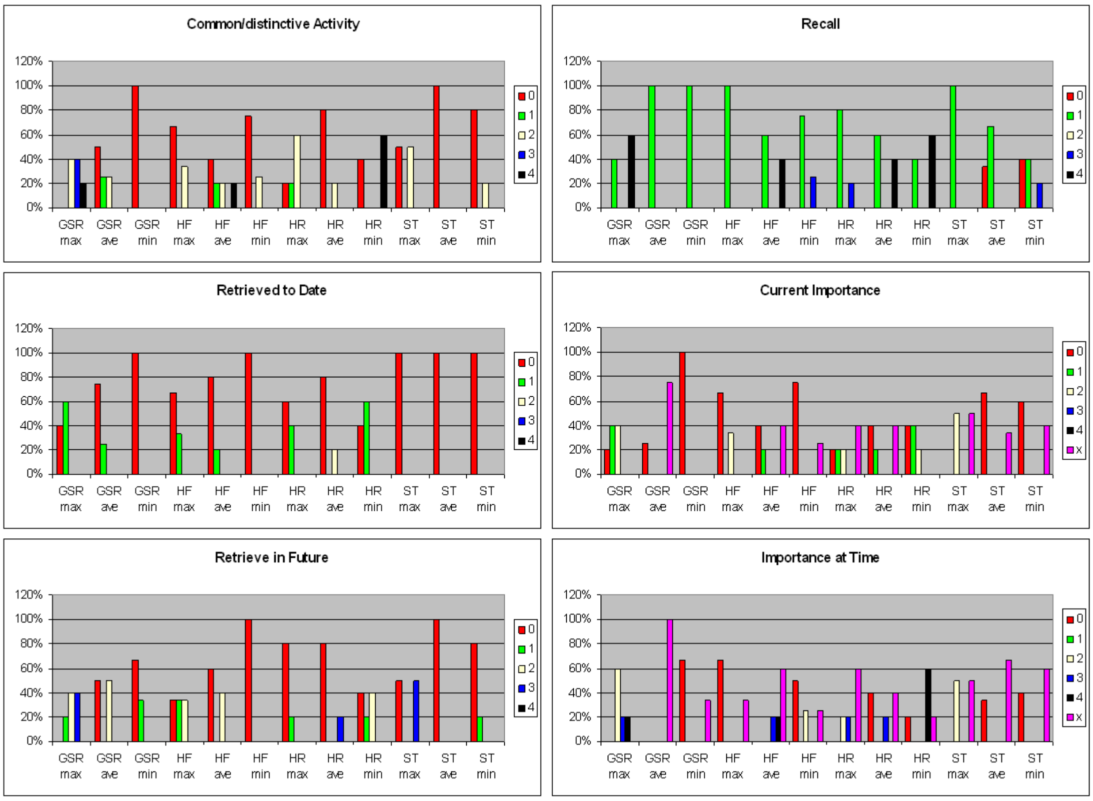


Figure 6.9: Questionnaire results - SenseCam event ratings for Subject 1 for max, average (ave) and min GSR, HR, HF and ST using the *removeEng* technique.

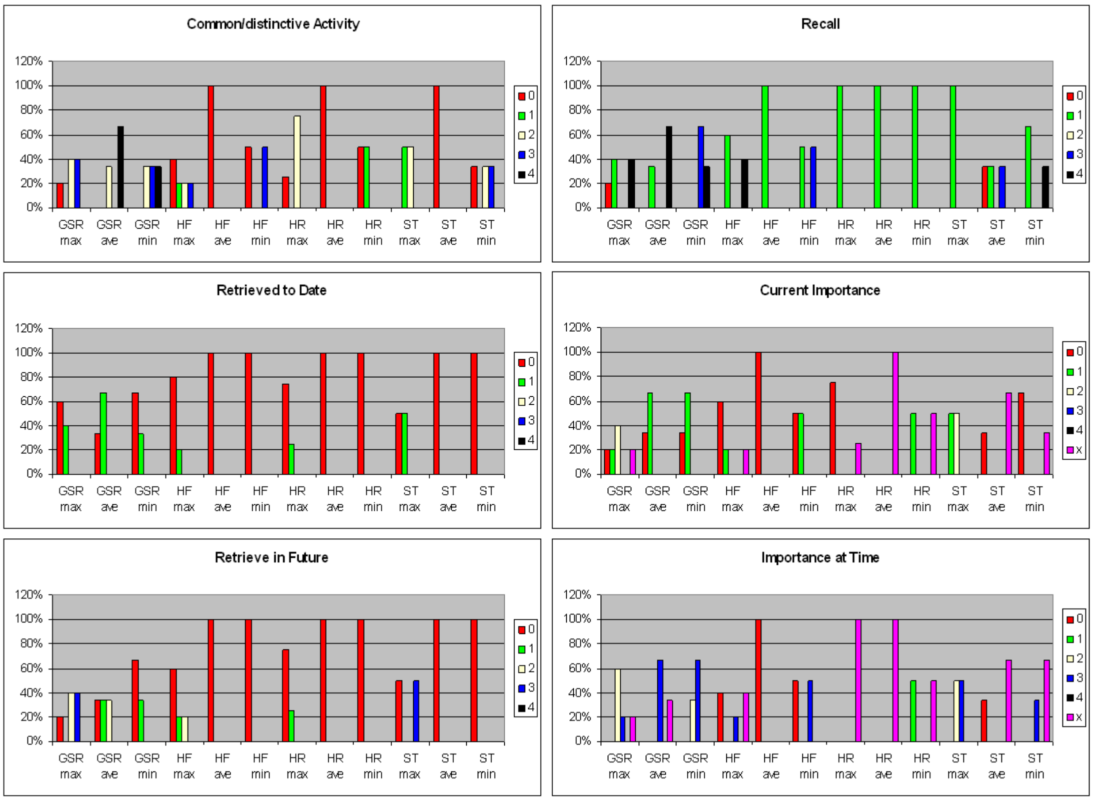


Figure 6.10: Questionnaire results - SenseCam event ratings for Subject 1 for max, average (ave) and min GSR, HR, HF and ST using the *divEng* technique.

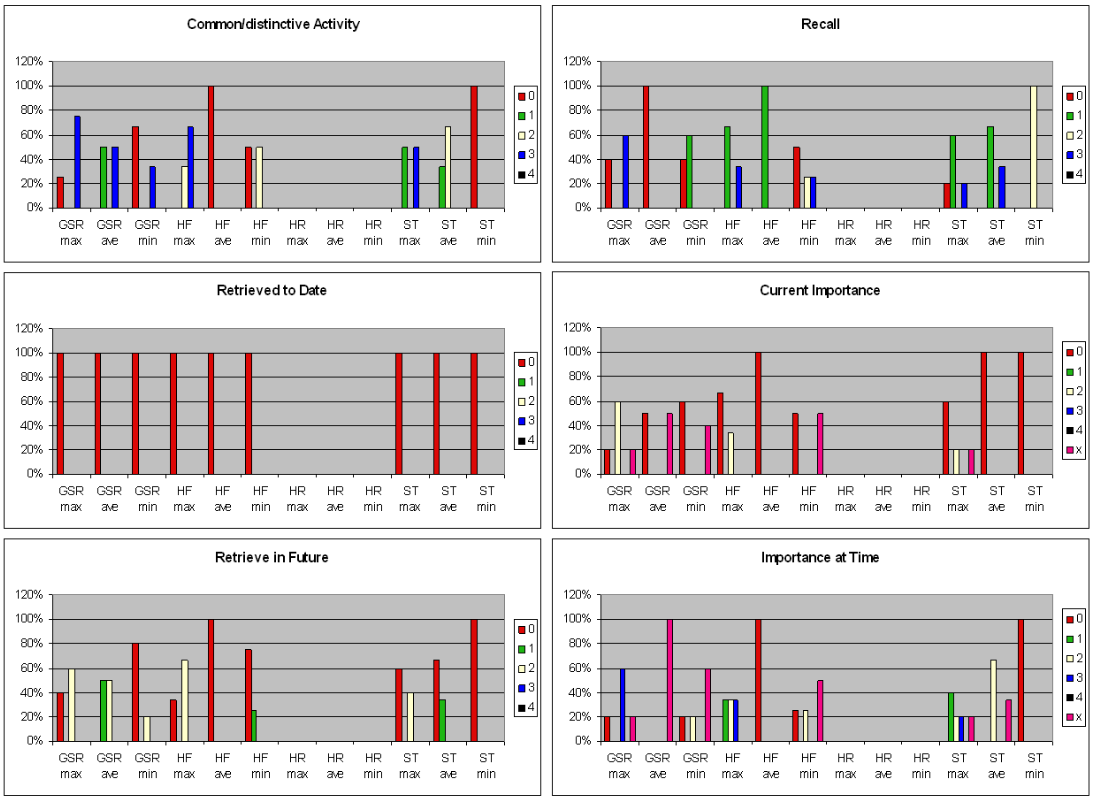


Figure 6.11: Questionnaire results - SenseCam event ratings for Subject 2 for max, average (ave) and min GSR, HR, HF and ST using the *removeEng* technique.

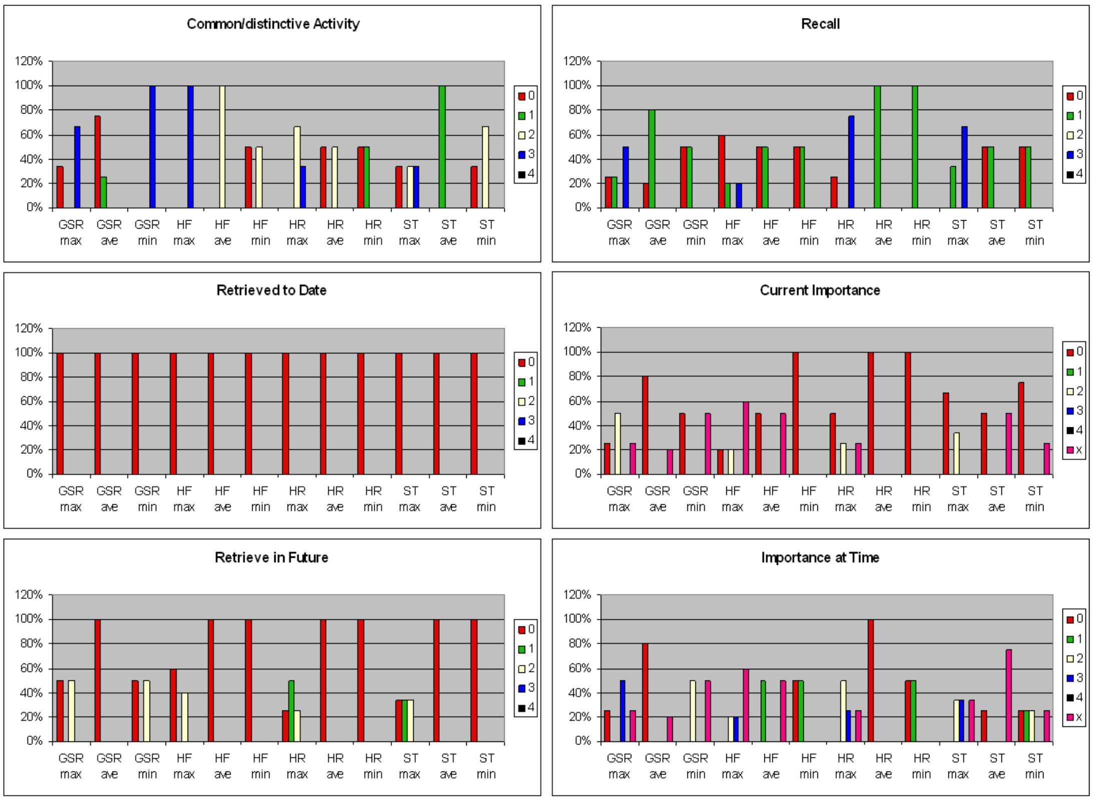


Figure 6.12: Questionnaire results - SenseCam event ratings for Subject 2 for max, average (ave) and min GSR, HR, HF and ST using the *divEng* technique.

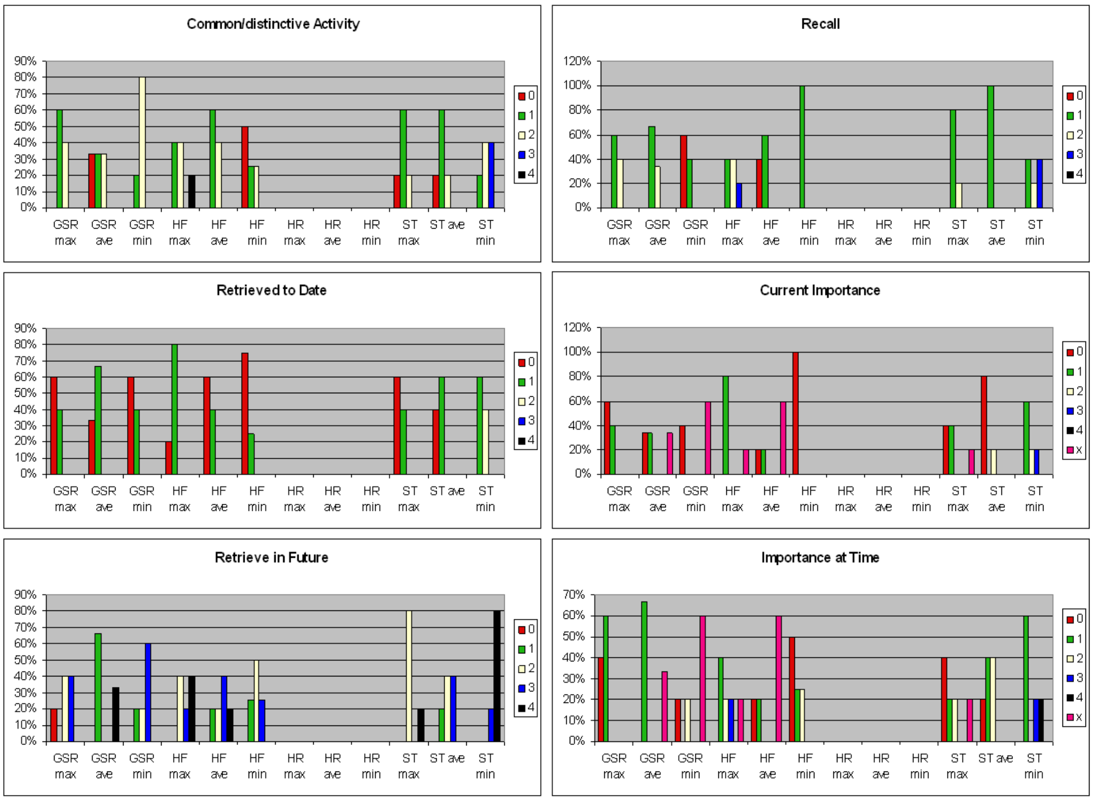


Figure 6.13: Questionnaire results - SenseCam event ratings for Subject 3 for max, average (ave) and min GSR, HR, HF and ST using the *removeEng* technique.

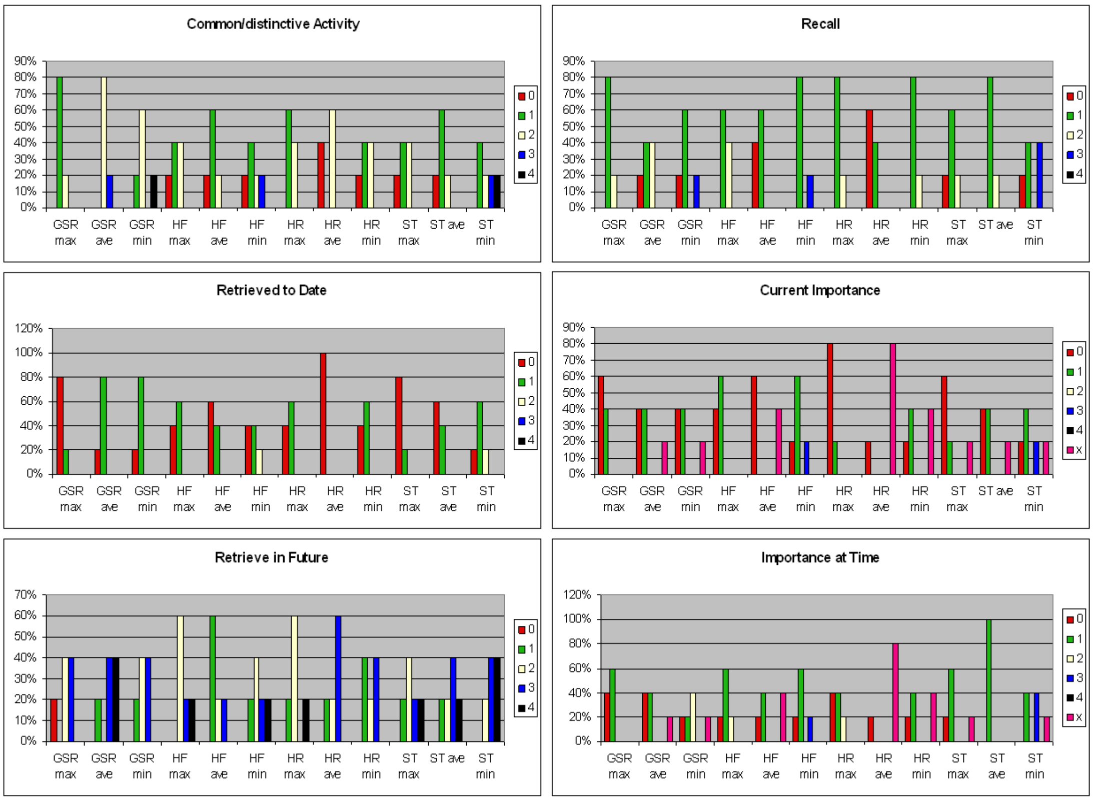
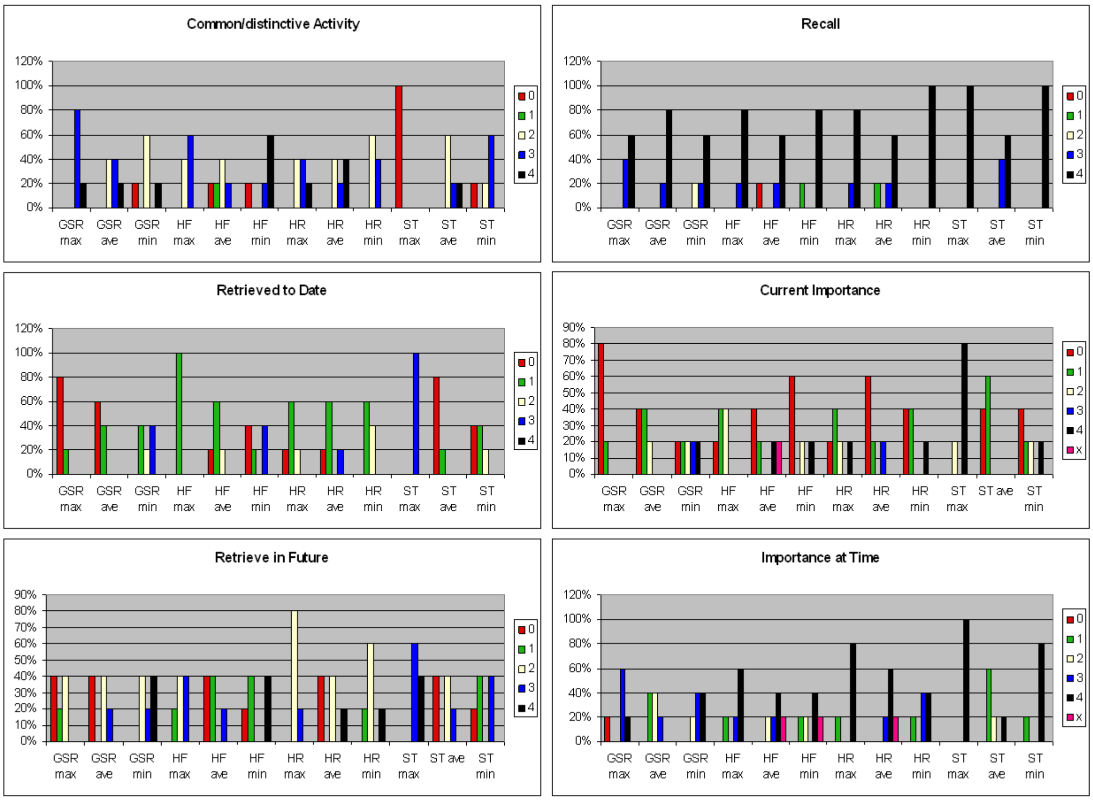


Figure 6.14: Questionnaire results - SenseCam event ratings for Subject 3 for max, average (ave) and min GSR, HR, HF and ST using the *divEng* technique.



