Title	Probabilistic metadata generation for places based on user data
Author(s)	Hegde, Vinod Kumar Gajanana
Publication Date	2016-08-16
Item record	http://hdl.handle.net/10379/5985

Probabilistic Metadata Generation for Places Based on User Data

Vinod Kumar Gajanana Hegde

Insight Centre for Data Analytics

College of Engineering and Informatics

National University of Ireland, Galway, Ireland

Ph.D. Thesis

16th August 2016

Galway, Ireland

SUPERVISOR:

Dr. Matthias Nickles

CO-SUPERVISOR:

Dr. Alessandra Mileo

INTERNAL EXAMINER:

Dr. Colm O'Riordan

EXTERNAL EXAMINER:

Prof. Dr. Pedro José Marrón

Abstract

In recent years, there has been a wide adoption of mobile devices such as smart phones and tablets. This wide adoption is supported by numerous mobile and Web applications which help users to consume and generate data on the go. Users generate large volumes of data using these applications which represent real time contextual information about them. Most of the current mobile and Web applications analyse user data such as social interests and physical presence of users at places to deliver better services and user experience in applications. However, studies have shown that spatial databases lack sufficient metadata for places as users are required to manually provide this information. Since this is time consuming work, users rarely annotate places in spite of having knowledge about them. Automatically generating annotations for places by exploiting user generated data on mobile and Web applications can potentially be used to overcome the lack of metadata for places. Rich metadata about places can be used by geospatial web services and location based services to provide accurate results.

Automatic generation of place metadata requires new sophisticated data mining algorithms. This thesis focuses on unsolved questions regarding the utilization of physical presence and social data of users to generate metadata for places. Specifically, we have developed probabilistic models and text processing algorithms for short text snippet or tag generation for locations using social interest profiles and check-ins of users at places. Then, we have studied how only the user presence data at places can be used to infer real world events at those places. To this end, we discuss a probabilistic outlier detection model and an algorithm to detect any unusual presence of huge crowds at places. We have then defined and implemented an approach to generate tags by analysing textual data generated during events conducted at locations. We have evaluated all the discussed models and algorithms with

both synthetic and real world data. Our experiments show that rich metadata for places can be derived by analysing user generated data.

Declaration

This dissertation is the result of my own work, except where explicit reference is made to the work of others, and has not been submitted for another qualification to this or any other university. This dissertation adheres to the rules and regulations that exist in the National University of Ireland, Galway and has taken all the suggestions from the respective Degree Committee.

Vinod Kumar Gajanana Hegde

Acknowledgements

First of all, I want to thank Prof. Dr. Manfred Hauswirth. Manfred has been the supervisor till the end of third year of my PhD studies. He always gave the freedom of picking the problems of interest to me and supported me during the research work. I also thank Dr. Josiane Xavier Parreira, who supported and motivated me by providing interesting results and the accompanying research techniques. I am grateful to Dr. Milovan Krnjajić who taught me two courses on probabilistic models and guided me in scientific publications. I have greatly benefited from the research discussions we have had during these years. I also learnt a lot from him while working as teaching assistant for the courses he taught. I have had the opportunity to learn about spatial data and analysis from Dr. Alexei Pozdnoukhov. I am grateful to him for the scientific discussions we had on spatial data analysis. I am thankful to Dr. Matthias Nickles for his valuable feedback. Matthias guided me as supervisor during the final phases of my studies and patiently guided me in structuring the thesis. I thank Dr. Alessandra Mileo for review of the chapters and detailed feedback. Further, I want to thank all the colleagues for the good times we always had during my stay in Galway.

I want to thank my parents Gajanana Hegde and Lakshmi Hegde for their immense love for me, the great confidence in me. I appreciate my wife Sahana's patience for accepting many weekends full of work and innumerable days where I had to go home very late from the lab. She has always been there with her love and support which enabled me to complete my doctoral research work.

Contents

x

Chapter 1

Introduction

Recent years have seen a huge rise in the number of users utilizing various services on the Web. This has led to an increase in the number of interesting Web applications which are diverse in the type of services they provide. At the same time the large number of mobile device users has prompted the creation of a vast number of mobile applications. Many of these mobile and Web applications address the creation and consumption of geospatial data. These geospatial applications can be classified as geospatial web services (GWS) and location based services (LBS) depending on whether they use real time location of a mobile device to deliver content. LBS are dependent on the real time physical location of a device which is obtained via global positioning system (GPS), cell triangulation method, and IP address analysis among other techniques. Geospatial applications enable users to contribute to the enrichment of geospatial databases by annotating geospatial objects with metadata. In this thesis, we specifically consider metadata for *places* or *locations* or *Points of Interests* (PoIs). Note that we use these terms interchangeably throughout the thesis.

Users enrich metadata in the geospatial databases by adding *geo-locations* or *places* to the maps, rating the places, annotating them with photos, videos, comments etc. Various types of metadata are used by geospatial applications mainly to provide accurate search and recommendation of places. Geospatial applications support search for places by place attributes such as name, category, distance from the user, rating etc. They also recommend famous places in real time, show the amount of people and predict the size of the crowd at places for various time periods. Location based services determine the physical location of users and enable users to express their visits to places via *check-ins* and share their experience. Personalised recommendations of places to users based on their place visiting patterns and social profiles are major

features of many location based services. There are numerous geospatial applications provided by commercial players such as Google[1], Facebook[2], FourSquare[3], Yelp[4] and many other players. There are also non-commercial projects like OpenStreetMap[5] (OSM), Wikimapia[6] among others. Most of the above mentioned geospatial application vendors rely on data provided by volunteers to build their own geospatial databases.

A huge increase in the number of applications critically dependent on geospatial data on both mobile devices and the Web have made the study of creation, storage and consumption of geospatial data crucial. It has been evident through various literature and research studies that the geospatial databases lack rich metadata. This motivates us to provide potential solutions to the problem of metadata generation for places, which we will present in this thesis. In what follows, we discuss the state-of-the-art in geospatial databases and major issues involved in creation and consumption of geospatial data. Specifically, we describe the current techniques of automatic and manual metadata generation for places that are adopted by major vendors. We analyse and discuss the sparsity of manually added metadata in these databases. We then discuss the challenges involved in automatically enriching geospatial databases, which sets the basis for our hypothesis and contributions.

Geospatial Databases

Geospatial databases are the databases that deal with creation, storage and retrieval of geospatial data. Geospatial data refers to any data about spatial objects that exist on the surface of the earth. Some of these spatial objects such as rivers, roads, and buildings can have well defined geographic boundaries. Places, which also represent spatial objects represent abstraction of a physical location and usually do not have well define boundary. Spatial objects can be defined by single or multiple location attributes. The most commonly used location attributes are latitude, longitude and altitude triplets to accurately identify spatial objects. The other common spatial attributes include textual address, ZIP codes. The term place is an abstract form of location as a place does not have exact geospatial boundaries defined, though a place exists on the surface of the earth.

---

[1]https://maps.google.ie/

[2]//www.facebook.com/places/

[3]https://foursquare.com/

[4]http://www.yelp.ie/la

[5]https://www.openstreetmap.org/

[6]http://wikimapia.org/

The geospatial data has been historically used for applications in mining, defence planning, natural resource mapping, environmental analysis, and weather prediction among others. The successful adoption of mobile devices in recent years by users has given rise to new ways of creation and consumption of geospatial data. The data about interaction and experience of users at places has become a major part of the geospatial databases. The extensive use of geospatial applications has increased the amount of geospatial data voluntarily contributed by users. In spite of these efforts, accurate metadata generation for places remains a challenge due to various technical and human limitations. We now discuss some of the research efforts and application techniques that have been employed to generate geospatial metadata. Our discussion includes both manual and automatic techniques for metadata generation.

The majority of the data in geospatial databases used by location based services and geospatial web services are contributed by volunteering users. These datasets are collectively termed as *Volunteered Geographic Information* (VGI). Mobile device users represent a major community of stakeholders in terms of geospatial data generation and consumption using both the mobile applications and Web applications on mobile devices. It has been noted that users provide geospatial data for various social and financial benefits [1, 2]. Note that populating geospatial databases demands extensive knowledge of the places under consideration. This demands utilizing the knowledge shared by residents and visitors of places. There has been extensive work on mining user generated data to derive metadata of various types for locations. The information in systems based on VGI is provided by locals and covers vernacular places and their names. Flanagin et al. [3] argues that credibility of such information is an issue and discusses how community based information generation can be useful. Wang et al. [4] discusses various approaches that can be adopted for manually tagging places using mobile phones. The semi-automatic approach to integrating such rich and timely information into the gazetteers has been discussed in Keßler et al. [5]. The work by Lin at al. [6] studies the naming preferences of people regarding the places they visit and shows that such preferences depend on the context of the person naming a place. It also finds that on an average places have very few description names (mean of number of place descriptions is 2.8).

Place tags can be classified into two categories based on their purpose. *Semantic tags* and *event tags* are the two major categories of metadata that can potentially be associated with a place. Semantic tags are short text snippets that semantically representative of a place. They are intended to describe the nature and abstract meaning

of a given place, that users can relate to. Event tags describe events that are held at places along with additional information of event such as start and end time, number of people at an event, purpose of event etc. There have been various efforts to derive the above mentioned metadata for places. Such efforts have successfully utilized information from location based services. Location based services at their core enable their users to broadcast visits to places via *check-ins*. Check-in data are descriptive of real time crowd movement at places along with experience of people at places. There have been other works that use information about geospatial coordinates (i.e., latitude and longitude) of user movements to infer categories of places. Similarly, sensor data captured about people's movements in buildings and road networks have been used to infer events at places. Now we describe some of the works that have exploited such user presence information to generate place metadata.

In Lian et al. [7], an automatic place naming technique based on user check-in activities is discussed. In Ye et al. [8], data analysis techniques to automatically derive the abstract category of location such as hotel, entertainment venue etc. have been presented. In Manasour et al. [9], places are modelled based on mentions of places in tweets to generate keywords or tags for places. The aim here is to derive those tags that closely align with query terms of place search queries by users. Commercial vendors are highly dependent on the manually curated VGI for populating their geospatial databases. They have successfully developed automatic metadata generation techniques based on data generated by mobile devices at various locations. The automatic metadata generation approaches adopted in OpenStreetMap [10] demand huge data about physical context of the mobile device users. OpenStreetMap uses the GPS traces of its mobile device users to create the road network maps for highly populated areas. Extraction of data from sensors such as GPS sensor, accelerometer etc. in mobile devices is one of the major challenges. Various solutions have been proposed in [11–14] to efficiently determine physical location of a user with minimal energy consumption. Foursquare uses the user check-ins to automatically derive the time periods during which a given location is actively visited by users. Pozdnoukhov et al. [15] show that events can be detected at geographic regions by analysing the number of geo-annotated micro-blog posts at a given time.

Through these examples, we can see that there is extensive work on metadata generation for locations so as to make geospatial applications more efficient and user friendly. Still, there are many challenges that need to be addressed to automatically generate geospatial metadata. For example, collecting the right type and right amount

of user data so as to run data mining algorithms on these data is a major challenge. Similarly, assessing the quality of automatically derived metadata when there is no manually generated ground truth poses another challenge. As discussed in the previous section, geospatial databases lack extensive metadata about locations due to various reasons. In this thesis, we mainly concentrate on automatically generating metadata for places so as to enrich geospatial databases. These metadata mainly in the form of tags, presence, duration of events at places etc. are crucially needed to build efficient location based services and geospatial web services mentioned earlier. In the following section, we describe the research questions and problems addressed by this thesis and , challenges that are involved in solving the stated problems.

1.1 Problem Statement

We address the problem of generating metadata for places by analysing place related user data captured on the Web. Fundamentally, the problem of place-metadata generation demands identifying appropriate data sets having references to locations and mining those data to derive metadata. Solving this problem needs addressing many fundamental research questions. The research questions we consider in order to solve the problem of automatic generation of place metadata, are as follows:

1. Which sources of data on the Web have to be chosen so that we can perform data mining on those to derive rich metadata about places?

2. How can we derive semantic tags which conceptually represent a place using these data sources?

3. How can we detect any events organized at a place by analysing the data about amount of crowd at that place?

The first research question involves many challenges few of which are listed below.