Common/distinctive Activity

Recall

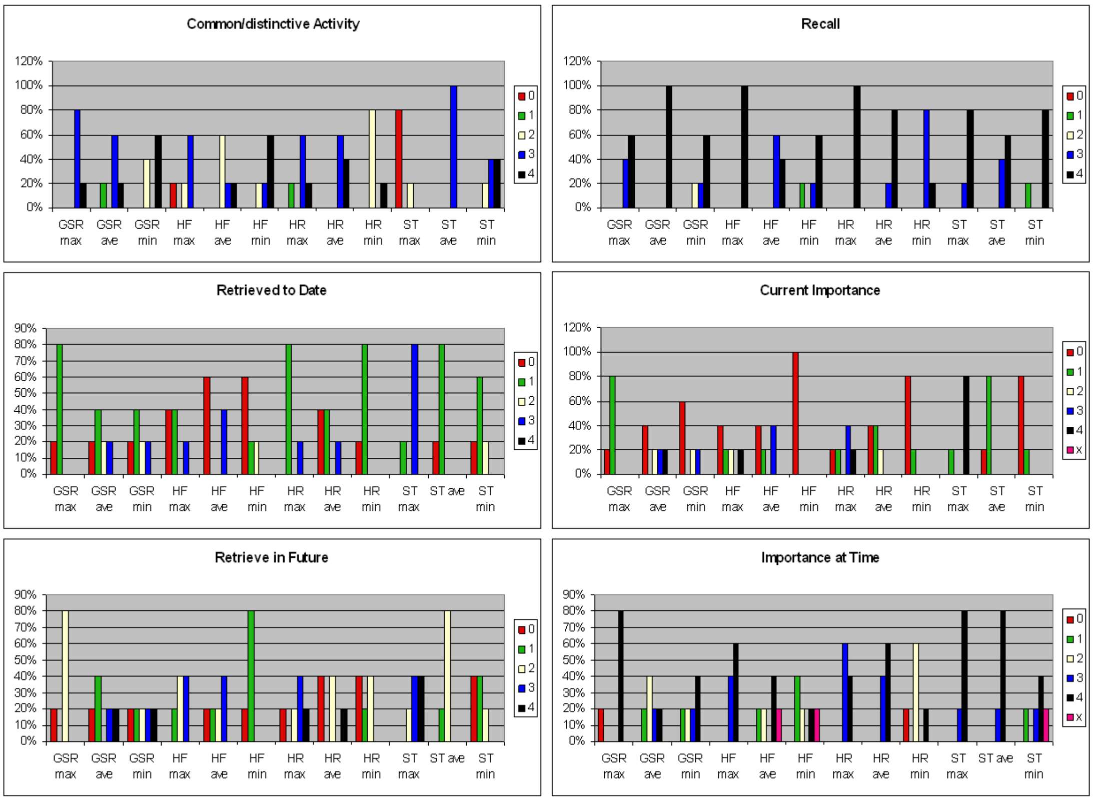
Retrieved to Date

Current Importance

Retrieve in Future

Importance at Time

Figure 6.15: Questionnaire results - computer item ratings for Subject 1 for max, average (ave) and min GSR, HR, HF and ST using the *removeEng* technique.



Common/distinctive Activity

Recall

Retrieved to Date

Current Importance

Retrieve in Future

Importance at Time

Figure 6.16: Questionnaire results - computer item ratings for Subject 1 for max, average (ave) and min GSR, HR, HF and ST using the *divEng* technique.

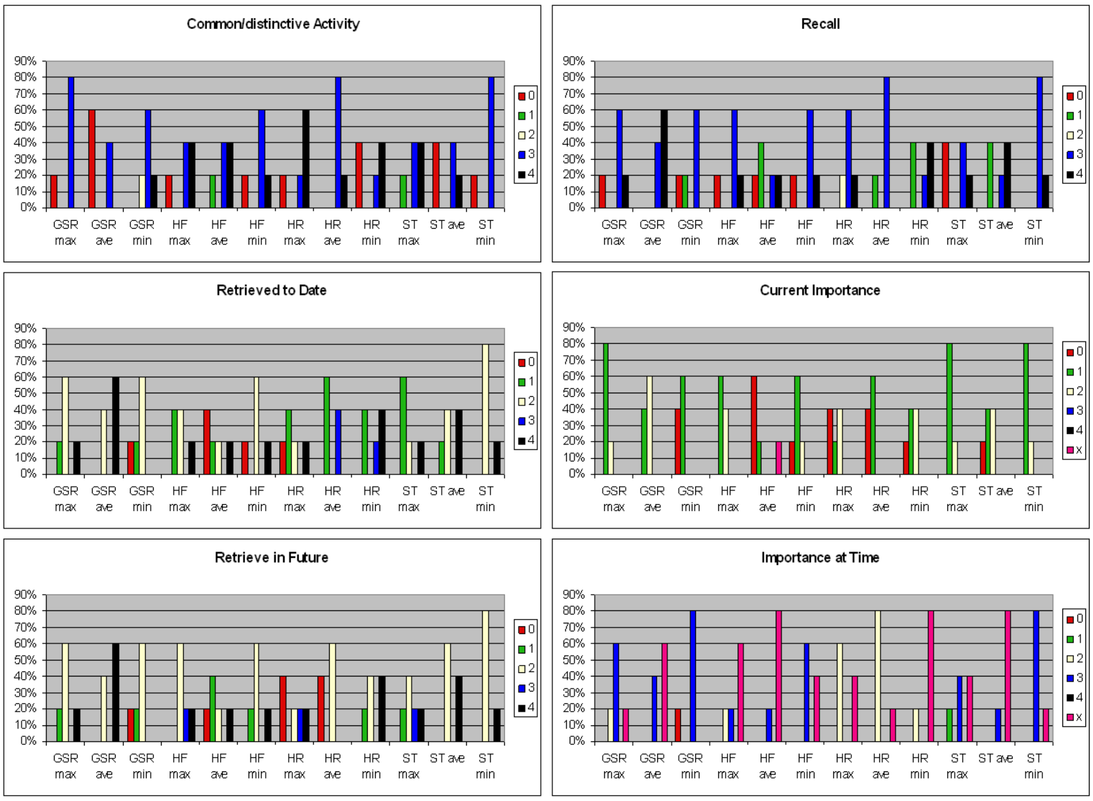


Figure 6.17: Questionnaire results - computer item ratings for Subject 2 for max, average (ave) and min GSR, HR, HF and ST using the *removeEng* technique.

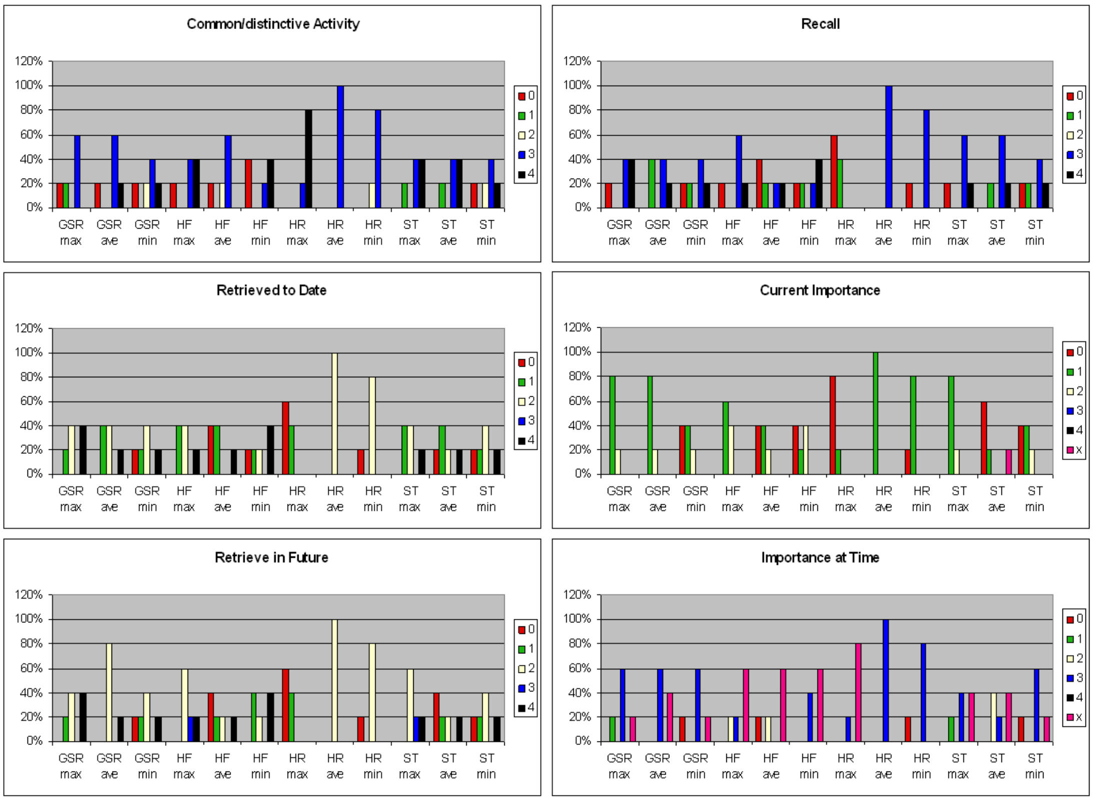


Figure 6.18: Questionnaire results - computer item ratings for Subject 2 for max, average (ave) and min GSR, HR, HF and ST using the *divEng* technique.

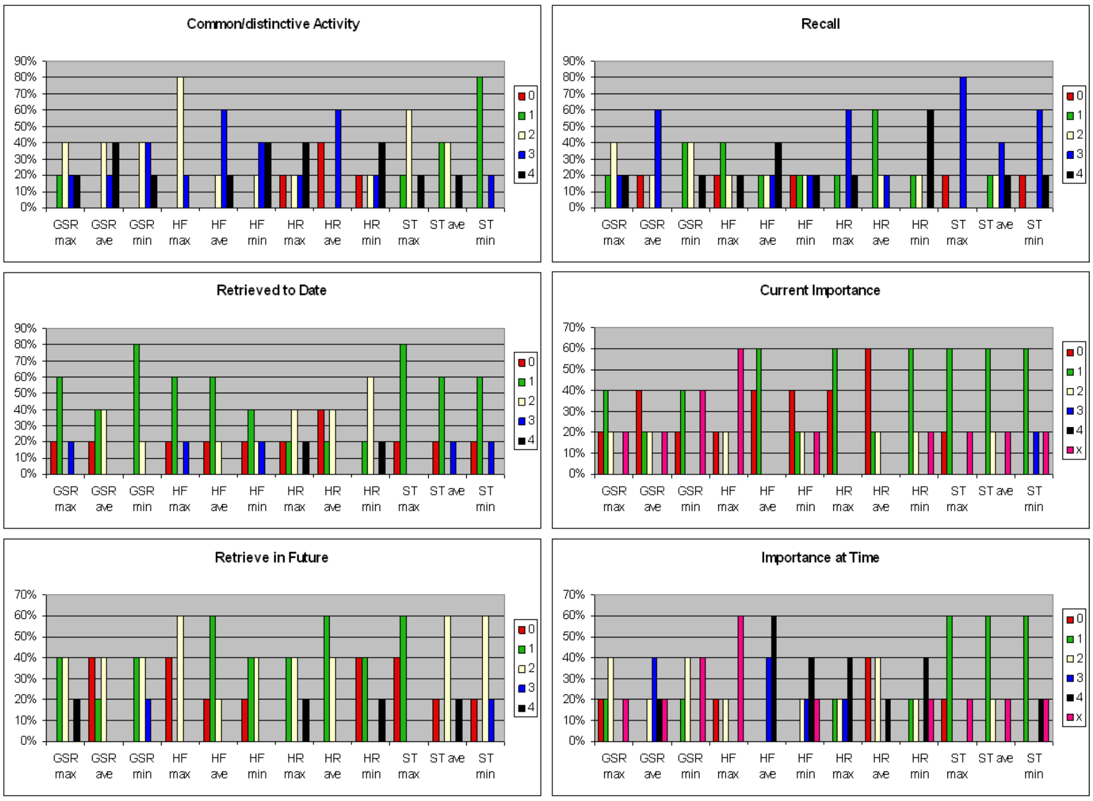


Figure 6.19: Questionnaire results - computer item ratings for Subject 3 for max, average (ave) and min GSR, HR, HF and ST using the *removeEng* technique.

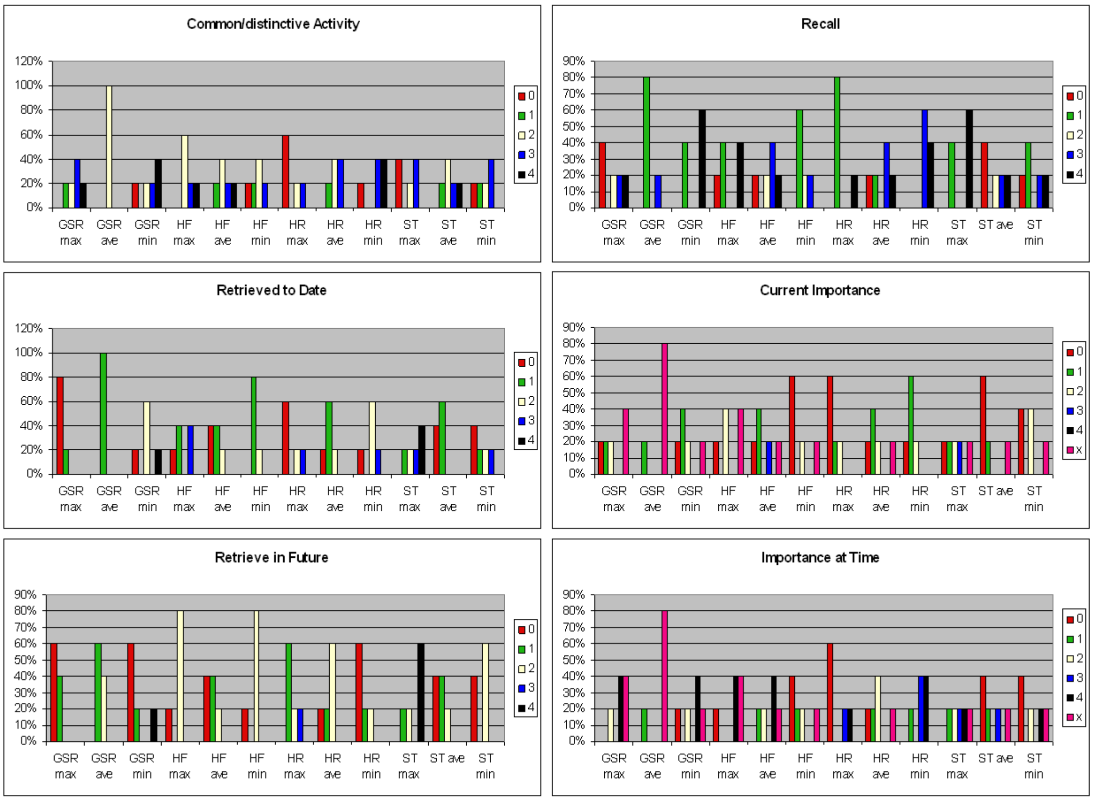


Figure 6.20: Questionnaire results - computer item ratings for Subject 3 for max, average (ave) and min GSR, HR, HF and ST using the *divEng* technique.

CHAPTER

SEVEN

Static Scores: Boosting Relevant Items in Result Lists using Past Biometric Response

Chapter Overview: Given the relationship observed in the previous chapter between biometric response at the time of experiencing lifelog items/events and the future importance of items to the individual, we wished to explore our hypothesis that use of biometric response as a static query factor boost in ranked content+context retrieval algorithms in the lifelogging domain would prove useful. In this chapter we investigate this hypothesis. Following the chapter introduction the setup of this investigation is described. We then provide a detailed analysis of the results of using Galvanic Skin Response, Heat Flux, Skin Temperature and Heart Rate as static scores in content+context-based retrieval algorithms for the lifelogging domain. This is followed by a discussion of the topic of this chapter, and concluding remarks.

7.1 Introduction

In Chapter 6 we showed that lifelog items which are important to an individual at the time they are experienced may be useful to the individual again in the future, and further that such incidents are associated with physiological responses when accessing these items that can be captured using digital sensors. Based on these findings we propose that recording biometric response as part of a lifelog may enable us to identify items which may be the most important in a future information searching task. We hypothesize that adding a query independent static score factor to items in lifelog IR result lists may improve ranked retrieval performance by promoting the rank of items with significant biometric responses.