Representative nature of data

There are huge streams of data constantly being pushed or pulled to the Web from various sources such as online social networks, blogs, micro-blogs, sensors, mobile devices etc. Some of these data are related to places and can possibly be utilized for mining and subsequent generation of metadata about places. Some of these data such as presence information broadcast through smart phones are directly associated with places and are unique in their size as well as rate at which they are pushed onto the Web. In contrast, though sources such as blogs have data related to places, they are

highly unstructured, smaller in size and are not periodically updated on the Web. This shows that the sources of data to derive metadata about places have to be appropriately analyzed and chosen.

Availability of data

Though many of the Web and mobile application users often generate data related to places, this data is not available for analysis. The reasons include, but are not limited to, issues such as privacy, security and trust. For example, though a user wishes to express her visit to a place, she wants it to be visible only to her friends for privacy reasons. Similarly, social interests of a user can have restricted access to the public w.r.t visibility and availability. These factors limit any integrated analysis of her presence information combined with the social interests she has expressed on the online social network profile or a micro-blog. These observations indicate that even though a data source is rich in geospatial data, the availability of such data openly on the Web for analysis is a challenge.

The second question has these prominent challenges.

Quality of derived tags compared to manual tags

As stated earlier, commercial geospatial databases lack metadata for places especially in the form of semantic tags and event tags. This is due to the fact that such metadata has to be contributed manually by volunteers and or obtained by commercial crowdsourcing. Note that manually generated semantic metadata is always limited by the vocabulary of the user adding the semantic tags. Similarly, quality of event tags annotated by users is limited by their knowledge of place being annotated and number of active annotators. So, any approach to automatically generate place-metadata would immediately face the challenge of evaluating the quality of automatically derived metadata. Specifically, this is due to the lack of extensively assigned tags which can be used as ground truth.

Variability in the derived tags

Any approach to derive semantic tags automatically has to consider the time dependency of semantic tags on a place. Semantic concepts associated with a place though are stable for a short duration of time, can change based on any logistical changes in a place. For example, if an auditorium has a conference on *Biological Sciences*, the semantic tags are relevant only during the period of the conference and changes based on other events happening there. Similarly, if a new engineering department is established in a building complex of engineering schools, an additional set of tags will

turn out to be relevant for that building complex. This possible evolution of semantic tags for a location over time is a critical question that needs to be addressed by any semantic tag generation technique.

The challenges associated with the third question are as follows.

Criteria for event detection

Events of various types happen at places. Places differ from each other in the types of events organized, the frequency with which those events are organized and the amount of people present for events. Though at a small scale, users do report some of the events and places they happen at, using various Web and mobile applications. Any mechanism to detect and predict events at places has to take the above factors into consideration together. This demands advanced statistical models that can flexibly model crowd at places and detect events.

Accuracy of prediction

In the previous section, we discussed some of the challenges faced in acquiring geolocation related data. We can observe that there is a lack of extensive data about the presence information of users at places. In spite of this limitation, any system built to predict the amount of crowd and events at a place has to be accurate enough. Only outputs from such a system can be used as a source of metadata for places.

Scalability

The current geospatial databases contain information about millions of places. So, any system built to predict the number of people at a place or crowdedness at a place at any given time has to be scalable. This demands that crowd and event prediction system needs to be light weight in terms of its resource consumption.

We have briefly described some of the challenges in generating metadata for places based on data generated by users on mobile and Web applications. Any successful attempt in automatically deriving metadata of places has many advantages. The rich metadata derived in the form of various types of tags enables geospatial application users to make very precise queries about places and get specific information about places. It also means that businesses have rich information about places which in turn can be used for recommendation and personalisation of the services they offer. Any of geospatial web applications and location based services can make use of metadata about places in order to enhance the quality of services.

We now list the assumptions we have made while conducting research experiments and studies.

1. We have an access to unbiased datasets comprising of user generated data through mobile and Web applications.

2. Users contributing their data voluntarily are fair in data generation and publication.

3. The privacy of those users is not violated whose data is used for automatic metadata generation about places.

In the following section, we discuss motivating facts about deriving metadata for places. We then describe our hypothesis and briefly state the results we have obtained to validate our hypothesis.

1.2 Hypothesis

The large scale adoption of online social networks, micro-blogging services and related mobile and Web applications by users has enabled service providers to access vast amount of fine-grained user data. A large proportion of these data are geospatial in nature. For example, micro-posts generated on micro-blogging sites such as Twitter[7] sometimes contain geolocation of the micro-post or *tweet* generation. Many of those micro-posts have references to places and geographic regions. Similarly, users on online social networks like Facebook often refer to places they have visited, events organized at places among others. Location based social networks such as Foursquare let users generate social data annotated with place information and broadcast them publicly or to a group of friends. Most of the above mentioned data are unstructured and do not contain metadata about places explicitly. Place metadata includes tags or text snippets that are descriptive of places, information about events at places, duration of those events, common interests of visitors of a place among others. These metadata are required for accurate search and recommendation of places on geospatial applications and online social networks. However, users on online social networks do not manually annotate places with metadata. There are few successful data mining efforts which have generated geospatial metadata based on data generated by users on the Web. For example, Noulas et al. [16] compute similarity scores for geographic regions based on user check-in activity on Foursquare. In [8], Ye et al. use the temporal

---

[7]https://twitter.com/

and check-in count data to derive category labels for places. In [9], Mansour et al. generate term distributions to represent business places by analysing data from social media. These prior works motivate us to hypothesise that place metadata can be derived from user generated data on mobile and Web applications. Specifically, we hypothesise and show that:

1. Descriptive text snippets or tags can be derived for places using textual data generated by a group of online users. To this end, we have used online social profiles and user presence data and derived large number of relevant tags for places considered in an experiment. Further, we have analysed text data related to events conducted at places and shown that highly relevant place tags can be derived.

2. Events organized at places can be detected effectively by analysing user presence data for places. We have developed a Bayesian statistical model to detect outliers in time series of univariate data, which we use to infer about events organized at places. This shows that data generated by users due to their passive interaction with sensors at a place can be used to infer events at that place.

We have used real world data from various types of social networks and physical sensors to validate these hypotheses. In the following section, we describe some of the application scenarios that can benefit from the availability of metadata for places. We also discuss how geospatial web applications and location based services can be made more precise and useful by utilising the automatically generated geospatial metadata.

1.3 Application Scenarios

The utilization of rich metadata about places plays an important role in the quality of geospatial web services and location based services. In this section, we consider three major application scenarios that can hugely benefit from the existence of metadata for places.

Location Search

Users of geospatial applications can benefit from the ability to query for locations with tags. There can be two types of semantic tags that can be annotated for a place. First type comprises of *categorical tags* that represent an abstract category that a place belong to. Examples of categorical tags are text snippets such as *educational institution*, *restaurant* etc. The other type of tags are *descriptive semantic tags* which

describe specific concepts a place is associated with. It is essential that a place has annotations of both types for effective search results. For example, consider a user query for places which are related to *computer science* in a geographic region. This query can only be successfully executed if relevant places have descriptive semantic tags such as *computer science research*, *mathematics* etc. along with categorical tag such as *educational institution*. If descriptive semantic tags are not available, a user might be shown no places or irrelevant places in a geographic region. This adversely affects the user experience and quality of service of a geospatial application. On the other hand, size of mobile devices and the way users depend on mobile devices for information needs make us realize that place query texts will be precise and short on mobile devices. So, annotating places with appropriate tags would facilitate geospatial applications to support complex and precise queries from users. Also, the number of places in a given geographic region can be huge. Hence, semantic tags should be specific enough so that a short list of highly relevant locations can be shown against a user query.

Personalised Recommendations of Places on Mobile Devices

Mobile devices are capable of capturing both social and physical context of a user. Since mobile device users use applications extensively, their social preferences, interests etc. are available on mobile devices which represent social contexts of users. A wide range of sensors such as GPS sensor, accelerometer on these devices let us determine physical contexts of users. This scenario makes a mobile device an ideal platform for personalised delivery of information. We now discuss some of the research works that show the crucial role of personalisation while delivering information on mobile devices.

There have been numerous studies on the impact of personalisation of the information delivered on mobile devices. Personalisation has been found to reduce the amount of data traffic on mobile devices, increase user satisfaction and also be an effective mechanism in mobile commerce [17, 18]. The limitations in terms of screen size, key board facilities, battery power etc. demand the need for precise and personalised information delivery on mobile devices. Irrelevant information delivered will be considered spam and be rejected by the user community.

It has been realized that if thoroughly personalised, commercial messages may be perceived as valuable information services as described by Bauer et al. [18]. There have been efforts towards personalisation of information delivered on mobile devices. Activities of a mobile device user along with geolocation can be inferred

effectively from the sensors of a mobile device. This has been used as a basis for personalisation in Partridge et al. [19]. The approach adopted in Miele et al. [20] enables users to explicitly mention their preferences which will be exploited by a system for personalisation of information delivered. User profiles existing in Online Social Networks (OSNs) along with proximity of mobile device users has been utilized in Pietilainen et al. [21] for opportunistic social network expansion. In [18], Bauer et al note that, if not personalised, users can reject information delivered and also change their mobile network operator. The presence of highly descriptive tags for places can be utilized along with social profiles of users to personalise the recommendation of places. For example, a user interested in *Art* can be shown places with semantic tags such as *Irish Music*, *Music*, *Art* etc. with a higher priority rather than showing a popular place with tag *Computer Science*. Similarly, places similar to the places frequently visited by a user can be recommended for the user by exploiting metadata available for places.

Detection of Events Organized at Places

Geospatial application users can greatly benefit from real time information about amount of crowd and events organized at places. The crowd information can be used for better scheduling of visits to places, notify any unusual crowd gathering at a place, deliver optimized advertisement campaigns etc. There has been extensive research on predicting amount of crowd at places using mobile device data and data generated from geospatial web applications and sensor data. The work in Ratti et al. [22] analyses movement of users through space and time in urban setting and applications are proposed for urban planning. Prediction techniques to determine future location visits of mobile device users based on GPS trajectories of mobile devices are discussed in Ying et al. [23]. Mobile phone signal data of users has been used in Calabrese et al. [24] to correlate human movements during social events. In [25], Zhou et al. show that bus arrival times can be predicted with high accuracy by detecting the locations of mobile device users using scalable techniques. We can see from the above mentioned research that the prediction of users at locations and detecting any unusual crowd at locations has huge implications for urban planning, event scheduling, location recommendation etc. There have been efforts to predict events at places based on the user generated data where the ground truth about the occurrence of events is not available. Accurate methods to detect events conducted at places are crucially needed as such event information are not systematically updated on the Web.

1.4 Contribution and Structure

The contributions made in this thesis are as follows:

1. We have shown that descriptive tags can be derived for places by analysing social interest profiles and location visiting patterns of a group of users. We have specifically studied whether common social interests of a group of users can be utilized to derive text snippets that are representative of a place. We have analysed the effectiveness of utilising implicit 'Wisdom of Crowd' that is present among a set of visitors of a place. We have defined an algorithm and a probabilistic model to achieve the same. We have used data from a well-known social network and a location based social network for our analysis. In Chapter 3, we discuss this in detail.

2. We have shown that textual data generated by user interactions while they participate in events organized at locations can be successfully used to derive descriptive tags for locations. We have applied a probabilistic model and performed statistical analysis to derive the tags and shown that tags corresponding to multiple topics relevant to a place can be derived. We have used data from a famous event based social network to derive our inferences. In Chapter 5, we describe our findings about this technique of deriving place tags.

3. We have developed an advanced probabilistic model to analyse time series data. We have used this to detect outliers in count data generated by user movements in buildings and subsequently identify hours during which events are conducted in buildings. We have applied our technique on both simulated and real world data and compared the performance of our technique against a state of the art technique. In Chapter 4, we describe our technique and discuss the advantages of our technique.

The work presented in this thesis has several advantages. We propose techniques for utilizing user generated content on mobile and Web applications to enrich information about places. Our techniques respect privacy of users whose data is used for geospatial data enrichment by not analyzing the data of any particular user but that of the crowd. We provide strong theoretical framework for analyzing user generated content which can be used to generate metadata about places and is independent of any platforms or services. We provide analysis of real world data sets, describe how our techniques can be efficiently used and discuss the performance of our techniques against those real

world datasets. We analyse the robustness of our techniques that automatically generate metadata about places. We also discuss on how such metadata can be effectively used in search and recommendation of places.