This chapter describes our study to investigate the utility of biometric response in re-ranking traditional information retrieval result lists. Since BM25F_mod2, described in Chapter 5, was our overall best performing content+context ranked retrieval model, we use BM25F_mod2 to generate the ranked result lists. We explore the use of the following biometric measures as static scores for re-ranking these retrieved result lists: galvanic skin response (GSR), heat flux (HF), heart rate (HR) and skin temperature (ST).

As discussed in Chapter 4.2.3 various approaches can be used to explore integrating static scores into ranked retrieval algorithms. In this chapter we explore using a linear combination of the ranked retrieval and static biometric scores. We also investigate various approaches for transforming the biometric response into a static score. In particular raw biometric scores and various nonlinear transformations of the biometric readings are explored. A promising technique for score transformation is presented in [Craswell et al., 2005a] where a sigmoid functional form is used to transform PageRank, link indegree, ClickDistance and URL length features into static scores (discussed in Chapter 4.2.3). This technique forms part of our investigation.

The next section presents our experimental setup. Section 7.3 provides the results of our experiment along with detailed analysis. We then discuss some key points on the topic of this chapter before concluding the chapter with a discussion of findings.

7.2 Experiment Procedure

In this section we describe the setup of our study to examine the utility of GSR, HR, HF and ST biometric data at the time of previous item access in re-ranking the output of a ranked retrieval result list.

Our generated indexes of the textual data in the one month subset of our 3 subjects' 20 month lifelogs which coincided with the period of biometric data capture (i.e. the biometric month), described in Chapter 4.3.1, are used in this investigation[1]. The biometric data is obtained from the biometrics table in the subjects' lifelog databases, described in Chapter 3.2.

The queries and result sets used for this investigation were those contained in the subset of the 100 test cases generated for each subject which contained items occurring during the biometric capture month. Full details of these queries and result sets are provided in Chapter 4.3.2.3 and 4.3.2.4. SMSs and imported emails were also excluded from these biometric month test cases, in this study. Removing these items does not greatly alter the makeup of the biometric month test cases: 1 of Subject 2's biometric month queries was for SMS only; 2 of the relevant items for Subject 2's biometric queries and 27 of the relevant items for Subject 3's biometric queries were SMSs and imported emails. The makeup of the resulting test cases, including total number of relevant items across the queries in the subjects' biometric month tests cases and the average number of relevant items for these queries, are shown in Table 7.1.

Since content+context-based retrieval using BM25F_mod2 was the overall best performing ranked retrieval approach for our collections (described in Chapter 5), we used BM25F_mod2 to obtain the queried content+context retrieval scores. Static biometric scores were added to the BM25F_mod2 scores (techniques used to obtain static biometric scores are described in Section 7.2.1). The rank of the relevant items in the result sets were noted.

The remainder of our experiment procedure is the same as that used for the ranked retrieval investigations described in Chapter 5.2.

[1] Note: The SMSs and imported emails in these indexes, which do not have "date-time" of access information available, were excluded from this experiment (imported emails are described in Chapter 3.3.1.1).

	Subject 1	**Subject 2**	**Subject 3**
Test Cases	22	7	36
Rel Items	154	67	533
Ave	7	9.57	14.81

Table 7.1: Total number of test cases, total number of relevant items (Rel Items) per subject across the test case queries and average number of relevant items per query (Ave).

7.2.1 Static Relevance Scores

Recall that with increased arousal levels GSR, HR and HF levels increase and ST levels decrease (described in Chapter 2.3.1). Since we seek to gain an understanding of the importance of the item to the individual in the collection as a whole, in these initial experiments into the utility of biometric response associated with past experience of items as a static score, we intuit that the maximum biometric response observed for an item across all past accesses to the item will indicate the importance of the item in the collection (or the importance of the item to the user in the collection)[2]. Thus each retrieved item for content+context retrieval was annotated with the maximum observed GSR, maximum observed HR, maximum observed HF and minimum observed ST across all accesses to the item[3]

As described in Chapter 2.3.1, increases in physical activity (detected through increases in energy expenditure) cause GSR, HR and HF levels to increase and ST levels to decrease. To discern changes in GSR, HR, ST and HF caused by changes in arousal level as opposed to changes in physical activity, we also tagged items with the maximum observed GSR, HF and HR with energy expenditure factored[4] and with the minimum observed ST with energy expenditure factored[5] across all accessed to the item. To factor energy expenditure into the biometric readings we divided GSR, HF and HR levels by their associated energy expenditure readings (i.e., $\frac{GSR}{engGSR}$, $\frac{HR}{engHR}$ and $\frac{HF}{engHF}$) and we multiplied ST levels by their associated energy expenditure readings (i.e. $ST \cdot engST$). As explained in Chapter 2.3.1, the lower the ST level the greater the arousal level, hence items were also tagged with the inverse of ST and the inverse

[2] This is the same premise as we took in Chapter 6 to explore the relationship between computer item importance and biometric response associated with previous item interaction.

[3] We acknowledge that biometric response when accessing files outside the biometric month, had it been recorded, may have resulted in different biometric levels being assigned to items, and that lack of this information may have negatively impacted on the results we will present in this chapter.

[4] Energy expenditure readings associated with GSR, HF and HR are referred to as engGSR, engHF and engHR respectively.

[5] Energy expenditure readings associated with ST are referred to as engST.

Tag Type	Subject1	Subject2	Subject3
GSR,ST,HF	35%	67%	60%
HR	53%	83%	85%

Table 7.2: Percentage of retrieved items missing galvanic skin response (GSR), skin temperature (ST), heat flux (HF) and heart rate (HR) biometric tags, across each subject's queries.

Type	Subject1	Subject2	Subject3
stNorm	0.5325	0.6253	0.4792
stMultEngNorm	0.0238	0.02566	0.0373
inversStNorm	0.3936	0.3240	0.4674
inversStMultEngNorm	0.8324	0.8775	0.8450
gsrNorm	0.1369	0.2623	0.3044
gsrDivEngNorm	0.2654	0.2498	0.3485
hrNorm	0.2877	0.25	0.3418
hrDivEngNorm	0.3367	0.5180	0.5172
hfNorm	0.3067	0.3734	0.2309
hfDivEngNorm	0.5172	0.6148	0.6047

Table 7.3: Default normalised biometric tags assigned to items with missing biometric tags.

of $ST \cdot engST$.

Items in the 'biometric month' test set which had no associated biometric readings, due to biometric recording devices being removed for data downloading purposes, the subjects need for mental break from wearing of devices, and in the case of the heart rate monitor errors in recorded readings, etc (as described in Chapter 3.3.4.4), were assigned default biometric tags. The default value used was the median of the biometric tag associated with retrieved items. Examining the items retrieved for each subject for BM25F_mod2 retrieval reveals that 35% of the items retrieved for Subject 1 had no GSR, HF and ST tags and 53% had no HR tags, for Subject 2 67% had no GSR, HF and ST tags and 83% were missing HR tags, and for Subject 3 60% were missing GSR, HF and ST tags and 85% were missing HR tags. These results are shown in Table 7.2. Table 7.3 provides the normalised default tags assigned to items with missing biometric tags for each subject.

All biometric tags associated with retrieved items were normalised using min-max normalisation. For example, to normalise each GSR tag associated with retrieved items we use: *(GSRtag - minGSRtag)/(maxGSRtag - minGSRtag)*. The following approaches for calculating static relevance scores using the normalised biometric data tags were investigated:

$$BIObase = w \cdot se \tag{7.1}$$

$$logBIO = w \cdot log(s) \tag{7.2}$$

$$logBIOeng = w \cdot log(se) \tag{7.3}$$

$$sigmBIO = w \cdot \frac{s^a}{k^a + s^a} \tag{7.4}$$

$$sigmBIOeng = w \cdot \frac{se^a}{k^a + se^a} \tag{7.5}$$

$$sigmIncST = w \cdot \frac{k^a}{k^a + st^a},\ where\ st = ST \tag{7.6}$$

$$sigmIncSTeng = w \cdot \frac{k^a}{k^a + st^a},\ where\ st = ST \times engST \tag{7.7}$$

In the above equations $s = \frac{1}{ST}$, *GSR*, *HR* or *HF* and $se = \frac{1}{ST \times engST}$ (i.e., $\frac{\frac{1}{ST}}{engST}$), $\frac{GSR}{engGSR}$, $\frac{HR}{engHR}$ or $\frac{HF}{engHF}$. For the remainder of this chapter we use *STbase* to refer to the the use of ST data in the *BIObase* equation, *logGSR* to refer to the use of GSR data in the *logBIO* equation, etc. Following parameter tuning using the full set of the 3 subjects' biometric month test cases, the static score's weight of importance (*w*) and parameters *k* and *a* (where applicable) were set for each equation. Our parameter tuning approach was the same as that used for tuning retrieval algorithms parameters, described in Chapter 5.2. That is, for each static scoring approach we manually tuned the weight (*w*) to give overall best retrieval performance, and where applicable the *k* and *a* parameters were set to 1 during this process. For the static scoring approaches which contained *k* and *a* parameters, we then manually tuned the *k* parameter to give overall best retrieval performance using the tuned weight (*w*) and leaving the *a* parameter set to 1. Finally using the tuned weight (*w*) and *k* parameter we manually tuned the *a* parameter to give overall best retrieval performance. Table 7.4 provides these tuned parameter

	w	**k**	**a**
STbase	0.04	-	-
logST	0.03	-	-
logSTeng	0.05	-	-
sigmST	1	9	3
sigmSTeng	1	0.1	0.2
sigmIncST	0.001	1	0.001
sigmIncSTeng	0.4	0.9	0.1
GSRbase	0.02	-	-
logGSR	0.002	-	-
logGSReng	0.003	-	-
sigmGSR	0.9	8	2
sigmGSReng	0.5	7	1.2
HFbase	0.04	-	-
logHF	0.09	-	-
logHFeng	0.03	-	-
sigmHF	0.8	1.2	1.1
sigmHFeng	0.9	0.2	0.1
HRbase	0.001	-	-
logHR	0.0001	-	-
logHReng	0.0001	-	-
sigmHR	2	8	3.5
sigmHReng	0.1	1	0.001

Table 7.4: Parameter tunings for static biometric functions.

values.

Equation 1 is our baseline static scoring approach, used to examine the effect of the raw ST, HR, HF and GSR values with energy expenditure factored in on re-ranking result lists. The remaining equations investigate the use of non-linear transformations of the biometric score. Equations 2 and 3 examine the effect of using logs of ST, HR, HF and GSR. The performance of our biometric scores using the transformation approach from [Craswell et al., 2005a] described in Chapter 4.2.3 is examined with Equations 4 and 5. This approach is used to generate static relevance scores for features where higher values indicate greater importance. An approach for calculating static relevance scores for features where lower values indicate greater importance is also provided in [Craswell et al., 2005a] and described in Chapter 4.2.3. This technique's performance using our ST data is investigated with Equations 6 and 7. The effect of accounting for energy expenditure is investigated in Equations 1, 3, 5 and 7.

The static scoring techniques presented in this section are added to content+context relevance scores generated using the BM25F_mod2 model, described in Chapter 5.5. The next section discusses results obtained using these approaches.

7.3 Results and Analysis

Retrieval effectiveness is measured here using average precision (AveP), P@5 and P@10. P@5 and P@10 show how effective our techniques were at moving relevant items towards the top of the result lists. Table 7.5 shows the retrieval scores for BM25F_mod2+static_score retrieval, averaged across the three experiment subjects, and percentage improvement over the BM25F_mod2 baseline that these scores correspond to. Table 7.6 presents the individual breakdown of results for each subject and Table 7.7 provides the percentage improvement over the BM25F_mod2 baseline that these scores correspond to. Table 7.9 presents results obtained for each subject when we consider only items with 'real' biometric tags (i.e. not considering items assigned default biometric tags) in retrieval. Table 7.10 presents the percentage improvement over the BM25F_mod2 baseline that these scores correspond to. The results presented in these tables suggest that adding biometric static scores to content+context IR scores is useful for re-ranking PL text-based collections. In this section we analyse these results.

7.3.1 Overall Static Score Performance

The total number of relevant items retrieved across all queries using BM25F_mod2 on the biometric month collections was: for Subject 1 144 items; for Subject 2 67 items; and for Subject 3 479 items.

Considering BM25F_mod2 content+context retrieval the addition of a static score using HF with energy expenditure levels factored resulted in the greatest percentage improvement over the content+context baseline. The three techniques which factor energy expenditure, i.e., *HFbase*, *logHFeng* and *sigmHFeng*, yielded 1%, 3% and 3% improvement in AveP, P@5 and P@10 respectively over the content+context baseline. Overall, when energy expenditure was not factored into the HF levels negative performance was obtained from the use of HF as a static score. Exceptions here are the AveP and P@10 results obtained using *logHF*, for which marginal improvement was noted, see Table 7.5.

Similar to the use of HF as a static score, factoring of energy expenditure was generally important when calculating static scores using ST, with 1% improvement in P@10

Static Technique	**Average Score**		
	AveP	**P@5**	**P@10**
BM2F_mod2	*0.382*	*0.290*	*0.229*
STbase	0.384 (0%)	0.290 (0%)	0.232 (1%)
logST	0.385 (1%)	0.289 (0%)	0.228 (0%)
logSTeng	0.383 (0%)	0.287 (-1%)	0.232 (1%)
sigmST	0.382 (0%)	0.287 (-1%)	0.229 (0%)
sigmSTeng	0.383 (0%)	0.287 (-1%)	0.232 (1%)
sigmIncST	0.382 (0%)	0.287 (-1%)	0.229 (0%)
sigmIncSTeng	0.383 (0%)	0.290 (0%)	0.232 (1%)
GSRbase	0.384 (0%)	0.294 (2%)	0.231 (1%)
logGSR	0.384 (0%)	0.294 (2%)	0.232 (1%)
logGSReng	0.384 (0%)	0.294 (2%)	0.231 (1%)
sigmGSR	0.384 (0%)	0.294 (2%)	0.232 (1%)
sigmGSReng	0.385 (1%)	0.294 (2%)	0.229 (0%)
HFbase	0.386 (1%)	0.299 (3%)	0.236 (3%)
logHF	0.384 (1%)	0.281 (-3%)	0.230 (0%)
logHFeng	0.387 (1%)	0.299 (3%)	0.236 (3%)
sigmHF	0.377 (-2%)	0.279 (-4%)	0.226 (-1%)
sigmHFeng	0.386 (1%)	0.299 (3%)	0.236 (3%)
HRbase	0.384 (0%)	0.292 (1%)	0.231 (1%)
logHR	0.385 (1%)	0.296 (2%)	0.232 (1%)
logHReng	0.384 (0%)	0.292 (1%)	0.231 (1%)
sigmHR	0.385 (1%)	0.296 (2%)	0.232 (1%)
sigmHReng	0.384 (0%)	0.292 (1%)	0.231 (1%)

Table 7.5: Average score and average percentage improvement, rounded to nearest whole number, for average precision (AveP), P@5 and P@10, by adding a static score to the BM25F_mod2 baseline.

observed using *STbase*, *logSTeng*, *sigmSTeng* and *sigmIncSTeng*. However, in the case of *logSTeng* and *sigmSTeng* 1% reduction in P@5 was noted. The only improvement noted when energy expenditure was not factored was 1% improvement for AveP using *logST*. These results are shown in Table 7.5.

Adding HR static scores also improved retrieval performance. More so than the use of ST although not quite to the same extent as the addition of HF static scores. Using HR greatest improvement in performance was obtained when energy expenditure was not factored. Specifically, using *logHR* and *sigmHR* yielded 1%, 2% and 1% improvement in performance for AveP, P@5 and P@10 respectively. The use of energy expenditure with HR to calculate static scores also showed utility, albeit not to the same extent as. These results are shown in Table 7.5.

Averaged across the 3 subjects all GSR static scoring techniques showed utility in reranking the content+context retrieval result lists. As shown in Table 7.5, 2% improvement in P@5 was noted for each static scoring technique using GSR, 1% improvement in P@10 for all techniques using GSR except *sigmGSReng*, and 1% improvement in

Static Technique	**Subject 1**			**Subject 2**			**Subject 3**		
	AveP	**P@5**	**P@10**	**AveP**	**P@5**	**P@10**	**AveP**	**P@5**	**P@10**
BM2F_mod2	*0.515*	*0.360*	*0.273*	*0.229*	*0.114*	*0.057*	*0.403*	*0.395*	*0.357*
STbase	0.519	0.360	0.280	0.230	0.114	0.057	0.404	0.395	0.360
logST	0.519	0.353	0.270	0.230	0.114	0.057	0.405	0.400	0.357
logSTeng	0.517	0.353	0.280	0.230	0.114	0.057	0.404	0.395	0.360
sigmST	0.514	0.353	0.273	0.229	0.114	0.057	0.404	0.395	0.357
sigmSTeng	0.517	0.353	0.280	0.230	0.114	0.057	0.404	0.395	0.360
sigmIncST	0.515	0.353	0.273	0.229	0.114	0.057	0.403	0.395	0.357
sigmIncSTeng	0.517	0.360	0.280	0.229	0.114	0.057	0.404	0.395	0.360
GSRbase	0.521	0.373	0.280	0.229	0.114	0.057	0.403	0.395	0.357
logGSR	0.521	0.373	0.283	0.227	0.114	0.057	0.403	0.395	0.357
logGSReng	0.521	0.373	0.280	0.229	0.114	0.057	0.403	0.395	0.357
sigmGSR	0.521	0.373	0.283	0.227	0.114	0.057	0.403	0.395	0.357
sigmGSReng	0.520	0.373	0.277	0.229	0.114	0.057	0.405	0.395	0.354
HFbase	0.524	0.387	0.293	0.231	0.114	0.057	0.404	0.395	0.357
logHF	0.510	0.340	0.273	0.235	0.114	0.057	0.408	0.389	0.360
logHFeng	0.524	0.387	0.293	0.231	0.114	0.057	0.406	0.395	0.357
sigmHF	0.501	0.327	0.267	0.228	0.114	0.057	0.401	0.395	0.354
sigmHFeng	0.524	0.387	0.293	0.231	0.114	0.057	0.404	0.395	0.357
HRbase	0.520	0.367	0.280	0.228	0.114	0.057	0.403	0.395	0.357
logHR	0.523	0.380	0.283	0.228	0.114	0.057	0.403	0.395	0.357
logHReng	0.520	0.367	0.280	0.228	0.114	0.057	0.403	0.395	0.357
sigmHR	0.523	0.380	0.283	0.228	0.114	0.057	0.403	0.395	0.357
sigmHReng	0.520	0.367	0.280	0.228	0.114	0.057	0.403	0.395	0.357

Table 7.6: Individual subjects scores for average precision (AveP), P@5 and P@10 by adding a static score to the BM25F_mod2 baseline.

AveP for *sigmGSReng*.

7.3.2 Performance Across Individual Subjects

Overview:

Analysing the individual results of each subject in Tables 7.6 and 7.7 we see that overall for Subject 1 HF with energy expenditure factored in provided the greatest utility in re-ranking the ranked retrieval result lists (see Subject 1's results for *HFbase*, *logHFeng* and *sigmHFeng* in these tables). Indeed as we look at the three subjects results in these tables we see that it was Subject 1's results for HF which influenced the superior utility observed for HF with energy expenditure considered in the average results reported in the previous section. For Subjects 2 and 3 greatest utility was also observed using HF. However, in these subjects' cases, greatest utility was observed when energy expenditure was not factored into the HF readings, and in particular through the use of *logHF* (see Subjects 2 and 3's results for HF in Tables 7.6 and 7.7).

Static Technique	**Subject 1**			**Subject 2**			**Subject 3**		
	AveP	**P@5**	**P@10**	**AveP**	**P@5**	**P@10**	**AveP**	**P@5**	**P@10**
BM2F_mod2	*0.515*	*0.360*	*0.273*	*0.229*	*0.114*	*0.057*	*0.403*	*0.395*	*0.357*
STbase	1%	0%	2%	0%	0%	0%	0%	0%	1%
logST	1%	-2%	-1%	1%	0%	0%	0%	1%	0%
logSTeng	0%	-2%	2%	0%	0%	0%	0%	0%	1%
sigmST	0%	-2%	0%	0%	0%	0%	0%	0%	0%
sigmSTeng	0%	-2%	2%	0%	0%	0%	0%	0%	1%
sigmIncST	0%	-2%	0%	0%	0%	0%	0%	0%	0%
sigmIncSTeng	0%	0%	2%	0%	0%	0%	0%	0%	1%
GSRbase	1%	4%	2%	0%	0%	0%	0%	0%	0%
logGSR	1%	4%	4%	-1%	0%	0%	0%	0%	0%
logGSReng	1%	4%	2%	0%	0%	0%	0%	0%	0%
sigmGSR	1%	4%	4%	-1%	0%	0%	0%	0%	0%
sigmGSReng	1%	4%	1%	0%	0%	0%	0%	0%	-1%
HFbase	2%	7%	7%	1%	0%	0%	0%	0%	0%
logHF	-1%	-6%	0%	3%	0%	0%	1%	-1%	1%
logHFeng	2%	7%	7%	1%	0%	0%	1%	0%	0%
sigmHF	-3%	-9%	-2%	0%	0%	0%	-1%	0%	-1%
sigmHFeng	2%	7%	7%	1%	0%	0%	0%	0%	0%
HRbase	1%	2%	2%	0%	0%	0%	0%	0%	0%
logHR	2%	6%	4%	0%	0%	0%	0%	0%	0%
logHReng	1%	2%	2%	0%	0%	0%	0%	0%	0%
sigmHR	2%	6%	4%	0%	0%	0%	0%	0%	0%
sigmHReng	1%	2%	2%	0%	0%	0%	0%	0%	0%

Table 7.7: Subjects' percentage improvement, rounded to nearest whole number, for average precision (AveP), P@5 and P@10 by adding a static score to the BM25F_mod2 baseline.

In the remainder of this section we analyse each subjects results in greater detail.

Subject 1:

Across the three subjects, Subject 1 benefited by far the most from the use of static scores in re-ranking content+context IR result lists. With each biometric measure (i.e., ST, GSR, HF and HR) showing utility as a static biometric score. Some form of improvement over BM25F_mod2 was observed using all techniques with the exception of *sigmST*, *sigmIncST*, *logHF* and *sigmHF*. The superior results observed for Subject 1 relative to the other two subjects is unsurprising given the much higher number of items assigned recorded biometric tags in this subjects result set (see Table 7.2).

For this subject, ST proved least useful in re-ranking the result lists relative to the other biometric measures, while HF proved most useful. Factoring energy expenditure into the biometric readings proved useful for HF, but not for HR and GSR. Use of energy was useful when using a sigmoid functional form to calculate the static score for ST.

Subject 2:

Only HF and ST proved useful as static scores in re-ranking content+context IR result lists for Subject 2. Of the two HF proved most useful. With all HF techniques except *sigmHF* yielding improvement in AveP. Greatest improvement here was noted when energy expenditure was not considered using *logHF*. The only improvement in performance noted through the use of ST as a static score was in AveP using the log approach which did not consider energy expenditure (*logST*).

Subject 3:

Similar to Subject 2, only HF and ST proved useful as static scores in re-ranking content+context IR result lists for Subject 3. Of the two biometric types, ST arguably proved most useful as a static score for this subject. All techniques which factored energy expenditure into the ST readings improved P@10. The logST technique (a technique which does not consider energy expenditure) yielded improvement in AveP. Using HF, only the log techniques were useful.

We next compare the utility of each of HF, ST, GSR and HR as static scores across the three experiment subjects.

HF:

As already noted Subjects 1 and 2 benefited the most from use of HF static scores. Indeed in the case of Subject 1 substantial improvement over the BM25F_mod2 baseline was noted. This is a promising result given that BM25F_mod2 itself out performed state of the art techniques, as shown in Chapter 5. For this subject each of the techniques which factored energy expenditure into the HF readings (i.e., *HFbase*, *logHFeng* and *sigmHFeng*) yielded a 2%, 7% and 7% improvement in AveP, P@5 and P@10 respectively.

Subject 2 did not benefit to the same extent as Subject 1 using HF as a static score. Similar to Subject 1 use of energy expenditure with the HF readings proved useful, with 1% improvement in AveP being observed using each of *HFbase*, *logHFeng* and *sigmHFeng*, and 3% improvement being observed using *logHF* (a technique which does not factor energy expenditure into the HF levels).

Subject 3 benefited to a certain extent using HF. However, while for the other two subjects improvement was noted using all the techniques which considered energy expenditure, only *logHFeng* yielded improvement for Subject 3 (1% improvement in AveP). Not considering energy expenditure, *logHF* resulted in 1% improvement in AveP and P@10.

ST:

Use of ST as a static score was arguably more beneficial for Subject 3 than use of HF. All ST techniques which considered energy expenditure (*STbase*, *logSTeng*, *sigmSTeng* and *sigmIncSTeng*) helped move relevant items upwards in the result list, with 1% improvement in P@10 being observed for this subject. Improvement was also observed using the *logST* technique, where 1% improvement in P@5 was obtained.

Use of ST as a static score also proved useful for Subjects 1 and 2, however not to the same extent as the best performing HF static scoring approaches. Best performance here was observed for Subject 1 using *STbase* with a 1% improvement in AveP and a 2% improvement in P@10. For Subject 2 using *logST* resulted in a 1% improvement in AveP.

GSR:

GSR static scores did not prove as useful for Subject 2 and Subject 3 as ST and HF static scores, with no improvement over BM25F_mod2 being observed. However, in the case of Subject 1 improvement in AveP, P@5 and P@10 was noted across all GSR static scoring functions. Interestingly for this subject, while HF readings required consideration of energy expenditure to provide utility, greatest improvement in performance using GSR was noted using the two techniques which do not consider energy expenditure, namely *logGSR* and *sigmGSR*. Using these two techniques 1%, 4% and 4% improvement in AveP, P@5 and P@10 was achieved.

HR:

HR static scores also proved useful for Subject 1. Similar to use of GSR, all techniques proved useful with HR in improving retrieval performance. As for GSR, the two techniques which did not consider energy expenditure proved most useful. Using these techniques (i.e. *logHR* and *sigmHR*) 2%, 6% and 4% improvement in AveP, P@5 and P@10 respectively were obtained. Which was better than the results observed using GSR. Although not as good as the results, shown earlier, obtained using HF with energy expenditure considered. In contrast to the positive results obtained using HR for Subject 1, no benefit was obtained using HR as a static score for the other subjects.

	Subject 1	Subject 2	Subject 3
Rel Items	154	67	533
Rel Retrieved [full]	144	67	479
Rel Retrieved [reduced] (for GSR, ST, HF)	99	36	226
Rel Retrieved [reduced] (for HR)	78	15	56

Table 7.8: Total number of relevant items (Rel Items) per subject across the biometric month test case queries and total number of relevant items retrieved using the full and reduced result lists. Note the reduced result lists only contain items which were tagged with recorded biometric response.

This is unsurprising given the exceptionally high number of retrieved items which were missing HR tags for Subjects 2 and 3, as was shown in Table 7.2.

7.3.3 Performance of Biometric Tagged Result Set

Given the percentage of retrieved items which were missing biometric tags and hence assigned default biometric tags, we wished to establish the impact of using default biometric tags on static scoring performance.

In this experiment we investigate performance of our static scoring functions on the items in the subset of each subject's result list which were assigned biometric tags corresponding to the subject's max observed biometric response when experiencing the item. That is, items which had been assigned default biometric tags are removed from the result list. Since less items were tagged with HR than GSR, ST and HF (discussed in Section 7.2.1) this will result in the reduced HR result list being smaller than the result list generated for items with GSR, HF and ST tagged items. The total number of relevant items retrieved across all queries using the GSR, HF and ST reduced result lists was: for Subject 1 99 items; for Subject 2 36 items; and for Subject 3 226 items. The total number of relevant items retrieved across all queries using the HR reduced result lists was: for Subject 1 78 items; for Subject 2 15 items; and for Subject 3 56 items. For ease of comparison we show the statistics for these reduced result lists along side the full result lists in Table 7.8. Tables 7.9 and 7.10 present the results of this experiment.

Results Overview:

Greater improvement over the BM25F_mod2 baseline was observed in this experiment; particularly in the case of Subjects 2 and 3. This is unsurprising given the higher percentage of items assigned default biometric tags for these two subjects, as shown

Static Technique	**Subject 1**			**Subject 2**			**Subject 3**		
	AveP	**P@5**	**P@10**	**AveP**	**P@5**	**P@10**	**AveP**	**P@5**	**P@10**
BM25F_mod2 (for GSR,) (HF,ST tags)	*0.304*	*0.267*	*0.203*	*0.117*	*0.029*	*0.114*	*0.213*	*0.281*	*0.241*
STbase	0.310	0.267	0.207	0.115	0.057	0.114	0.214	0.281	0.246
logST	0.303	0.260	0.197	0.118	0.057	0.129	0.214	0.287	0.249
logSTeng	0.308	0.267	0.207	0.116	0.057	0.114	0.214	0.281	0.249
sigmST	0.302	0.260	0.197	0.115	0.029	0.114	0.214	0.281	0.241
sigmSTeng	0.308	0.267	0.207	0.116	0.057	0.114	0.214	0.281	0.249
sigmIncST	0.303	0.260	0.197	0.116	0.029	0.114	0.213	0.281	0.241
sigmIncSTeng	0.310	0.267	0.207	0.115	0.057	0.114	0.214	0.281	0.249
GSRbase	0.313	0.280	0.203	0.115	0.057	0.114	0.213	0.281	0.241
logGSR	0.313	0.280	0.207	0.115	0.029	0.114	0.213	0.281	0.241
logGSReng	0.313	0.280	0.203	0.115	0.057	0.114	0.213	0.281	0.241
sigmGSR	0.313	0.280	0.207	0.115	0.057	0.114	0.213	0.281	0.241
sigmGSReng	0.312	0.280	0.203	0.115	0.057	0.114	0.214	0.281	0.246
HFbase	0.311	0.293	0.213	0.116	0.057	0.114	0.214	0.281	0.241
logHF	0.298	0.260	0.200	0.121	0.086	0.129	0.216	0.297	0.251
logHFeng	0.311	0.293	0.213	0.116	0.057	0.114	0.214	0.281	0.241
sigmHF	0.293	0.260	0.200	0.113	0.057	0.114	0.208	0.292	0.249
sigmHFeng	0.311	0.293	0.213	0.116	0.057	0.114	0.214	0.281	0.241
BM25F_mod2 (for HR tags)	*0.238*	*0.253*	*0.163*	*0.028*	*0.086*	*0.114*	*0.055*	*0.119*	*0.103*
HRbase	0.243	0.260	0.170	0.028	0.086	0.114	0.056	0.119	0.103
logHR	0.246	0.273	0.173	0.028	0.086	0.114	0.055	0.119	0.103
logHReng	0.243	0.260	0.170	0.028	0.086	0.114	0.055	0.119	0.103
sigmHR	0.246	0.273	0.173	0.028	0.086	0.114	0.055	0.119	0.103
sigmHReng	0.243	0.260	0.170	0.028	0.086	0.114	0.055	0.119	0.103

Table 7.9: Individual subjects scores for average precision (AveP), P@5 and P@10 for BM25F_mod2+static_score approaches using the subset of the full result set which was tagged with recorded biometric response.

in Table 7.2. We observe that HF with energy expenditure considered is still overall the best performing static scoring technique for Subject 1. Similarly HF without energy expenditure remains the best performing technique for Subject 2. For Subject 3, while use of ST resulted in the greatest improvement over the BM25F_mod2 baseline on the full result lists, using the reduced result lists HF, with energy expenditure not considered, provided greatest utility.

HF:

Similar to the results observed in the previous section, overall best performance was obtained through the use of HF as a static score. What causes this superior performance using HF is not known. As for the previous section, factoring of energy expenditure was important here for Subject 1 but the reverse held true for Subjects 2

Static Technique	Subject 1			Subject 2			Subject 3		
	AveP	**P@5**	**P@10**	**AveP**	**P@5**	**P@10**	**AveP**	**P@5**	**P@10**
BM25F_mod2 (for GSR,) (HF,ST tags)	*0.304*	*0.267*	*0.203*	*0.117*	*0.029*	*0.114*	*0.213*	*0.281*	*0.241*
STbase	2%	0%	2%	-1%	100%	0%	0%	0%	2%
logST	0%	-3%	-3%	1%	100%	13%	0%	2%	3%
logSTeng	2%	0%	2%	-1%	100%	0%	0%	0%	3%
sigmST	0%	-3%	-3%	-1%	0%	0%	0%	0%	0%
sigmSTeng	2%	0%	2%	-1%	100%	0%	0%	0%	3%
sigmIncST	0%	-3%	-3%	-1%	0%	0%	0%	0%	0%
sigmIncSTeng	2%	0%	2%	-2%	100%	0%	0%	0%	3%
GSRbase	3%	5%	0%	-1%	100%	0%	0%	0%	0%
logGSR	3%	5%	2%	-1%	0%	0%	0%	0%	0%
logGSReng	3%	5%	0%	-1%	100%	0%	0%	0%	0%
sigmGSR	3%	5%	2%	-1%	100%	0%	0%	0%	0%
sigmGSReng	3%	5%	0%	-1%	100%	0%	0%	0%	2%
HFbase	3%	10%	5%	0%	100%	0%	0%	0%	0%
logHF	-2%	-3%	-2%	4%	200%	13%	1%	6%	5%
logHFeng	2%	10%	5%	0%	100%	0%	0%	0%	0%
sigmHF	-3%	-3%	-2%	-4%	100%	0%	-3%	4%	3%
sigmHFeng	3%	10%	5%	0%	100%	0%	0%	0%	0%
BM25F_mod2 (for HR tags)	*0.238*	*0.253*	*0.163*	*0.028*	*0.086*	*0.114*	*0.055*	*0.119*	*0.103*
HRbase	2%	3%	4%	0%	0%	0%	1%	0%	0%
logHR	3%	8%	6%	0%	0%	0%	0%	0%	0%
logHReng	2%	3%	4%	0%	0%	0%	0%	0%	0%
sigmHR	3%	8%	6%	0%	0%	0%	0%	0%	0%
sigmHReng	2%	3%	4%	0%	0%	0%	0%	0%	0%

Table 7.10: Subjects' percentage improvement, rounded to nearest whole number, for average precision (AveP), P@5 and P@10 over BM25F_mod2 for BM25F_mod2+static_score approaches using the subset of the full result set which was tagged with recorded biometric response.

and 3. However, it should be noted that these improvements in retrieval performance on the subjects' collections are not statistically significant for AveP, P@5 or P@10 (100 samples, Welch two sample t-test, p>0.05).

Increases in the percentage improvement, relative to use of the full result lists in the previous section, were noted for AveP and P@5 using *HFbase*, *logHF* and *sigmHF* for Subject 1. However, P@10 disimproved.

For Subject 2, AveP results using *logHF* improved from 3% in the previous sections experiment to 4% in this experiment. While *HFbase*, *logHFeng*, *sigmHF* and *sigmHFeng* did not yield improvements for P@5 for this subject in the previous experiment using the full result lists, in this experiment using the reduced result lists a substantial 100% improvement in P@5 was noted using this technique. However, using these techniques on the reduced result lists resulted in decreases in AveP.

Subject 3, similar to Subject 2, still benefited the most from the *logHF* technique. On the reduced result list this benefit was much greater than that observed on the full result list. 1%, 6% and 5% improvement in AveP, P@5 and P@10 respectively were obtained using *logHF*. *sigmHF* also proved benefical at moving relevant items upwards in the retrieved result list with 4% improvement in P@5 and 3% improvment in P@10.

HR:

Use of HR did not prove useful for Subject 2, with no improvement over BM25F_mod2 being observed. This is the same as we observed using the full result lists in Table individPercent.

However, for Subject 3 improvement over the BM25F_mod2 baseline which was not present using the full result lists was noted, in particular using the *HRbase* technique.

Using the reduced result lists, HF results for Subject 1 showed improvement over those observed on the full result list, with the two techniques which do not consider energy expenditure (*logHR* and *sigmHR*) still proving most useful.

GSR:

We observe improvements over BM25F_mod2 in AveP and P@5 across all techniques which used GSR for Subject 1. Similar to the results observed on the full result lists, not factoring energy expenditure into the GSR readings proved most useful. 3% and 5% improvement in AveP and P@5 respectively were obtained using *logGSR* and *sigmGSR*. However, while 4% improvement in P@10 was noted using these techniques on the full result lists, only 2% improvement in P@10 was obtained here.

For Subjects 2 and 3 for whom no utility was observed using GSR as a static score when using the full result set, improvement over the BM25F_mod2 baseline was obtained using the reduced result list. 100% improvement in AveP using *GSRbase*, *logGSReng*, *sigmGSR* and *sigmGSReng* was noted for Subject 2. 2% improvement in P@10 using *sigmGSReng* was obtained for Subject 3.

ST:

The techniques which yielded improvement in P@10 using ST as a static score noted on the full result lists for Subject 1 remained the same using the reduced result lists, with all techniques which considered energy expenditure yielding 2% improvement in P@10 using the reduced result lists. However, for this subject the improvements in AveP increased to 2% for all techniques which considered energy expenditure using the reduced result list.