The rest of the thesis is organized as follows. In Chapter 2, we will provide the background of the thesis which is the basis for theoretical and experimental contributions made in this thesis. In Chapter 3, we consider the scenario of automatically generating tags as metadata for places based on the location visiting patterns and social interest profiles of mobile device users. We describe a simple probabilistic model along with a text processing algorithm to derive tags for locations. We derive sets of tags for locations on a well-known location based social network (LBSN) using the social profiles and check-ins data of a group of users. We study the effectiveness of our technique using various information theoretic measures. We also evaluate the performance of our technique by comparing the derived tags against a set of tags provided by volunteers.

In Chapter 4, we describe how the presence information of users at locations can be utilized to understand any unusual crowd gathering at locations. We present a Bayesian network which is a probabilistic graphical model along with an appropriate version of Markov chain Monte carlo (MCMC) algorithm for the inference. We compare our technique with a state of the art event detection technique based on Markov modulated Poisson process (MMPP) using a simulated data set. We show that our technique has better accuracy in detecting events by analysing the performance of our technique on a real world events data set. We also show that our technique can be used to predict the amount of *check-ins* at locations accurately. In Chapter 5, we analyse the interaction data of users at events conducted at locations and show that such data can be used to generate descriptive tags for locations. We derive tags for a set of locations represented on an event based social network (EBSN) using textual data generated due to events conducted at those locations. We use the manual tags annotated for corresponding locations on a different social network to evaluate our performance. We show that we are able to derive larger number of highly relevant location tags with our technique. We conclude the thesis and discuss the future work in Chapter 6.

1.5 Research Outcome and Impact

Publications

We now list relevant publications we have produced:

- In Hegde et al. [26], we have discussed how the social interest profiles of visitors of a place can potentially be used to derive descriptive tags for places. We have used the results of this work in Chapter 3.

- In Hegde et al. [27], we have shown that amount of people at a place can be analysed to infer any events that happen at that place. We have proposed an advanced Bayesian algorithm to analyse time series of count data to infer outliers which we in turn use to detect events. The proposed algorithm is more accurate than an advanced event detection model and has many advantages which we describe in detail in Chapter 4.

- In Hegde et al. [28], we have shown that textual data generated by online social network users during various events organized at a place can be analysed to derive descriptive semantic tags for that place. Our proposed methodology considers thousands of possible tags and comes up with a handful of highly relevant tags for a place. We describe our methodology and outcomes in Chapter 5.

- In Weth et al. [29], we have discussed the user movements across locations in urban settings and showed the importance of annotating locations with appropriate websites. We have also shown that very few locations are annotated with websites thus hindering any reliable use of location websites for tag generation.

Invention Disclosures and Patent Applications

- Data Analysis and Event Detection Method and System
  Vinod Kumar Gajanana Hegde, Milovan Krnjajic, Manfred Hauswirth
  Patent Application, European Patent Office, 2015

- Unsupervised Outlier Detection in Count Data with a Bayesian Nonparametric Model
  Vinod Kumar Gajanana Hegde, Milovan Krnjajic, Manfred Hauswirth
  Invention Disclosure, Technology Transfer Office, NUI Galway, 2014

Research Grants for Commercialisation

- Commercialisation Grant for a project entitled 'DAATIC, A **D**at**A A**naly**T**ics platform for customer **I**ntelligen**C**e based on big data'.
  Vinod Kumar Gajanana Hegde has been the principal investigator. This project has been funded in 2015 by Enterprise Ireland and has been co-funded by the Eu-

ropean Regional Development Fund (ERDF) under Ireland's European Structural and Investment Funds Programmes 2014-2020. The grant amount is €166553.

- Commercialisation Feasibility Study Grant for a project entitled 'The need for a data analytics platform for location based intelligence'.
Vinod Kumar Gajanana Hegde has been the principal investigator. This project has been funded in 2013 by Enterprise Ireland. The grant amount is €13300.

Chapter 2

Background

In this chapter, we introduce some of the theoretical concepts and user data that form the basis for the models developed and used in the thesis. First we discuss the interaction of users on location based social networks (LBSN) and event based social networks (EBSN) where some of the data generated by users is related to places. We will discuss some of the data mining efforts that analyse data generated on LBSN and EBSN. We will specifically focus on research efforts and application systems that have exploited data generated on these types of social networks to generate metadata for places. The findings and shortcomings of current state-of-the-art techniques have motivated us to research further and provide potential solutions. We describe these contributions in next three chapters. Specifically, in Chapters 3 and 5, we have used data from these categories of social networks for analysis to derive place metadata and to validate our hypothesis. In the following sections of this chapter, we discuss probabilistic mixture models which can model underlying sub population structures in any dataset. These models have been used in Chapters 4 and 5 for probabilistically deriving place metadata. We finally describe two major categories of probabilistic mixture models and discuss specific instances of these models that have been successfully utilized in many applications. We have used and further developed these specific model instances for probabilistically generating place metadata which we discuss in Chapters 4 and 5.

2.1 Location Based Social Networks (LBSN)

In recent years, location based social networks on mobile devices have become very popular. These social networks are centred on physical and social context of their users. They provide many functionalities to users for online networking. The users

can describe their social interests along with their summary profiles. They can connect as friends with people who have similar interests or who have visited the geolocations. LBSN enable the users to share their current geolocation with friends or broadcast to the public depending on their privacy preferences. They can manage a list of places they intend to visit in the future. The users can rate and comment about the places they have visited or know about. They can also share the videos and photos that they have taken at those places. The growing number of mobile device users demands for more LBSN services that can provide rich experience to users. Interestingly, as a result of the user interactions on location based social networks, a huge amount of geolocation related data is generated by users. Various data mining and information retrieval tasks have been defined and developed using these data. We discuss some of these data mining tasks in the remainder of this subsection. Our discussion considers three major categories of tasks namely **place recommendation**, **place-centric opportunistic networking** and **metadata generation for places**. Some of the prominent location based service providers are FourSquare[1], Yelp[2] and Facebook[3] among others. In Chapters 3 and 4, we analyse the nature of the data generated on LBSN.