The 1% improvement in AveP using logST observed on the full result lists for Subject 2 remained unchanged. However, further improvements, which were not observed on the full collection, were also gained using the reduced result lists for this subject. Specifically, 100% improvement in P@10 using *STbase*, *logST*, *logSTeng*, *sigmSTeng* and *sigmIncSTeng*. 13% improvement in P@10 was also obtained for *logST*.

For Subject 3, the 1% improvements in P@10 observed on the full result lists, increased to 2% using *STbase* and 3% using *logSTeng*, *sigmSTeng* and *sigmIncSTeng* on the reduced result lists. However, *logST* was the best performing ST technique for this subject on the reduced result lists, yielding 2% improvement in P@5 and 3% improvement in P@10.

7.3.4 Further Analysis

Based on the results in the previous sections, we can arguably say that overall across the three subjects greatest utility was found using HF as a static score to re-rank ranked retrieval result lists. Greatest utility is observed for Subject 1 when energy expenditure is factored and greatest utility is observed for Subjects 2 and 3 when energy expenditure is not factored in. Specifically, using *HFbase*, *logHFdivEng* and *sigmHFdivEng* gave the best performance for Subject 1 and using *logHF* yielded the best performance for Subjects 2 and 3.

What causes HF to provide greater utility than GSR, ST and HR is unclear, and analysis of this topic is beyond the scope of our current work. What causes the factoring of energy expenditure into the HF readings for Subject 1 to be important and not for the other subjects is also unclear. To analyse this further, we examine the heat flux and energy expenditure readings recorded for each subject, shown in Table 7.11. In this table we see variations in the energy expenditure readings across the three subjects, but nothing that makes Subject 1 stand out as different from the other two subjects. Subject 3 is observed to have the highest maximum energy reading and the largest range. Subject 2 has the lowest maximum energy reading and the smallest range. However, for HF Subject 1 has a notably larger range of readings than the other two subjects. If

Measure	Subject1	Subject2	Subject3
Energy Expenditure			
Range	9.4210	6.2306	11.3546
Min	0.9611	0.9097	1.3364
Max	10.3821	7.1403	12.6910
Median	1.2261	1.0743	1.5294
Average	1.7459	1.4159	1.8698
Standard deviation	1.2056	0.9199	0.9875
Heat Flux			
Range	418.6397	184.2164	292.8851
Min	0.0623	13.2509	50.3201
Max	418.7020	197.4673	343.2052
Median	122.9309	89.7858	104.2617
Average	127.6795	91.3807	106.8151
Standard deviation	26.5487	14.3210	19.0009

Table 7.11: Statistics on heat flux and energy expenditure readings recorded for subjects.

Measure	Subject1	Subject2	Subject3
Heat Flux			
Range	343.5474	49.5711	180.9575
Min	44.7479	87.4216	61.4442
Max	388.2953	136.9927	242.4017
Median	150.1228	97.9087	103.2233
Average	162.4652	98.8280	105.2820
Standard deviation	51.9503	8.8796	17.2874
Heat Flux divided by Energy Expenditure			
Range	182.6907	39.7705	83.4845
Min	17.3788	82.3389	16.7783
Max	200.0696	122.1094	100.2628
Median	111.8734	90.5645	67.2640
Average	114.6962	91.8826	66.4827
Standard deviation	24.3706	7.9073	12.0835

Table 7.12: Statistics on heat flux and heat flux divided by energy expenditure readings captured by the biometric device and assigned to items retrieved for subjects.

and why this results in the necessity to factor energy expenditure into HF readings for this subject is not clear. Further, if and why this would lead to factoring energy expenditure readings into HF for the other subjects not being as effective as considering HF levels on their own is unknown based on the study.

We next analyse the HF and HF/engHF readings which were captured by the biometric device (calculated based on data captured by the biometric device in the case of HF/engHF) and tagged to items retrieved for BM25F_mod2 content+context retrieval. This revealed the statistics shown in Table 7.12. We observe that the HF/engHF statistics for Subject 1 are more inline with the HF statistics for Subjects 2 and 3. For all subjects dividing HF by the energy expenditure reduces the variance from the average; in the case of Subject 1 this reduction is dramatic. This leads us to the supposition that Subject 1's HF readings may have been more affected by unknown external factors than the other subjects, in particular when interacting with computer items (as the HF levels under investigation here were tagged to retrieved computer items). Factoring of energy expenditure into the HF readings appears to counter these external factors. Again though why the reverse would be true for Subjects 2 and 3 and why this would result in consideration of energy expenditure in their HF readings resulting in inferior performance is not known. Of course, other unknown factors could also be playing a part here. Further analysis on this topic and understanding of the role of energy expenditure in HF readings could form the basis of future physiological research using larger groups of subjects.

As a final note on this topic we observe, in Tables 7.9 and 7.10, for Subject 1 that consideration of energy expenditure in the ST readings is also important. Suggesting that whatever external factors are causing the increases in HF are also present for ST readings.

7.4 Discussion

Prior experiments similar to those presented in this chapter were conducted [Kelly and Jones, 2010a]. These experiments also showed potential utility for biometric response associated with past experience of lifelog items as a static score in ranked retrieval result lists. However, these experiments suffered a number of short comings which the experiments presented in this chapter addressed. Further these experiments did not go as far, in analysing the use of biometric response as a static score. In particular in our prior experiments:

- An earlier version of the test sets was used, from which all noise/problems for computer items discussed in Chapter 3.3.1.3 were not removed. Specifically several items for which we subsequently obtained content data did not have content data in these experiments. This would have impacted on the ranked retrieval performance.

- Our prior experiments explored adding the biometric static scores to BM25 content only retrieval and to BM25 structured content+context retrieval algorithms, whereas in this chapter we use our best performing content+context retrieval algorithm namely BM25F_mod2.

- Our prior experiments only used a subset of the context types used in the experiments presented in this chapter for content+context ranked retrieval.

- The biometric data from the beginning and end of each period of wearing the biometric devices which was skewed (as discussed in Chapter 3.3.4.4) was not removed from the dataset.

- Energy expenditure values were not calculated on the given subject's personal details (i.e. age, weight and height), due to the devices not being set with the subject's personal details.

- Items assigned default biometric readings in the experiment were assigned the average of available biometric readings as opposed to the median used in this chapter's experiments.

- In the previous experiments when tagging items with biometric data with energy expenditure factored in items were tagged with their highest GSR, HR, HF and 1/ST response across all accesses to the items, and these highest GSR, HR, HF and 1/ST readings were divided by their corresponding energy expenditure readings. On consideration, this breaks with the spirit of factoring energy expenditure readings into the GSR, HF, HR and 1/ST readings, where the purpose of factoring energy expenditure readings into the biometric readings is to get at the true biometric levels with certain external factors removed. Hence in the experiments presented in this chapter we tagged items with their highest $\frac{GSR}{engGSR}$, $\frac{HR}{engHR}$, $\frac{HF}{engHF}$ and $\frac{\frac{1}{ST}}{engST}$ readings.

- Heat flux was not considered in the prior experiments.

- Experiments to explore the effect of assigning default biometric tags to items missing biometric readings were not conducted. This issue was examined in the experiments presented in this chapter in Section 7.3.3.

7.5 Conclusions

In this chapter we set out to investigate the role of biometric response in lifelog item retrieval. We presented a novel approach for calculating a static score factor based on an individual's biometric response combined with ranked retrieval relevance scores. Results obtained show some support for the utility of biometric response as a score factor in lifelog search. Greatest overall improvement in performance was found by the addition of a static HF score, with greatest improvement being observed for Subject 1 when energy expenditure was factored into the HF levels and for Subjects 2 and 3 when energy expenditure was not factored into the HF levels. These results are in contrast to those observed in Chapter 6 using biometrics to extract important computer items, where ST levels with energy expenditure factored (i.e. the *divEng* technique) proved most beneficial for Subjects 1 and 3, and GSR levels with energy expenditure factored (i.e. the *divEng* technique) proved useful for Subject 2. These differences in results across experiments and subjects suggest that: there may not be one overarching 'best' biometric measure for detecting item importance based on biometric levels associated with past experience of items; that the same biometric measures may not be useful across different retrieval scenarios; and that individual difference is an important factor affecting the utility of different biometric measures. Future studies on these topics are merited.

That being said, the results observed in this chapter are promising. However it is acknowledged that this study was conducted on a limited number of subjects over a relatively short period of time. Further experiments with larger numbers of subjects are required to establish the scalability of the technique presented in this chapter. However, due to the large psychological burden placed on subjects wearing the biometric devices for extended periods of time, and the difficulty in gaining participants willing to partake in experiments which log their personal data, this initial study formed a good means to establish if further research in this domain is warranted. Given the results presented in this chapter we believe it is worth investing in further research in this space using larger collections of subjects. In the future work section of the next chapter we discuss some avenues for such research.

Part IV

Conclusions

CHAPTER

EIGHT

Conclusions and Future Work

Technological developments are enabling individuals to store increasing amounts of digital data pertaining to their lives. As these personal archives grow ever larger, reliable ways to help individuals locate required items from these collections become increasingly important. In this thesis we set out to further research this space through consideration of the backend retrieval challenge. Our exploration focused on the role of recalled context data and biometric indicators of item importance in lifelog retrieval. We developed algorithms which allow individuals to query their PLs based on their recalled content and context associated with lifelog items, and which also factor biometric response associated with past experience of items into the retrieval process. In the next section we conclude this thesis, providing an overview of each chapter and the main outcomes from them. We also overview how the thesis objectives listed in the introduction chapter have been met. In Section 8.2 we move on to provide several avenues for future research in the scope of the work presented in this thesis.

8.1 Conclusions

In this thesis we set out to explore the utility of implicitly recorded and derived context types in lifelog retrieval. In the next section we provide a summary of key points in the thesis, along with key observations and findings. The following section then revisits the thesis objectives presented in Chapter 1, and overviews how these objectives have been met.

8.1.1 Summary of Presented Work

Chapter 2:

We presented a discussion of research related to this space and explained how it motivates the studies presented in this thesis. A summary of existing research was provided showing that memory associated with past experience of digital items plays a vital role in retrieval. It was also shown that this memory can be harnessed in the form of context data associated with items, e.g. date of last access to an item. We could however, not find evidence that this memory of past experience with items has been used in previous work on advanced retrieval algorithms for the personal space. We put forward that this is an important, previously unexploited opportunity to help people locate required data in personal collections.

We highlighted past research which showed that significant or important events tend to increase arousal levels and that in turn biometric levels increase. Based on this, we proposed that moments of increased biometric response associated with interaction with lifelog items, or based on life experiences captured by the SenseCam, might indicate the events/items in lifelogs which would be most interesting for the lifelog owner to view or most important to them. We further supposed that this might result in such biometric response being useful in re-ranking ranked result lists.

We also highlighted the challenges associated with evaluation in this space. A review of evaluation approaches taken to date was provided, which led to the evaluation approach taken in this thesis. Based on this research, we decided to adopt the approach of creating lifelogs by recording subjects digital activity over an extended period of time. This approach was taken given the unavailability of existing test sets and test cases for research in this space.

Chapter 3:

We presented the test sets created for experimentation in this study. These test sets consisted of laptop and PC files interacted with, emails sent and received, web pages viewed, SMS messages sent and received, and SenseCam images captured. These lifelog items were annotated with several sources of automatically generated context data. Specifically, each item was annotated with title, path to file, URL, extension type, to and from information. The following context data types associated with each access to these items were also included: begin date and time; end date and time; year; season; month; day of week; whether access happened during the week or at weekend; whether access happened during the beginning of week, mid week or at end of week; whether the access happened in the morning, afternoon, evening or night; device used to access the item (e.g. laptop); geo-location when accessing the item; light status when accessing the item; weather conditions when accessing the item; people present when accessing the item; and biometric response when accessing the item.

To our knowledge these are the largest and most heterogeneous test sets of their nature in existence. We described the process by which these test sets were created, along with the challenges and problems encountered in creating them. We also showed the resulting size of our test sets and statistics on their make up and suggested how this might impact on retrieval from them.

These test sets were just for three subjects, all of whom were post-graduate students within our department. Despite this similarity in our test subjects, variation was observed in the make up of the three subjects collections. These variations were highlighted and indicators presented as to how this might impact on the utility of context data in retrieval on the subjects collections. As an aside, given the individual differences observed in our three, quite similar, subjects' collections one can expect much larger variations to occur across the personal collections of the wider populous. This leads to the need for flexible retrieval techniques which can cater for individual differences.

Chapter 4:

We explored retrieval algorithms of interest to our research, and described the state-of-the-art in this space which we used as a basis for our retrieval algorithm investigations presented in Chapters 5 and 7. We also described how we created indexes of the textual content of our generated lifelog test sets for the retrieval approaches we wished to investigate. The means by which we generated user queries and result sets for these queries was also described. These user queries were based on subjects' perception of items they would want to retrieve from their lifelogs and their recalled content and context associated with these items. Result sets were generated based on manual user ratings of pooled result lists generated using a combination of different retrieval approaches.

Chapter 5:

In Chapters 5-7 we described our experimental studies, the justification for these studies, and provided a detailed analysis of results obtained. Specifically, in Chapter 5 we explored developing ranked retrieval algorithms for the lifelog domain. Ranked retrieval approaches explored allowed individuals to search based on both their recall of the content in their collections and several rich sources of associated context data. As part of this exploration we analysed the utility of recalled context in improving content only based retrieval for our three experiment subjects. We looked at a structured querying based retrieval approach and several flat querying based retrieval approaches. Our primary contribution in this chapter was the suggestion and validation of a novel flat retrieval technique. This technique was an extension of the state-of-the-art BM25F retrieval approach. This technique modified BM25F at the field term scoring level, with the aim of accounting for the structure of lifelogs. Our first modification, consisted of a term frequency normalization of the term score. This approach weighted term field scores based on the frequency of the term in the document field relative to its frequency across all fields of the document. This approach did not show utility in our collections in which context terms were also liable to occur in the content field of documents. Our second, and overall most successful, modification consisted of a field length normalization of the term score. This approach weighted term field scores based on the length of the field relative to the length of the document as a whole. In this approach the occurrence of terms in shorter fields received a higher boost than the occurrence of terms in longer fields. Overall, this approach showed utility in the experimental lifelog collections, in which context query terms could also occur in the content of documents. However, with this approach we noted the least improvement in performance for the one subject who used context query terms in the content field of their queries. For this subject our adopted approach was liable to give incorrect extra weighting to content query terms which occurred in the context fields of non-relevant items. We hypothesized that combining our term frequency normalization and length normalization approach in weighting the BM25F terms scores would help dampen this effect. This supposition was shown to be correct in the case of this subject. However, for the other two subjects using this approach was not as effective as using length normalization to weight the field level scores in isolation.

Chapter 6:

In Chapter 6 we began our exploration into the use of biometric response associated with past experience of lifelog items as static scores in ranked retrieval algorithms. To begin this exploration we attempted to establish if such biometric response indicated future importance of items in lifelogs. We were particularly interested in the textual content in subjects' lifelogs, since this was the data for which we were developing retrieval techniques. However, given the multimedia nature of lifelogs, and the particular benefit we felt might be achieved using biometric response in detecting important SenseCam events and the large need for extracting interesting SenseCam events from amongst the potentially vast quantities of images available, we also explored the utility of biometrics in extracting such events.

We proposed a mechanism to extract possibly important or interesting SenseCam events and computer items from lifelogs using biometric response levels observed from past experience of lifelog items. The biometric measures we explored were galvanic skin response (GSR), heat flux (HF), skin temperature (ST) and heart rate (HR). This represented the first time that biometric response had been used in this way, i.e. as a future indicator of item/event importance. We also proposed that energy expenditure associated with biometric response could be used to remove some external factors (e.g., motion, eating) that affect biometric response. In other words, that factoring associated energy expenditure readings into biometric readings could help in 'getting at' the true biometric response associated with interesting items/events. Results obtained validated our approaches, but highlighted the need for further research in this space. Different biometric measures and approaches showed greater utility for different subjects. That is, no clear 'best' biometric measure for important item/event detection was observed. Further, in the case of one subject who regularly browsed their SenseCam collection and found everything from the mundane to the novel interesting to view, past biometric response did not provide much utility as an enabling technology for interesting SenseCam event detection.

Chapter 7:

Having gained positive results in Chapter 6 investigations, in Chapter 7 we moved on to explore the utility of integrating past biometric response as a static score in the overall best performing ranked retrieval technique which we developed in Chapter 5.

In this investigation we used the item biometric tagging approach used in Chapter 6. Our investigation showed utility for using biometric static scores, but again no clear candidate biometric measure for use in this approach was observed. HF was found to be the overall most useful biometric measure for static score creation across the three subjects. However, for two subjects greatest utility was observed when energy expenditure was not factored into the HF readings. For the other subject the reverse was true. We suggested that this was due to the unexplained large HF level variations in this subject's collection relative to the other subjects' collections, which factoring in of energy expenditure appeared to remove.

Remarks:

As acknowledged at several points throughout this thesis, while the results of our experiments are promising, our studies were conducted on only three subjects over a 20 month period using 100 queries per subject for our ranked retrieval investigations presented in Chapter 5, and using a one month subset of our three subjects lifelogs for the biometric investigations presented in Chapters 6 and 7. That being said, these collections are larger and more heterogeneous than any other collections used to date in this space. Further, this is a new space of research, in which retrieval algorithm development using rich context data has not previously been explored. Indeed even across other spaces of research, no work exists to our knowledge which explores the integration of the rich context types we looked at into ranked retrieval algorithms. Our studies also present the first use of biometric response associated with past experience of digital content as a future indicator of that contents importance. And subsequently the first studies examining the integration of such biometric response as static scores in ranked retrieval approaches. Importantly in these experiments we suggested and showed utility in using energy expenditure as a means to remove some external influencing factors from biometric readings - this appears to be important in helping get at a truer indication of underlying arousal levels using biometric response in some cases. We believe that the work presented in this thesis motivates further research in this space using larger scale studies and that it sets the foundations for such further research. We also believe that we have opened the doors to exciting new spaces of research, from which there is much scope for future research. In the future work section of this chapter we highlight some of these avenues for future research.

8.1.2 Achievement of Thesis Objectives

The primary contributions of this thesis can be summed up as:

- Review of related work in personal information systems, context data in retrieval and the use of biometrics in retrieval, from which we showed:

  Support for using recalled context data in retrieval; Lack of sophisticated retrieval algorithms for the personal space; Lead up to the use of biometric response associated with experience of lifelog items as a future indicator of the item's importance through review of biometric response research and current state of explorations in biometric response in the digital environment.

- Test set generation:

  Creation of the largest known most heterogeneous test sets for experimentation in this space, and details of how these collections were created along with the challenges in creating such collections. A unique insight into the make up of such collections was also provided.

- Generation of lifelog IR algorithms which cater for content+context queries:

  We explored the utility of recalled context data in IR retrieval algorithms. We examined this from two perspectives: 1) structured content+context queries; and 2) flat content+context queries. We showed support for the use of queried context data in lifelog retrieval. We introduced a novel flat ranked retrieval approach to account for the make up of our subjects' collections and querying approaches. Preliminary support for this retrieval approach was shown.

- Generation of techniques to extract important items from lifelog collections using biometric response associated with past experience of items:

  We provided evidence, through experimentation using our generated test sets, to support our hypothesis that biometric response associated with previous experience of lifelogged items/events can be used to detect items/events that our subjects may wish to view in the future.

- Generation of techniques to integrate biometric measures associated with past experience of items into lifelog retrieval algorithms:

Based on our observed utility of past biometric response in detecting important computer items in lifelog collections, we explored the utility of past biometric response in re-ranking ranked result lists. We found support for the addition of maximum observed biometric levels, associated with past interaction with computer items, as a static score to our developed content+context ranked retrieval approach.

8.2 Future Work

PL retrieval research is a new and exciting domain. This is just the beginning of the story. There is much work still to be done. Further work on applications to allow individuals to manage, browse through and search their lifelogs; examination of security and privacy issues to be accounted for; establishing provisions for sharing of lifelog content; the list goes on. The future in this domain of research is set to be interesting! Specifically from the point of view of the retrieval element of PL research, this and other existing work just forms the beginning of the story. Some possibilities for the next chapter of the story follow.

8.2.1 Evaluation Techniques

Given the complexity of creating lifelogs and the problems we encountered with our approach (described in Chapter 3), we would recommend adopting a different lifelog creation approach in future lifelog studies. One possibility is to create an integrated lifelog capturing solution from the ground up. Such a solution should automate data capture, data integration, time aligning and writing to final lifelog database.

Alternatively, existing systems could be extended. For example, for studies conducted on the Mac operating system, the Slife application is now available in open source for this operating system. This could be easily extended, towards a lifelogging solution, to capture item content, file path and the ’to’ ’from’ fields of emails.

Whatever solution is adopted, reducing the computer speed of the user while logging is not an acceptable option, as we found with even our dedicated lifeloggers. In generating a lifelogging solution, information such as file path, webpage URL, email title and timestamp could be logged live. With further details, such as item content, context tagging, time aligning and database updates conducted during system idle time. Complete control by subjects of their personal logged data, using local to the subject stores only, and zero personal data access to investigators are also imperative factors to incorporate into a lifelog generation and evaluation solution using these lifelogs.

Given the personal nature of such lifelog collections and resulting difficulties associated with gaining subjects, the over head involved in creating lifelogs for experimentation purposes and the lack of cross comparability of techniques developed on the same lifelogs across institutions, there is need to move towards standardization in evaluation in this domain. A point which formed the focus of the ECIR 2011 workshop on evaluating personal search[1]. We propose that one possible solution to this problem and avenue for future research is through generation of pseudo collections which exhibit the characteristics of 'real' user collections. To create such collections a detailed understanding of the make up of real users desktop collections, items they retrieve from these collections and query formation styles is required. Part of such an analysis could take the form of observations, user studies, diary studies, etc, as are carried out in the personal information management (PIM) community [Jones, 2006, Teevan and Jones, 2008, Barreau et al., 2009]. However, a detailed statistical analysis of the make up of the collections and querying behaviour of a large cross section of the populous is also required in order to move to a situation where real users collections can be replicated in a pseudo way. To understand the make up of individuals desktop collections, statistics need to be built up on the volume of different information types in these collections, the volume of topics covered, the amount of similarity between items, etc. This analysis could potentially be conducted through a drive within the research community, with either clear guidelines on the statistics to gather or crawlers to automatically generate statistics from participants PCs provided. We propose that required statistics for target result items would include: extension type of target item, distinctiveness of target item in collection as a whole, recency of last access to target item, etc; and that required statistics for user queries would include: query length, frequency of query terms in target item, frequency of query term in collection as a whole, etc. Similar to gaining statistics on the content of individuals desktops, a stand alone search application or a tool which plugs into individuals current search application (e.g. Google Desktop[2]) could be provided to the research

[1] http://www.cdvp.dcu.ie/iCLIPS/EPS2011/ (September 2011)

[2] http://desktop.google.com/ (September 2011)

community to log statistics on the nature of queries performed and items retrieved on subjects computers.

Using the statistics gathered for each individuals desktop contents, query format and items retrieved, we believe the techniques developed in [Azzopardi et al., 2007] and [Kim and Croft, 2009] provide a strong foundation from which to build pseudo test collections which mimic the characteristics of 'real' test collections. We propose mimicking desktop content by using the statistics gathered on the make up of individuals' desktop content to lay user profiles on top of an extension to the pseudo desktop collection creation approach proposed in [Kim and Croft, 2009]. In extending this approach, other information which could be mined in creating these collections includes the details provided by people on their homepage, e.g. many people provide lists of personal and work interests and details on co-workers (either explicitly or through inferred means, e.g. co-authorship of papers in the case of academics) on their homepages. We also envisage possibilities to extend the content gathering approach to include other item types and items generated from web content using existing summarization, extraction and rephrasing approaches, for example. Having created pseudo desktop collections we propose extracting target result sets from each user's collection using the available statistics on what the 'real' user retrieves from their collection. To form the queries for the target items, a query generation process which uses the statistics on the 'real' users query formation for the given target item is required. We envisage the query generation approach proposed by [Azzopardi et al., 2007] and refined to facilitate multi-field retrieval by [Kim and Croft, 2009], coupled with the information gained by our proposed statistical analysis would form a good starting point for development of a query generation process for this space.

8.2.2 Context Data in Retrieval

There is much scope for future research in the space of using recalled context data in IR algorithms in the lifelogging domain. In our investigations we used a wide selection of automatically generated context types. However, there are possibly many other forms of rich context which people recall associated with items. Possibilities here might include items accessed around the same time as the required item, SMS messages sent/received or phone calls made around the same time as the required item, etc. Qualitative research exploring this topic, along with means to capture or derive ’new’ context types would be useful.

In our studies we did not differentiate between recalled context associated with creation of items and that associated with subsequent accesses to items. Future studies could look at the impact of making such a distinction, both from the point of view of individuals’ recall of creation and access context and the impact on retrieval performance of making this distinction. Further, in our retrieval experiments, presented in Chapter 5, we acknowledged that it was possible that errors in subjects’ memories of content and context associated with required items could have occurred. Future research could explore sophisticated techniques to account for the possibility of incorrect recollection. We further highlighted in our retrieval experiments that retrieval is most probably impacted by missing context data in our collections. Future, more reliable lifelogging solutions will be more robust and should not suffer from missing data to the same extent. However, there will most likely still remain a certain amount of missing context data in collections. It would be useful to investigate means to account for such missing data. Perhaps through the filling in of missing data using surrounding available data with predetermined confidence rules.

Another issue affecting the querying using context experiments of Chapter 5, is that predefined context terms were used (e.g. the term ’web’ was used in the ’extension type’ field to indicate web pages). In the future simple mappings would be required to disambiguate where an individual enters for example the query term ’webpage’ that they mean ’web’. This is an achievable aim in the structured query space. However, presented with a flat query, in which the target fields for query terms are not known generation of such mappings will be a complex task.

The results obtained from our ranked retrieval investigations in Chapter 5 suggest that it will not be possible to develop one overarching technique that is most successful across all PL collections and collection types. Rather a personalised retrieval approach, such as a suite of functions for retrieval of different item types and different functions which work under different conditions are required. Such functions would respond differently to queries of varying length and might down or up weight different fields depending on the volumes of different types of data in a field (for example, a geo-location field could be down weighted for an individual who rarely moves between locations). Retrieval functions of this nature might be informed by a statistical analysis of the nature of individuals’ personal collections, of the type described in the previous evaluation approach section.

A final point on the use of recalled context in retrieval. While it was beyond the scope of this thesis to explore retrieval algorithms for all types of items which may be contained in a PL, development of retrieval techniques for these item types is necessary. An initial exploration of this topic could explore the utility of the retrieval techniques presented in this thesis across other item types (i.e. retrieval of non-textual items), by allowing people to query based on their recalled context associated with these items. Further context types could also be added to these items, for example people in photographs, artist associated with music, etc. The possibility of linking non-textual items to textual items in lifelogs to gain content for these items could also be explored. Such linking might be based on weighted temporal links for example.

8.2.3 Biometrics in Retrieval

The results of our biometric experiments presented in Chapters 6 and 7 indicate that biometric readings serve as a useful tool for aiding extraction of important items from long-term lifelogs. Our adopted approaches for using biometric levels in this regard were intended as a first pilot venture into this space. Future research needs to be carried out to gain greater understanding of the patterns of biometric response observed when individuals are experiencing interesting or important events (or other types of lifelogged events), which they will wish to retrieve in the future. Given the mixed utility we observed from factoring energy expenditure into biometric levels in our experiments, future research should also seek to gain greater understanding of the use of energy expenditure with biometric levels. Findings from such studies could form the basis of improved techniques for using biometrics to aid lifelog item extraction.

Additionally, beyond the lifelogging domain, we envisage several possible applications of the biometric techniques presented in this thesis both in the archive searching, recommendation space and in particular to SenseCam images helping locate important events for memory impaired individuals engaged in SenseCam focused memory therapy [Baecker et al., 2007, Berry et al., 2007]. Indeed in a future where biometric recording is prevalent, the same patterns of biometric response may be observed across individuals for the same items in shared archives (e.g., digital libraries, photograph archives, retail websites), which might allow such items to be given query independent boosts for all users of the archive. Current research exploring development of less cumbersome biometric recording devices, for example research at MIT Media Lab[3], provides indication that reliable unobtrusive biometric devices embedded in individuals clothes or bracelets for example will be widely available for use by such tools.

8.3 Summary

Research into retrieval from personal lifelogs is becoming increasingly important. The Ph.D. research programme presented in this thesis, aimed to contribute to this rapidly growing area of research by exploring methods to integrate implicitly recorded and derived context data types into traditional IR algorithms for the lifelog retrieval domain.

We postulated that context data can be used to harness the way people remember items in their lifelogs and that past biometric context can be used to locate important lifelog items. If this context information is exploited correctly, we demonstrated that it may be possible to create a system that retrieves items based on both an individual user's unique information needs and on what they remember about items.

Personal digital archives are increasingly becoming part of our present. In the near future we believe it will be hard for people to imagine a world where personal lifelogs did not exist. In this thesis we embraced this future by investigating retrieval techniques that integrated recalled content and context with biometric response associated with past experience of lifelog items to address some of the unique retrieval requirements of personal lifelogs.

[3] http://affect.media.mit.edu/index.php (September 2011)

Part V

Appendix

APPENDIX

A

List of Publications

This Ph.D. research was carried out as part of the Science Foundation Ireland Research Frontiers Programme 2006 funded iCLIPS project. Further details of the iCLIPS project are available at: http://www.cdvp.dcu.ie/iCLIPS/ (September 2011). The following publications arrive wholly or partially out of this Ph.D. research. However, much of their content does not appear in this thesis, rather they were explorations conducted in this space prior to writing the thesis and conducting thesis experiments. These publications inform much of the thesis. Copies of these publications can be obtained from the iCLIPS project website at: http://www.cdvp.dcu.ie/iCLIPS/publications.html (September 2011).

A.1 Towards Context Data in Lifelog Retrieval

These publications were written early in the Ph.D. They propose and explore the idea of using context data and memory in lifelog retrieval. Many of the ideas expressed in these publications explore the premises of this thesis.

- L. Kelly. "The Information Retrieval Challenge of Human Digital Memories". In Proceedings of the BCS IRSG Symposium: Future Directions in Information Access (FDIA 2007), pg. 114-122, Glasgow, Scotland, 28-29 August 2007, British Computer Society.
- L. Kelly and G.J.F. Jones. "Venturing into the Labyrinth: the Information Retrieval Challenge of Human Digital Memories". Proceedings of the Workshop on Supporting Human Memory with Interactive Systems, at HCI 2007: The 21st

British HCI Group Annual Conference, pg. 37-40, Lancaster, U.K., 3-7 September 2007.

- L. Kelly. "Searching Heterogeneous Human Digital Memory Archives". In K-Space Jamboree Workshop, Berlin, Germany, 14 September 2007.
- L. Kelly. "Context and Linking in Retrieval from Personal Digital Archives". SIGIR 2008 - Doctoral Consortium, 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pg. 899, Singapore, 20-24 July 2008, ACM.

A.2 Context Data in Lifelog Retrieval

These publications present the first explorations into using several rich sources of context data to retrieve items from one subject's sample personal collection consisting of six weeks of computer activity. The first publication is a result of the M.Sc. project of Marguerite Fuller, where retrieval effectiveness using recalled content and context was tested directly after the six weeks of data capture. The second publication presents a follow up study conducted six months later to examine the change in recalled memory of content and context, and resulting change in retrieval effectiveness after the six month interval.

- M. Fuller, L. Kelly and G.J.F. Jones. "Applying Contextual Memory Cues for Retrieval from Personal Information Archives". In Proceedings of Personal Information Management (PIM 2008), workshop at CHI 2008, Florence, Italy, 5-6 April 2008.
- L. Kelly, Y. Chen, M. Fuller and G.J.F. Jones. "A Study of Remembered Context for Information Access from Personal Digital Archives". In Proceedings of the 2nd International Symposium on Information Interaction in Context (IIiX 2008), pg. 44-50, London, U.K., 14-17 October 2008.

The following publications propose ideas for developing applications which allow people search their lifelogs using their episodic memory of lifelog items. This is the work of Yi Chen also engaged as a researcher on the iCLIPS project, and hence is not discussed in this thesis. My contribution to these publications is provision of the backend content+context retrieval algorithms. These algorithms are those described in Chapter 5. However precise implementation details of these algorithms, nor evaluation of them are discussed in these papers.

- Y. Chen, L. Kelly and G.J.F. Jones. "Memory Support for Desktop Search". In Proceedings of Desktop Search Workshop at SIGIR 2010, pg. 13-16, Geneva, Switzerland, 23 July 2010.
- Y. Chen, L. Kelly and G.J.F. Jones. "Supporting Episodic Memory from Personal Lifelog Archives using SenseCam and Contextual Cues". Poster presentation at SenseCam Symposium 2010, Dublin, Ireland, 16-17 September 2010.

This paper explores the distinguishing ability of geo-location information on a subset of our 3 subjects 20 month lifelogs. This paper is joint work with Daragh Byrne. It should however be noted that the statistics presented in this paper are not comparable with our 20 month lifelogs as these experiments were conducted on the earlier version of our lifelogs (described in Chapter 7.4). This paper raises some important issues for the use of geo-location in lifelog item retrieval and the underlying premise of the paper remains valid for the 20 month lifelogs used in this thesis.

- L. Kelly, D. Byrne and G.J.F. Jones. "The Role of Places and Spaces in Lifelog Retrieval". In Proceedings of Personal Information Management (PIM 2009), Workshop at ASIST 2009, Vancouver, Canada, 7-8 November 2009.

A.3 Biometrics in Lifelog Item Importance Detection

This paper presents the concept of using biometrics to extract important lifelog events (SenseCam images and computer files), which individuals may wish to view using biometric response at the time of experiencing the lifelogged events. To do this the first approach presented in Chapter 6 was used. Specifically, removing periods of high energy expenditure. This experiment differs from that presented in Chapter 6 in a number of ways (as discussed in Chapter 6.4) and formed part of our initial explorations on this topic, from which further insight into the use of biometrics in important item selection was obtained and the approach for doing this refined (refined approach presented in Chapter 6 of this thesis). Given the results of this experiment we were motivated to explore this topic further, as described in Chapter 6, and in particular explore the long-term utility of biometrics in locating interesting lifelog events. Comparison between the experiment in this paper and that presented in this thesis is drawn in Chapter 6.4.

- L. Kelly and G.J.F. Jones. "Examining the Utility of Affective Response in Search of Personal Lifelogs". In Proceedings of the 5th Workshop on Emotion in HCI, British HCI Conference 2009, Cambridge, U.K., 1 September 2009.

The following paper presents a related study on biometrics not directly relevant to this thesis. In this paper another possible use of SenseCam and computer events extracted from lifelogs based on biometric response levels is explored. Specifically the utility of these events in self-reflection. Although this paper was just published in 2010, the experiments were conducted in 2008, one month after the subjects had collected the biometric data. These experiments too use an earlier version of our lifelogs (described in Chapter 6.4).

- L. Kelly and G.J.F. Jones. "An Exploration of the Utility of GSR in Locating Events from Personal Lifelogs for Reflection". In Proceedings of the 4th Irish Human Computer Interaction Conference (iHCI2010), Dublin, Ireland, 2-3 September 2010.

The annual SenseCam symposium provided an opportunity for us to share the SenseCam element of the aforementioned biometrics in lifelog event extraction experiments with the SenseCam community. This was done in the following poster presentations.

- L. Kelly and G.J.F. Jones. "Examining the Utility of Biometric Response for SenseCam Archive Browsing". Poster presentation at SenseCam Symposium 2009, Chicago, USA, 16-17 October 2009.
- L. Kelly and G.J.F. Jones. "An Exploration of the Utility of Affective Response in SenseCam Archives". Poster presentation at SenseCam Symposium 2010, Dublin, Ireland, 16-17 September 2010.

The following paper explored the use of biometric response levels associated with past experience of lifelog items as a static score in ranked retrieval algorithms. It too used an earlier version of our lifelogs. The results of this paper raised many interesting points and gave us greater insight into use of biometrics as static scores. Comparison between this static scoring experiment and that presented in this thesis is drawn in Chapter 7.4.

- L. Kelly and G.J.F. Jones. "Biometric Response as a Source of Query Independent Scoring in Lifelog Retrieval". In Proceedings of 32nd European Conference on Information Retrieval (ECIR 2010), pg. 520-531, Milton Keynes, UK., 28-31 March 2010, Springer-Verlag.

A.4 Supporting Evaluation in the Lifelogging Domain

This paper was an effort exploring issues of evaluation for applications in the lifelogging domain.

- G.J.F. Jones, C. Gurrin, L. Kelly, D. Byrne and Y. Chen. "Information Access Tasks and Evaluation from Personal Lifelogs". In Proceedings of the 2nd International Workshop on Evaluating Information Access (EVIA 2008), pg. 75-86, A Satellite Workshop of NTCIR-7, Tokyo, Japan, 16 December 2008.