Recommendation of places based on geolocation and social interests of a user is one of the important problems to be addressed by location based services and geospatial web services. Ye et al. [30] explores effective approaches to collaborative recommendation of locations based on datasets of user visits or *check-ins* on famous location based social networks. Noulas et al. [16] describe an approach to exploit the semantic annotations of places to cluster geographic regions and users. They also explain the use of such analysis for recommending appropriate places to users. In Zhang et al. [31], authors have proposed a model to detect neighbourhood boundaries. They have used this model to build a system that recommends activities in various neighbourhoods. Predicting the movements of people precisely is a challenge and accurate predictions have many applications. In [32], Cho et al. show that human movements can be attributed to the periodic behaviour and social relationships of the users on an LBSN and predicts people movements with high accuracy. Gao et al. [33] use geo-social correlations among users to overcome the *cold start* problem in prediction of location visits of users. Zheng et al. [34] show that even the raw GPS trajectories can be utilized to provide place recommendations for users. Noulas et al.

---

[1] https://foursquare.com/

[2] http://www.yelp.ie/la

[3] //www.facebook.com/places/

[35] show that check-ins data on LBSN can be used to accurately predict the places a user will visit in the future.

Data generated on the LBSNs potentially contains information about real world events. Motivated by this, there have been various research works on inferring events conducted at places. De et al. [36] show that spatio-temporal information about forest fires can be derived using the publicly available data generated on LBSNs. In [37], Ferrari et al. describe the application of a probabilistic model to infer the most frequent activity patterns in major cities. Data generated on location based social networks have been used to generate metadata for locations. Mansour et al. [9] have generated and used term distributions to represent business places by analysing data from social media. These term distributions are found to be effective in handling user search queries regarding business places. Ye et al. [8] use temporal and check-in count data to derive category labels for places. Location based social network data has been successfully exploited for other types of inference tasks. In [38], Karamshuk et al. show that optimal locations for establishing new stores can be effectively inferred using data like type of places, inflow of distant users to a place among others which are generated on LBSN by users. [15, 39] show that geographic region specific topics that are trending at any time can be obtained by analysing check-ins and interaction data of LBSN users. Opportunistic networking with other users based on location visiting patterns of users is a service that has been studied in Ying et al. [40]. Scellato et al. [41] address the problem of predicting new social links among users based on their location visiting patterns on LBSN. Historical data about location visits of users have been used in Zheng et al. [42] to infer the user similarity and recommend friends.

2.2 Event Based Social Networks (EBSN)

The wide adoption of Web applications in everyday life of people has a significant impact on the way people communicate and interact with each other. A huge number of Web applications which are centred on the organization and participation of people in real world events have successfully fused the offline and online worlds. Specifically, event based social networks (EBSNs) are online social networks that are centred on the events and social aspects of event participants. They act as online platforms to conduct offline events in a systematic manner where users can physically participate in events of their interest at a given place. They let users group among themselves based on their interests and location. Users can also collaborate on conducting events through active discussions on EBSN forums. EBSNs let users explore events using keyword searches

on event categories and names. Users generate rich metadata about events and places where events are conducted by commenting about user experience at events, rating the events, uploading event related photos and videos, etc. Huge amounts of event specific data generated by users potentially represent a wealth of information about events. Motivated by this, many data mining and information retrieval solutions have been proposed which exploit data generated on EBSNs. Some of the prominent EBSN platform providers are Meetup[4], Plancast[5], Eventfull[6] among others. In Chapter 5, we discuss the nature of data generated on EBSN platforms.

Liu et al. [43] describe how EBSNs have played an important role in merging the online connections and offline interactions. They show that networks on EBSNs are more cohesive than other types of social networks such as LBSNs. They also infer that information diffusion among users is best predicted by considering both online and offline interactions. In [44], Feng et al. describe a solution to find the most influential event organizers to determine the overall success of conducted events. Peifeng et al. [45] show that there is no correlation between offline and online interactions thus reiterates the importance of analysing the offline interactions involved during events with event related data. Number of people intending to attend an event can be huge on EBSN platforms. Those users usually have varying social interests and potential some common interests. So, Li et al. [46] study the problem of organizing events efficiently and propose a solution by analysing the friendship network of users on EBSNs.

2.3 Mixture Models

Mixture models are probabilistic models used for inference on data where there is partial or complete lack of information about sub populations present in a dataset. They are used to represent a dataset using a set of parametric probability distributions governing the sub populations. The usual choices of probability distributions include Gaussian, Poisson, Multinomial distributions. When the number of mixture components is fixed, the mixture model is known as a finite mixture model (FMM). The finite mixture models are used when there is prior information about the number of sub populations in the dataset. It is necessary to use infinite mixture models (IMM) when there is no prior information about the number of sub populations involved in generating the data. The sub populations are represented by mixture components or

---

[4]http://www.meetup.com/
[5]http://plancast.com/
[6]http://eventful.com/

clusters. They let us model the joint probability distributions governing the population from which a sample i.e., a dataset is observed. We now discuss them in detail and the some of their successful applications in probabilistic inference.

2.3.1 Finite Mixture Models

A finite mixture model can be described using the following variables.

1. N data items $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$

2. K clusters which is the number of sub populations

3. π which is a vector of mixture weights of the clusters in the population

4. θ which is a vector of parameters for each of the clusters

A finite mixture model is a convex combination of more than one probability density function of the same family. FMMs are used for inference when there is strong prior information about the number of sub populations for a given dataset. The probability density function of an FMM is defined as

$$Pr(x|\theta, \pi) = \sum_{k=1}^{K} \pi_k Pr(x|\theta_k) \tag{2.1}$$

$$\sum_{k=1}^{K} \pi_k = 1 \tag{2.2}$$

So, the parameter K is known before the inference is carried out on a dataset. The inference problem will involve considering the data at hand as samples from a population and estimating π and θ. Here π denotes cluster weights or mixing proportions of data belonging to each sub population. θ denotes a vector of parameters of the probability distributions governing the sub populations. In a Bayesian setting, π and θ are random variables and have prior distributions. So, the generative model of the finite mixture model can be described as follows.