The following book chapter was written in collaboration with Daragh Byrne as a review of our experiences in collection of large scale long-term personal lifelogs. In it challenges associated with creating lifelogs are explored, along with possible solutions and guidelines for those wishing to create lifelogs for experimentation purposes. While the challenges associated with creating lifelogs are described in Chapter 3 of this thesis, the guidelines presented for future lifelog test set creators are those of Daragh Byrne, and hence do not feature in this thesis. I would recommend reading this book chapter to anyone considering working with the lifelogging technologies discussed in this thesis.

- D. Byrne, L. Kelly and G.J.F. Jones. "Multiple multimodal mobile devices: Lessons Learned from Engineering Lifelog Solutions". In: Handbook of Research on Mobile Software Engineering: Design, Implementation and Emergent Applications, IGI Publishing, 2010.

In this publication a means to move towards standardized evaluation in the personal search space is presented. The ideas presented in this paper are included in the future work section of this thesis (Chapter 8.2.1).

- L. Kelly, G.J.F. Jones. "Information Access Tasks and Evaluation from Personal Lifelogs". In Proceedings of Evaluating Personal Search Workshop, ECIR 2011, 18 April 2011.

A.5 SIGIR 2010 Desktop Search Workshop

The following proceedings are those of the SIGIR 2010 Desktop Search Workshop and the report is a write-up of the outcomes of this workshop. These provide a concise insight into the current state of research in the desktop search space.

- D. Elsweiler, G.J.F. Jones, L. Kelly and J. Teevan (editors). "Proceedings of the SIGIR 2010 Workshop on Desktop Search". Geneva, Switzerland, July 2010.
- D. Elsweiler, G.J.F. Jones, L. Kelly and J. Teevan. "Report on Desktop Search Workshop". SIGIR Forum, Vol 44(2), December 2010.

A.6 ECIR 2011 Evaluating Personal Search Workshop

The following proceedings are those of the ECIR 2011 Evaluating Personal Search Workshop. At the time of writing (September 2011), a write-up of the outcomes of this workshop is due to appear later this year. These provide a good insight into where this emerging space is currently at.

- D. Elsweiler, L. Kelly and J. Kim (editors). "Proceedings of the ECIR 2011 Workshop on Evaluating Personal Search". Dublin, Ireland, April 2011.

BIBLIOGRAPHY

[Ainsworth et al., 1993] Ainsworth, B. E., Haskell, W. L., Leon, A. S., Jacobs, D. R., Montoye, H. J., Sallis, J. F., and Paffenbarger, R. S. (1993). Compendium of physical activities: classification of energy costs of human physical activities. *Med Sci Sports Exerc.*, 25(1):71–80.

[Ainsworth et al., 2000] Ainsworth, B. E., Haskell, W. L., White, M. C., Irwin, M. L., Swartz, A. M., Strath, S. J., O'Brien, W. L., Bassett, D. R., Schmitz, K. H., Emplaincourt, P. O., Jacobs, D. R., and Leon, A. S. (2000). Compendium of physical activities: an update of activity codes and MET intensities. *Med Sci Sports Exerc.*, 32(9 Suppl):498–504.

[Aizawa et al., 2004] Aizawa, K., Tancharoen, D., Kawasaki, S., and Yamasaki, T. (2004). Efficient Retrieval of Life Log Based on Context and Content. In *The 1st ACM Workshop on Capture, Archival and Retrieval of Personal Experiences (CARPE 2004)*, pages 22–31. New York, New York, USA.

[Andre et al., 2006] Andre, D., Pelletier, R., Farringdon, J., Safier, S., Talbott, W., Stone, R., Vyas, N., Trimble, J., Wolf, D., Vishnubhatle, S., Boehmke, S., Stivoric, J., and Teller, A. (2006). The Development of the SenseWear armband, a Revolutionary Energy Assesment Device to Assess Physical Activity and Lifestyle. Technical report, BodyMedia, Inc.

[Anttonen, 2002] Anttonen, J. (2002). Electrophysiologically Interactive Computer Systems. *Computer*, 35(3):60–65.

[Anttonen and Surakka, 2005] Anttonen, J. and Surakka, V. (2005). Emotions and heart rate while sitting on a chair. In *CHI'05 Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 491–499, New York, NY, USA. ACM.

[Arapakis et al., 2009] Arapakis, I., Konstas, I., and Jose, J. (2009). Using Facial Expressions and Peripheral Physiological Signals as Implicit Indicators of Topical Relevance. In *Proceedings of the ACM International Conference on Multimedia*, pages 461–470, New York, NY, USA. ACM Press.

[Azzopardi et al., 2007] Azzopardi, L., de Rijke, M., and Balog, L. (2007). Building simulated queries for known-item topics: an analysis using six european languages. In *Proceedings of the 30th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR'07).*

[Baecker et al., 2007] Baecker, R. M., Marziali, E., Chatland, S., Easley, K., Crete, M., and Yeung, M. (2007). Multimedia Biographies for Individuals with Alzheimer's Disease and Their Families. In *Second International Conference on Technology and Aging*. Toronto, Canada.

[Bailey et al., 2003] Bailey, P., Craswell, N., and Hawking, D. (2003). Engineering a Multi-purpose Test Collection for Web Retrieval Experiments. *Information Processing and Management: an International Journal*, 39(6):853–871.

[Balabanovic, 1998] Balabanovic, M. (1998). An Interface for Learning Multi-topic User Profiles from Implicit Feedback. In *AAAI-98 Workshop on Recommender Systems*. Madison, Wisconsin.

[Bannon, 2006] Bannon, L. J. (2006). Forgetting as a feature, not a bug: the duality of memory and implications for ubiquitous computing. *CoDesign*, 2(1):3–15.

[Barreau et al., 2009] Barreau, D., Teevan, J., and Gwizdka, J., editors (2009). *Proceedings of The Fourth Personal Information Management Workshop (PIM 2009)*.

[Belkin et al., 2000] Belkin, N. J., Cool, C., Head, J., Jeng, J., Kelly, D., Lobash, S. L. L., Park, S., Savageknepshield, P., and Sikora, C. (2000). Relevance Feedback versus Local Context Analysis as Term Suggestion Devices: Rutgers' TREC-8 Interactive Track Experience. In *Proceedings of the 8th Text Retrieval Conference*, pages 565–574.

[Belkin et al., 1995] Belkin, N. J., Kantor, P., Fox, E. A., and Shaw, J. A. (1995). Combining the evidence of multiple query representations for information retrieval. *International Journal of Information Processing and Management*, 31(3).

[Bell, 2004] Bell, G. (2004). Challenges in Using Lifetime Personal Information Stores based on MyLifeBits. *Alpbach Forum*.

[Bell and Gemmell, 2007] Bell, G. and Gemmell, J. (2007). A Digital Life. *Scientific American*.

[Bell and Gemmell, 2009] Bell, G. and Gemmell, J. (2009). *Total Recall - How the E-Memory Revolution Will Change Everything*. Dutton.

[Berry et al., 2007] Berry, E., Kapur, N., Williams, L., Hodges, S., Watson, P., Smyth, G., Srinivasan, J., Smith, R., Wilson, B., and Wood, K. (2007). The use of a wearable camera, SenseCam, as a pictorial diary to improve autobiographical memory in a patient with limbic encephalitis. *Neuropsychological Rehabilitation*, 17(4):582–601.

[Black et al., 1996] Black, A. E., Coward, W. A., Cole, T. J., and Prentice, A. M. (1996). Human energy expenditure in affluent societies: analysis of 574 doubly labelled water measurements. *Eur. J. Clin. Nutr.*, 50:72–92.

[Blanc-Brude and Scapi, 2007] Blanc-Brude, T. and Scapi, D. L. (2007). What do people recall about their Documents? Implications for Desktop Search Tools. In *IUI'07*. Honolulu, Hawaii, USA.

[Blum et al., 2006] Blum, M., Pentland, A., and Troster, G. (2006). InSense: Interest-Based Life Logging. *IEEE Multimedia*, 13(4):40–48.

[Bradley et al., 2001a] Bradley, M. M., Codispoti, M., Cuthbert, B. N., and Lang, P. J. (2001a). Emotion and Motivation I: Defensive and Appetitive Reactions in Picture Processing. *American Psychological Association Inc.*, 1(3):276–298.

[Bradley et al., 2001b] Bradley, M. M., Codispoti, M., Cuthbert, B. N., and Lang, P. J. (2001b). Emotion and Motivation II: Sex Differences in Picture Processing. *American Psychological Association Inc.*, 1(3):300–319.

[Brockway, 1987] Brockway, J. M. (1987). Derivation of formulae used to calculate energy expenditure in man. *Hum. Nutr. Clin. Nutr.*, 41C:463–472.

[Brusilovsky, 1996] Brusilovsky, P. (1996). Methods and Techniques of Adaptive Hypermedia. *Journal of User Modelling and User adapted Interaction*, 6(2-3):87–129.

[Bush, 1945] Bush, V. (1945). As We May Think. *The Atlantic Monthly*, pages 101–108.

[Byrne et al., 2010] Byrne, D., Kelly, L., and Jones, G. (2010). *Multiple multimodal mobile devices: Lessons Learned from Engineering Lifelog Solutions*. IGI Publishing.

[Byrne et al., 2007] Byrne, D., Lavelle, B., Doherty, A., Jones, G. J. F., and Smeaton, A. F. (2007). Using Bluetooth and GPS Metadata to Measure Event Similarity in SenseCam Images. In *IMAI'07 - 5th International Conference on Intelligent Multimedia and Ambient Intelligence*, pages 18–24. Salt Lake City, Utah, USA.

[Cai et al., 2004] Cai, D., He, X., Wen, J., and Ma, W. (2004). Block-level link analysis. In *Proceedings of the 27th annual international conference on Research and development in information retrieval (SIGIR'04)*, pages 440–447. Sheffield, South Yorkshire, UK, ACM Press.

[Canini et al., 2010] Canini, L., Gilroy, S., Cavazza, M., Leonardi, R., and Benini, S. (2010). Users' Response to Affective Film Content: a Narrative Perspective. In *Proceedings of the 8th International Workshop on Content-Based Multimedia Indexing*, pages 50–55. Grenoble.

[Carmel et al., 2003] Carmel, D., Maarek, Y. S., Mandelbrod, M., Mass, Y., and Soffer, A. (2003). Searching XML Documents via XML Fragments. In *Proceedings of the 26th annual internation ACM SIGIR conference on research and developments in information retrieval*, pages 151–158, New York, NY, USA. ACM Press.

[Chen and Segall, 2009] Chen, X. Y. and Segall, Z. (2009). Xv-pod: An emotion aware, affective mobile video player. In *WRI World Congress on Computer Science and Information Engineering*, pages 277–281.

[Chirita, 2007] Chirita, P. (2007). *Emerging Applications of Link Analysis for Ranking*. PhD thesis, Universitat Hannover, Hannover, Germany.

[Chirita et al., 2006] Chirita, P., Ghita, S., Nejdl, W., and Paiu, R. (2006). Beagle++: Semantically Enhanced Searching and Ranking on the Desktop. In *The 3rd European Semantic Web Conference (ESWC)*. Budva, Montenegro.

[Chirita and Nejdl, 2006] Chirita, P. and Nejdl, W. (2006). Analyzing User Behavior to Rank Desktop Items. In *The 13th International Symposium on String Processing and Information Retrieval (SPIRE)*. Glasgow, United Kingdom.

[Chowdhury et al., 2003] Chowdhury, A., Aljlayl, M., Jensen, E., Beitzel, S., Grossman, D., and Frieder, O. (2003). Linear Combinations Based on Document Structure and Varied Stemming for Arabic Retrieval. In *TREC 2002, The Eleventh Text REtrieval Conference*.

[Cleverdon and Keen, 1966] Cleverdon, C. and Keen, E. (1966). Factors Determining the Performance of Indexing Systems (Volume 1: Design; Volume 2: Results). Technical report, Cranfield, UK.

[Cole et al., 2004] Cole, P. J., LeMura, L. M., Klinger, T. A., Strohecker, K., and McConnell, T. R. (2004). Measuring Energy Expenditure In Cardiac Patients Uwing The BodyMedia Armband Versus Indirect Calorimetry. A Validaton Study. *Journal of Sports Medicine and Physical Fitness*, 44(3):264–271.

[Craswell and Hawking, 2003] Craswell, N. and Hawking, D. (2003). Overview of the TREC-2002 Web Track. In *The Eleventh Text REtrieval Conference*.

[Craswell et al., 2005a] Craswell, N., Robertson, S., Zaragoza, H., and Taylor, M. (2005a). Relevance Weighting for Query Independent Evidence. In *Proceedings of the Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 416–423.

[Craswell and Vries, 2005] Craswell, N. and Vries, A. P. D. (2005). Overview of the TREC-2005 Enterprise Track. In *The Fourteenth Text REtrieval Conference*.

[Craswell et al., 2005b] Craswell, N., Zaragoza, H., and Robertson, S. (2005b). Microsoft Cambridge at TREC-14: Enterprise track. In *TREC-2005: The Fourteenth Text REtrieval Conference*.

[Croft, 2009] Croft, B. (2009). Query Evolution. In *Keynote at ECIR 2009*.

[Cutrell et al., 2006a] Cutrell, E., Dumais, S. T., and Teevan, J. (2006a). Searching to Eliminate Personal Information Management. *Communications of the ACM. Personal Information Management*, 49(1):58–64.

[Cutrell et al., 2006b] Cutrell, E., Robbins, D. C., Dumais, S. T., and Sarin, R. (2006b). Fast, Flexible Filtering with Phlat - Personal Search and Organization Made Easy. In *CHI 2006: Conference companion on Human factors in computing systems*, pages 261–270, New York, NY, USA. Montreal, Quebec, Canada, ACM Press.

[Denzer and Young, 2003] Denzer, C. M. and Young, J. C. (2003). The Effect of Resistance Exercise on the Thermic Effect of Food. *Internation Journal of Sport Nutrition and Exercise Metabolism (IJSNEM)*, 13(3):396–402.

[Doherty and Smeaton, 2008] Doherty, A. and Smeaton, A. (2008). Combining face detection and novelty to identify important events in a visual lifelog. In *Workshop on Image- and Video-based Pattern Analysis and Applications, at the 8th International Conference on Computer and Information Technology*.

[Doherty et al., 2007] Doherty, A., Smeaton, A., Lee, K., and Ellis, D. (2007). Multimodal segmentation of lifelog data. In *RIAO 2007 - Large-Scale Semantic Access to Content (Text, Image, Video and Sound)*.

[Dumais et al., 2003] Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R., and Robbins, D. C. (2003). Stuff I've seen: a system for personal information retrieval and re-use. In *SIGIR '03: The 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 72–79, New York, NY, USA. Toronto, Canada, ACM Press.

[Ellis and Lee, 2006] Ellis, D. P. W. and Lee, K. (2006). Accessing Minimal-Impact Personal Audio Archives. *IEEE Multimedia*, 13(4):30–38.

[Elsweiler et al., 2010] Elsweiler, D., Jones, G., Kelly, L., and Teevan, J. (2010). Report on Desktop Search Workshop. *SIGIR Forum*, 44(2):28–34.

[Elsweiler et al., 2011] Elsweiler, D., Kelly, L., and Kim, J., editors (2011). *Proceedings of The ECIR 2011 Evaluating Personal Search Workshop*.

[Elsweiler et al., 2005] Elsweiler, D., Ruthvan, I., and Jones, C. (2005). Dealing with Fragmented Recollection of Context in Information Management. In *Context-Based Information Retrieval (CIR-05) Workshop in Fifth International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT-05)*.

[Elsweiler and Ruthven, 2007] Elsweiler, D. and Ruthven, I. (2007). Towards task-based personal information management evaluations. In *Proceedings of the 30th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR '07)*.

[Elsweiler et al., 2007] Elsweiler, D., Ruthven, I., and Jones, C. (2007). Towards Memory Supporting Personal Information Management Tools. *Journal of the American Society for Information Science and Technology*.

[Ferry et al., 1999] Ferry, B., Roozendaal, B., and McGaugh, J. (1999). Basolateral Amygdala Noradrenergic Influences on Memory Storage Are Mediated by an Interaction between beta- and alpha1-Adrenoceptors. *Journal of Neuroscience*, 19(12):5119–5123.

[Freeman and Gelernter, 1996] Freeman, E. and Gelernter, D. (1996). Lifestreams: A Storage Model for Personal Data. *SIGMOD Record*, 25(1):80–86.

[Fruin and Rankin, 2004] Fruin, M. L. and Rankin, J. W. (2004). Validity of a Multi-Sensor Armband in Estimating Rest and Exercise Energy Expenditure. *Medicine and Science in Sports and Exercise*, 36(6):1063–1069.

[Fuller et al., 2008] Fuller, M., Kelly, L., and Jones, G. J. F. (2008). Applying Contextual Memory Cues for Retrieval from Personal Information Archives. In *PIM 2008 - Proceedings of Personal Information Management, Workshop at CHI 2008*. Florence, Italy.

[Gazzaniga et al., 2002] Gazzaniga, M. S., Ivry, R. B., and Mangun, G. R. (2002). *Cognitive Neuroscience (Second Edition)*. Norton.

[Gemmell et al., 2006] Gemmell, J., Bell, G., and Lueder, R. (2006). MyLifeBits: A Personal Database for Everything. *Communications of the ACM. Personal Information Management*, 49(1):88–95.

[Gemmell et al., 2002] Gemmell, J., Bell, G., Lueder, R., Drucker, S., and Wong, C. (2002). MyLifeBits: Fulfilling the Memex Vision. In *ACM Multimedia '02*. Juan Les Pins, France.

[Gemmell and Sundaram, 2004] Gemmell, J. and Sundaram, H., editors (2004). *Proceedings of The First ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE 2004)*.

[Gemmell and Sundaram, 2005] Gemmell, J. and Sundaram, H., editors (2005). *Proceedings of The Second ACM Workshop on Continuous Archival and Retrieval of Personal Experiences (CARPE 2005)*.

[Gomes et al., 2010] Gomes, P., Gama, S., and Gonçalves, D. (2010). Using Personal Information Visualization for Document Retrieval. In *Proceedings of the SIGIR 2010 Desktop Search Workshop*, pages 17–20.

[Gurrin et al., 2008] Gurrin, C., Smeaton, A. F., Byrne, D., O'Hare, N., Jones, G. J. F., and O'Connor, N. (2008). An Examination of a Large Visual Lifelog. In *4th Asia Information Retrieval Symposium (AIRS)*, pages 537–542. Springer-Verlag.

[Harper et al., 2008] Harper, R., Randall, D., Smyth, N., L., C. E., Heledd, and Moore, R. (2008). The past is a different place: they do things differently there. In *Proceedings of the 7th ACM conference on Designing interactive systems*, pages 271–280.