$$\pi|\alpha \sim Dir(\alpha) \tag{2.3}$$

$$z_i|\pi \sim Discrete(\pi) \tag{2.4}$$

$$\theta_k|\lambda \sim G_0(\lambda) \tag{2.5}$$

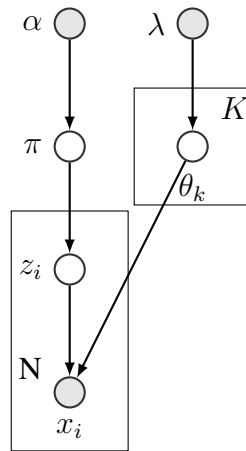$$x_i|z_i, \{\theta_k\}_{k=1} \sim F(\theta_{z_i}) \tag{2.6}$$

Figure 2.1: Generative model of the finite mixture models

Here, z_i is a auxiliary variable used in the sampling step of inference. π is a discrete distribution with a dirichlet prior distribution defined by α. θ_k are parameters of a probability distribution and they generally belong to an exponential family as all the members of exponential family have a conjugate distribution. The prior distribution for θ_k is defined by the base distribution with parameters λ. The inference problem now turns out to be deriving the posterior distributions of these variables using the likelihood of the data. Any Markov Chain Monte Carlo (MCMC) algorithm can be applied to infer these random variables as discussed in Andieu et al. [47]. The generative model of the finite mixture models is shown in Figure 2.1.

Finite mixture models have been successfully used for probabilistic inferences in various applications. Bailey et al. [48] have effectively used the two component mixture model to analyse the protein sub-sequences and estimate thresholds of Bayesian classifiers for further analysis. In [49], Deb et al. discuss how a finite mixture of negative binomial distributions can be used to infer the heterogeneity in the health care spend of citizens. Image segmentation is a well-studied problem in the area of computer vision. Alfo et al. [50] find that a finite mixture of Gaussian distributions can be effectively applied to perform image segmentation and achieve better segmentation results even for images with high noise. More recently, latent dirichlet allocation (LDA) [51] has been successfully employed in various text mining problems. We now explain latent dirichlet allocation which forms the basis of our data analysis in Chapter 5. Latent dirichlet allocation is a finite mixture model which is used to probabilistically model a collection of data items. Here each data item has a finite mixture of latent random variables where each random variable has its own probability distribution with sample space comprising of individual elements of all data items.
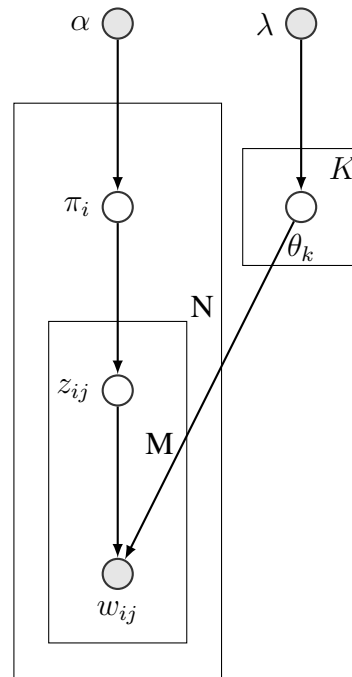
Figure 2.2: Generative model of latent dirichlet allocation (LDA)

2.3.2 Latent Dirichlet Allocation

The primary version of latent dirichlet allocation proposed by Blei et al. [51] is an unsupervised text clustering technique by inferring the latent topic and word distributions for a set of documents in a document collection. It infers the latent topic distributions for each of the documents. Each latent topic is also analysed for the word distributions over the set of all words in the document collection. The topic distributions of a document can be used to infer the most probable topic indices for the document and subsequently the most probable words representing the document. It can be used to cluster documents, collaborative filtering on document collections and text representation [51]. Various versions of LDA technique have been successfully applied in data mining and machine learning tasks. These include localizing the software bugs [52], mining business topics from source code [53], fraud detection in telecommunication networks by analysing the call behaviour [54], unsupervised entity resolution [55]. There are many research works that have been motivated by the primary version of LDA and have proposed extended models of LDA for various probabilistic inference tasks. Wang et al. [56] have studied the task of image classification and text annotation by applying sLDA model. Titov et al. [57] show that aspects that are rated by users can be derived along with the topics from the collection

of user reviews using an extended model MG-LDA. The variables involved in the LDA model are as follows.

1. N is the number of documents in the collection

2. K is the number of latent topics

3. M is the number of words in a given document

4. w_{ij} is the jth word in i^{th} document

5. π_i is the topic distribution for i^{th} document

6. θ_k is the word distribution for k^{th} topic

The generative model of LDA can be described as below.

$$\pi_i | \alpha \sim Dir(\alpha) \tag{2.7}$$
$$\theta_k | \lambda \sim Dir(\lambda) \tag{2.8}$$
$$z_{ij} | \pi_i \sim Discrete(\pi_i) \tag{2.9}$$
$$w_{ij} | z_{ij}, \{\theta_k\}_{k=1} \sim Discrete(\theta_{z_{ij}}) \tag{2.10}$$

Here, π_i indicates the topic distribution for i^{th} document. It is a discrete distribution over topic indices for the documents. θ_k are parameters of a discrete distribution with the sample space comprising of all the words in the document collection. The prior distribution for θ_k is defined by dirichlet distribution with parameters λ. z_{ij} is the latent topic indicator for word w_{ij}. This model can be used to infer per document topic distributions π_i and per topic word distributions θ_k using any variational Bayes approximation [51], Gibbs sampling an MCMC algorithm [58]. LDA is a finite mixture model where number of topics is known a priori. A log-likelihood based technique has been proposed in Griffiths et al. [58] to find the optimal number of topics. Unsupervised versions such as hierarchical dirichlet process [59] have been applied by Teh et al. to address the problem of document clustering. These models have performance comparable to LDA but do not require any model selection. The generative model for latent dirichlet allocation is as shown in Figure 2.2.