[Hettema et al., 2000] Hettema, J., Leidelmeijer, K., and Greenen, R. (2000). Dimensions of information processing: physiological reactions to motion pictures. *Eur. J. Personality*, 14(1):39–63.

[Hori and Aizawa, 2003] Hori, T. and Aizawa, K. (2003). Context-based Video Retrieval System for the Life-Log Applications. In *The 5th ACM SIGMM International Workshop on Multimedia Information Retrieval, MIR 2003*, pages 31–38. Berkeley, California, USA.

[Jaimes et al., 2004] Jaimes, A., Omura, K., Nagamine, T., and Hirata, K. (2004). Memory Cues for Meeting Video Retrieval. In *CARPE'04*, pages 74–85. New York, New York, USA.

[Jain and Ross, 2004] Jain, A. K. and Ross, A. (2004). Multibiometric systems. *Communications of the ACM*, 47:34–40.

[Jakicic et al., 2004] Jakicic, J. M., Marcus, M., Gallagher, K., Randall, C., Thomas, E., Goff, F. L., and Robertson, R. J. (2004). Evaluation of the SenseWear Pro Armband to Assess Energy Expenditure During Exercise. *Medicine and Science in Sports and Exercise*, 36(5):897–904.

[Joki et al., 2007] Joki, A., Burke, J., and Estrin, D. (2007). Campaignr - a framework for participatory data collection on mobile phones. Technical report, Centre for Embedded Network Sensing, University of California, Los Angeles.

[Jones, 2006] Jones, W., editor (2006). *Proceedings of The Second Personal Information Management Workshop (PIM 2006)*.

[Kang and Kim, 2003] Kang, I. and Kim, G. (2003). Query type classification for web document retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval (SIGIR'03)*, pages 64–71. ACM Press.

[Karger et al., 2005] Karger, D., Bakshi, K., Huynh, D., Quan, D., and Sinha, V. (2005). Haystack: A Customizable General-Purpose Information Management Tool for End Users of Semistructured Data. In *CIDR 2005. Conference on Innovative Data Systems Research*.

[Kataoka et al., 1998] Kataoka, H., Kano, H., Yoshida, H., Saijo, H., Yasuda, M., and Osumi, M. (1998). Development of a Skin Temperature Measuring System for Non-contact Stress Evaluation. *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2:940–943.

[Kelly et al., 2008] Kelly, L., Chen, Y., Fuller, M., and Jones, G. J. F. (2008). A study of remembered context for information access from personal digital archives. In *2nd International Symposium on Information Interaction in Context (IIiX)*, pages 44–50. British Computer Society.

[Kelly and Jones, 2009] Kelly, L. and Jones, G. J. F. (2009). Examining the utility of affective response in search of personal lifelogs. In *5th Workshop on Emotion in HCI, British HCI Conference 2009*.

[Kelly and Jones, 2010a] Kelly, L. and Jones, G. J. F. (2010a). Biometric response as a source of query independent scoring in lifelog retrieval. In *Proceedings of 32nd European Conference on Information Retrieval (ECIR 2010)*, pages 520–531. Springer-Verlag.

[Kelly and Jones, 2010b] Kelly, L. and Jones, G. J. F. (2010b). An exploration of the utility of gsr in locating events from personal lifelogs for reflection. In *4th Irish Human Computer Interaction Conference (iHCI2010)*.

[Kim and Andre, 2008a] Kim, J. and Andre, E. (2008a). Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[Kim and Andre, 2008b] Kim, J. and Andre, E. (2008b). Multi-channel biosignal analysis for automatic emotion recognition. In *International Conference on Bio-inspired Systems and Singal Processing (Biosignals 2008)*.

[Kim and Croft, 2009] Kim, J. and Croft, W. B. (2009). Retrieval experiments using pseudo-desktop collections. In *In Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*, pages 1297–1306. ACM.

[Kim et al., 2009] Kim, J., Xue, X., and Croft, W. B. (2009). A probabilistic retrieval model for semistructured data. In *In Proceedings of the 31st European Conference on Information Retrieval (ECIR 2009)*, pages 228–239. Springer.

[Kim et al., 2004] Kim, K. H., Bang, S. W., and Kim, S. R. (2004). Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, 42:419–427.

[King et al., 2005] King, G. A., Deemer, S. E., Franco, B. M., Potter, C., and Coleman, K. J. (2005). Accuracy of Three Physical Activity Monitors to Measure Energy Expenditure During Activities of Daily Living. *Medicine and Science in Sports and Exercise*, 37(5).

[Klein et al., 2002] Klein, J., Moon, Y., and Picard, R. W. (2002). This computer responds to user frustratioin: Theory, design, and results. *Interacting with Computers*, 14(2):119–140.

[Kleinberg, 1999] Kleinberg, J. M. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–362.

[Kraaij et al., 2002] Kraaij, W., Westerveld, T., and Hiemstra, D. (2002). The importance of prior probabilities for entry page search. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'02)*, pages 27–34. ACM Press.

[Kurland, 2006] Kurland, O. (2006). *Inter-Document Similarities, Language Models, and ad hoc Information Retrieval*. PhD thesis, Department of Computer Science, Cornell University.

[Kurland and Lee, 2006] Kurland, O. and Lee, L. (2006). Respect My Authority! HITS Without Hyperlinks, Utilizing Cluster-Based Language Models. In *29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*. Seattle, Washington, USA.

[Lalmas, 2009] Lalmas, M. (2009). XML Retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–111.

[Lang, 1995] Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist*, 50(5):372–385.

[Lang et al., 1993] Lang, P. J., Greenwald, M. K., Bradley, M. M., and Hamm, A. O. (1993). Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30:261–273.

[Lavelle et al., 2007] Lavelle, B., Byrne, D., Gurrin, C., Smeaton, A. F., and Jones, G. J. F. (2007). Bluetooth Familiarity: Methods of Calculation,Applications and Limitations. In *MIRW 2007 - Mobile Interaction with the Real World, Workshop at the MobileHCI07: 9th International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 55–58.

[Lieberman, 1997] Lieberman, H. (1997). Autonomous Interface Agents. In *Conference proceedings on Human factors in computing systems*. Atlanta, GA, USA.

[Lisetti et al., 2003] Lisetti, C., Nasoz, F., LeRouge, C., Ozyer, O., and Alvarez, K. (2003). Developing Multimodal Intelligent Affective Interfaces for Tele-Home Health Care. *International Journal of Human-Computer Studies*.

[Lisetti and Nasoz, 2004] Lisetti, C. L. and Nasoz, F. (2004). Using Noninvasive Wearable Computers to Recognize Human Emotions from Physiological Signals. *EURASIP Journal on applied Signal Processing, Hindawi Publishing Corporation*, pages 1672–1687.

[Luhn, 1957] Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 4:309–317.

[Maltzman and Boyd, 1984] Maltzman, I. and Boyd, G. (1984). Stimulus significance and bilateral scrs to potentially phobic pictures. *Journal of Abnormal Psychology*, 93:41–46.

[Manning et al., 2009] Manning, C., Raghavan, P., and Schutze, H. (2009). *An Introduction to Information Retrieval*. Cambridge University Press.

[Mase, 2006] Mase, K., editor (2006). *Proceedings of The Third ACM Workshop on Capture, Archival and Retrieval of Personal Experiences (CARPE 2006)*.

[Mase et al., 2006] Mase, K., Sumi, Y., Toriyama, T., Tsuchikawa, M., Ito, S., Iwasawa, S., Kogure, K., and Hagita, N. (2006). Ubiquitous Experience Media. *IEEE Multimedia*, 13(4):20–29.

[McGaugh, 2003] McGaugh, J. (2003). *Strong memories are made of this Memory and Emotion: The Making of Lasting Memories*. Columbia University Press.

[Mealey et al., 2007] Mealey, A. D., Jakicic, J. M., Mealey, L. M., Davis, K. K., and McDermott, M. D. (2007). Validation of the SenseWear Pro Armband to Estimate Energy Expenditure during a Simulation of Daily Activity. In *Abstract presented at the Annual meeting of the American College of Sports Medicine*. New Orleans, LA, USA.

[Mooney et al., 2006] Mooney, C., Scully, M., Jones, G. J. F., and Smeaton, A. F. (2006). Investigating Biometric Response for Information Retrieval Applications. In *The 28th European Conference on Information Retrieval (ECIR 2006)*, pages 570–574, Berlin Heidelberg. London, UK, Springer-Verlag.

[Nack, 2005] Nack, F. (2005). You Must Remember This. *IEEE Multimedia*, 12:4–7.

[Nakayama et al., 1977] Nakayama, T., Ohnuki, Y., and Niwa, K. (1977). Fall in Skin Temperature During Exercise. *Jpn J Physiol.*, 27(4):423–37.

[Nichols, 1997] Nichols, D. (1997). Implicit Rating and Filtering. In *DELOS Workshop on Filtering and Collaborative Filtering*. Budapest, Hungary.

[Norman, 1998] Norman, D. A. (1998). *The Design of Everyday Things*. The MIT Press, London, England.

[Ogilvie and Callan, 2003] Ogilvie, P. and Callan, J. (2003). Combining Document Representations for Known Item Search. In *Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA. ACM Press.

[O'Hare et al., 2005] O'Hare, N., Gurrin, C., Jones, G. J. F., and Smeaton, A. F. (2005). Combination of Content Analysis and Context Features for Digital Photograph Retrieval. In *2nd IEE European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*. London, U.K.

[O'Hare et al., 2006] O'Hare, N., Lee, H., Cooray, S., Gurrin, C., Jones, G. J. F., Malobabic, J., O'Connor, N. E., Smeaton, A. F., and Uscilowski, B. (2006). MediAssist: Using Content-Based Analysis and Context to Manage Personal Photo Collections. In *CIVR2006 - 5th International Conference on Image and Video Retrieval*, pages 529–532. Springer-Verlag.

[Page et al., 1998] Page, L., Brin, S., Motwani, R., and Winograd, T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford InfoLab.

[Partala and Surakka, 2004] Partala, T. and Surakka, V. (2004). The effects of affective interventions in human-computer interaction. *Interacting with Computers*, 16(2):295–309.

[Pérez-Iglesias et al., 2009] Pérez-Iglesias, J., Pérez-Agüera, J. R., Fresno, V., and Feinstein, Y. Z. (2009). Integrating the Probabilistic Models BM25/BM25F into Lucene. *CoRR*, abs/0911.5046.

[Picard, 2000] Picard, R. W. (2000). Towards computers that recognize and respond to user emotion. *IBM Systems Journal*, 39(3.4):705–719.

[Picard et al., 2001] Picard, R. W., Vyzas, E., and Healey, J. (2001). Towards Machine Emotional Intelligence: Analysis of Affective Physiological State. *IEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191.

[Richardson et al., 2006] Richardson, M., Prakash, A., and Brill, E. (2006). Beyond PageRank: Machine Learning for Static Ranking. In *Proceedings of the International World Wide Web Conference*, pages 707–715. Edinburgh, Scotland, ACM Press.

[Ringel et al., 2003] Ringel, M., Cutrell, E., Dumais, S., and Horvitz, E. (2003). Milestones in Time: The Value of Landmarks in Retrieving Information from Personal Stores. In *Interact 2003*.

[Robertson and Sparck-Jones, 1976] Robertson, S. E. and Sparck-Jones, K. (1976). Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*, 27(3):129–146.

[Robertson et al., 1992] Robertson, S. E., Walker, S., Hancock-Beaulieu, M., Gull, M., and Lau, M. (1992). Okapi at TREC. In *Proceedings of the First Text REtrieval Conference (TREC)*, pages 21–30. Gaithersburg, MD. NIST.

[Robertson et al., 1993] Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. (1993). Okapi at TREC-2. In *Proceedings of the Second Text Retrieval Conference (TREC-2)*, pages 21–34. Gaithersburg, MD. NIST.

[Robertson et al., 2004] Robertson, S. E., Zaragoza, H., and Taylor, M. (2004). Simple BM25 extension to multiple weighted fields. In *13th ACM International Conference on Information and Knowledge Management*, pages 42–49, Washington D.C., U.S.A.

[Rothwell et al., 2006] Rothwell, S., Lehane, B., Chan, C. H., Smeaton, A. F., O'Connor, N. E., Jones, G. J. F., and Diamond, D. (2006). The cdvplex biometric cinema: Sensing physiological responses to emotional stimuli in film. In *Adjunct Proceedings of Pervasive*.

[Ruthven, 2005] Ruthven, I. (2005). *Integrating approaches to relevance*, volume 19 of *Information Retrieval Series*, pages 61–80. Springer.

[S. E. Robertson and S. Jones, 1994] S. E. Robertson and S. Jones (1994). Simple, proven approaches to text retrieval. Technical Report 356, University of Cambridge.

[Sakamoto et al., 2006] Sakamoto, R., Nozawa, A., Tanaka, H., Mizuno, T., and Ide, H. (2006). Evaluation of the driver's temporary arousal level by facial skin thermogram-effect of surrounding temperature and wind on the thermogram. *IEEJ Trans. EIS*, 126(7):804–809.

[Salton and Buckley, 1988] Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, pages 513–523.

[Salton et al., 1975a] Salton, G., Wong, A., and Yang, C. S. (1975a). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

[Salton and Yang, 1973] Salton, G. and Yang, C. S. (1973). On the specification of term values in automatic indexing. *Journal of Documentation*, 29(4):351–372.

[Salton et al., 1975b] Salton, G., Yang, C. S., and Yu, C. T. (1975b). A theory of term importance in automatic text analysis. *Journal of the ASIS*, 26(1):33–44.

[Savoy and Rasolofo, 2003] Savoy, J. and Rasolofo, Y. (2003). Report on TREC-11 experiments: Arabic, Named Page and Topic Distillation searches. In *TREC 2002, The Eleventh Text REtrieval Conference*.

[Scheirer et al., 2002] Scheirer, J., Fernandez, R., Klein, J., and Picard, R. W. (2002). Frustrating the user on purpose: a step toward building an affective computer. *Interacting with Computers*, 14(2):93–118.

[Smeaton et al., 2007] Smeaton, A. F., Over, P., and Kraaij, W. (2007). Evaluation campaigns and TRECVid. In *The 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA. ACM Press.

[Smeaton and Rothwell, 2009] Smeaton, A. F. and Rothwell, S. (2009). Biometric Responses to Music-Rich Segments in Films: The CDVPlex. In *Seventh International Workshop on Content-Based Multimedia Indexing*, pages 162–168.

[Soleymani et al., 2008] Soleymani, M., Chanel, G., Kierkels, J. J., and Pun, T. (2008). Affective ranking of movie scenes using physiological signals and content analysis. In *MS '08: Proceedings of the 2nd ACM workshop on Multimedia Semantics*, pages 32–39, New York, NY, USA. ACM Press.

[Soules, 2006] Soules, C. A. N. (2006). *Using context to assist in personal file retrieval*. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA.

[Soules and Ganger, 2005] Soules, C. A. N. and Ganger, G. R. (2005). Connections: Using Context to Enhance File Search. In *20th ACM Symposium on Operating Systems Principles (SOSP'05)*, pages 119–132. Brighton, United Kingdom.

[Sparck-Jones, 1972] Sparck-Jones, K. (1972). A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*, 28(1):11–20.

[St-Onge et al., 2007] St-Onge, M., Mignault, D., Allison, D. B., and Rabasa-Lhoret, R. (2007). Evaluation of a portable device to measure daily energy expenditure in free-living adults1-3. *Am J Clin Nutr*, 85:742–749.

[Stumpf et al., 2005] Stumpf, S., Bao, X., Dragunov, A., Dietterich, T. G., Herlocker, J., Johnsrude, K., Li, L., and Shen, J. (2005). Predicting User Tasks: I Know What You're Doing! In *20th National Conference on Artificial Intelligence (AAAI-05), Workshop on Human Comprehensible Machine Learning*. Pittsburgh, PA.

[Tancharoen et al., 2005] Tancharoen, D., Yamasaki, T., and Aizawa, K. (2005). Practical Experience Recording and Indexing of Life Log Video. In *The 2nd ACM Workshop on Capture, Archival and Retrieval of Personal Experiences (CARPE 2005)*. Singapore.

[Teevan et al., 2004] Teevan, J., Alvarado, C., Ackerman, M. S., and Karger, D. R. (2004). The Perfect Search Engine Is Not Enough: A Study of Orienteering Behavior in Directed Search. In *CHI 2004: Conference companion on Human factors in computing systems*, pages 415–422. Vienna, Austria.

[Teevan and Jones, 2008] Teevan, J. and Jones, W., editors (2008). *Proceedings of The Third Personal Information Management Workshop (PIM 2008)*.

[Torii et al., 1992] Torii, M., Yamasaki, M., Sasaki, T., and Nakayama, H. (1992). Fall in Skin Temperature of Exercising Man. *Br J Sports Med*, 26(1):29–32.

[Upstill, 2004] Upstill, T. (2004). *Document ranking using web evidence*. PhD thesis, Australian National University, Australia.

[Upstill et al., 2003] Upstill, T., Craswell, N., and Hawking, D. (2003). Query-independent evidence in home page finding. *ACM Transactions on Information Systems*, 21(3):286–313.

[van Rijsbergen, 1979] van Rijsbergen, C. J. (1979). *Information Retrieval, 2nd Edition*. Butterworth-Heinemann, Newton, MA, US.

[W. Boucsein, 1992] W. Boucsein (1992). *Electrodermal activity*. Plenum Press, New York.

[Ward et al., 2002] Ward, R. D., Cahill, B., Marsden, P. H., and Johnson, C. A. (2002). Physiological Responses to HCI Events - what produces them and how detectable are they? In *Proceedings of HCI 2002*, pages 90–93.

[Ward and Marsden, 2003] Ward, R. D. and Marsden, P. H. (2003). Physiological Responses to Different WEB page designs. *International Journal of Human-Computer Studies*, 59(1/2):199–212.

[Wilkinson, 1994] Wilkinson, R. (1994). Effective Retrieval of Structured Documents. In *Research and Developments in Information Retrieval*, pages 311–317.

[Xu et al., 2003] Xu, H., Yang, Z., Wang, B., Liu, B., Cheng, J., Liu, Y., Yang, Z., Cheng, X., and Bai, S. (2003). TREC-11 experiments at CAS-ICT: Filtering and Web. In *TREC 2002, The Eleventh Text REtrieval Conference*.

[Zaragoza et al., 2004] Zaragoza, H., Craswell, N., Taylor, M., Saria, S., and Robertson, S. (2004). Microsoft Cambridge at TREC-13: Web and HARD tracks. In *TREC 2004, The Thirteenth Text REtrieval Conference.*