2.3.3 Infinite Mixture Models

Infinite mixture models (IMM) also known as dirichlet process mixture models are probabilistic models which are highly flexible in modelling the sub populations within

the data. The flexibility is due to the assumption that an unknown number of mixture components exist in the data with non-zero mixture weight. The number of mixture components K as described in Section 2.3.1, is set to be infinite in these models. In Figure 2.3, we show the corresponding graphical model. The number of such components is inferred every time a new sample is observed. The ability of infinite mixture models to infer and adapt to the changing number of mixture components in the data has resulted in wide application of them in various probabilistic inference tasks in various fields. Chen et al. [60] show that Gaussian IMM can be used to accurately infer the confidence bounds during statistical process performance monitoring. Shin et al. [61] discuss the application of Gaussian IMM on astronomical time series data to cluster the data and detect outliers. The problem of inferring about number of sub populations and assignment of individuals to sub populations based on the genetics data has been discussed in Huelsenbeck et al. [62]. They conclude that infinite mixture model can be utilized for inference and sensitivity of the inference can be unaffected by prior choice on the number of sub populations when the data is sufficiently large. Kottas et al. [63] discuss how IMM can be used for modelling mortality count data with spatial attribution for accurate predictions and discuss the limitations for parametric models for modelling such data. IMM based hierarchical clustering has been proposed in Heller et al. [64]. This overcomes the traditional problems in hierarchical clustering using ad hoc number of clusters, limitations on making any probabilistic statements about cluster membership, cluster cohesion etc. Some more prominent works that have applied infinite mixture models to various probabilistic inference tasks are [65, 66]. The generative model of the infinite mixture model can be described as follows.

$$\pi|\alpha \sim GEM(1, \alpha) \tag{2.11}$$

$$z_i|\pi \sim Discrete(\pi) \tag{2.12}$$

$$\theta_k|\lambda \sim G_0(\lambda) \tag{2.13}$$

$$x_i|z_i, \{\theta_k\}_{k=1}^{\infty} \sim F(\theta_{z_i}) \tag{2.14}$$

2.3.4 Infinite Poisson Mixture Model

An infinite Poisson mixture model is a Bayesian nonparametric model which can be used when the probabilistic inference needs to be carried out on count data. Count data arises often in the real world situations such as vehicular traffic data, user visits or check-ins data at locations, number of Web clicks of any Web site etc. The following
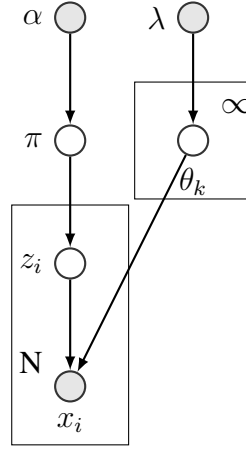
Figure 2.3: Generative model of the infinite mixture models

equations represent the generative model of infinite Poisson mixture model. Here, GEM refers to the stick breaking process discussed in Sethuraman et al. [67].

$$\pi|\alpha \sim GEM(1,\alpha) \tag{2.15}$$

$$z_i|\pi \sim Discrete(\pi) \tag{2.16}$$

$$\theta_k|\lambda \sim Gamma(\lambda) \tag{2.17}$$

$$x_i|z_i, \{\theta_k\}_{k=1}^{\infty} \sim Poisson(\theta_{z_i}) \tag{2.18}$$

The probabilistic inference will include clustering the count data into different sub populations or *clusters* each with its own Poisson rate parameter. A gamma prior distribution is used as base distribution to model the rate parameters. The proportion of each cluster in the population corresponding to the count data is also inferred. In [68], author discusses on how an Infinite Poisson mixture model can be used for unsupervised learning where there are multiple features occurring multiple times in a data point. Krnjajić et al. [69] show the advantages of modelling count data with Bayesian nonparametric approach over the parametric approach. The authors analyse the data of a randomized control trial and show that such count data can be accurately modelled using the Bayesian nonparametric modelling compared to parametric approaches. However, if there is significant Poisson over dispersion or under dispersion in the inferred clusters, an infinite negative binomial mixture model would be appropriate [70].

2.4 Summary

In the first two sections of this chapter, we described two major types of social networks that capture geospatial related social data. We specifically discussed the nature of the data generated on these social networks. We also listed numerous commercial players that provide some of the prominent LBSN and EBSN platforms. Furthermore, various data mining techniques that have used on these data sources to generate additional insights were also presented. Then, we described a prominent class of probabilistic modelling techniques that are sophisticated and have gained huge popularity in the recent years due to their performance. Since the aim of this research work is to automatically generate metadata for places and uncertainty is involved in any automated inference, we have developed and applied these probabilistic techniques to analyse geospatial data generated on EBSN and LBSN to derive metadata for places. In the following three chapters, we discuss these in detail.

Chapter 3

Semantic Tagging of Places with Social Profiles of Users

Online forums such as online social networks (OSNs), location based social networks (LBSNs) let the users interact with each other. Some of the data generated due to these interactions inherently has a geographic aspect. Even though these online forums enable users to manually tag places they have visited or have knowledge about, users rarely do so. Moreover, the available information attached to places (e.g., their names, category, textual address) is often ambiguous or insufficient for service providers to automatically generate tags. On the other hand, users often provide information about their interests in online profiles via online social networks. They also express the locations they visit by broadcasting the visits to the public or friends via online forums. Motivated by these facts, we address the following question in this chapter. *Can the common interests of a group of people that has visited a particular place be used to derive tags for that place?*

Specifically, we focus on deriving descriptive semantic tag annotations for places using social interest profiles and location visits or *check-ins* data of users. We present an approach that automatically assigns semantic tags to places, based on interest profiles of users and their check-ins at places. The approach consists of: (i) an interest profile expansion algorithm to derive semantic concepts related to the user interests; (ii) a model to determine the probability that a particular semantic concept describes a place, based on the check-in activities of users; and (iii) a noise removal approach, using a hierarchical clustering technique, which is applied on the top-probable semantic concepts to derive the final semantic tags for places. We have evaluated our approach with real world datasets from popular social networking service, against a set of

manually assigned tags. The experimental results show that not only we are able to automatically derive meaningful tags for different places, but also that the sets of tags assigned to places are expected to stabilise as more unique users check-in at places. This indicates that top-probable tags derived can be consistently assigned to places after check-ins by a threshold number of users. In the next few paragraphs, we provide the motivation and the state of the art techniques for annotating locations with tags.

In recent years, numerous Geospatial Web services (GWS) have enabled the users to generate and consume geospatial data. On the other hand, mobile devices have become highly ubiquitous. Equipped with sophisticated sensors such as GPS sensors and cameras, they now enable a new range of location based services (LBS). These services determine the physical location of their users and provide a number of functionalities. The physical location of a user can be described at different levels of accuracy. Cell triangulation techniques [71–73] determine the approximate location of a mobile device whereas GPS sensors on mobile devices can provide much accurate location with latitude, longitude data. For instance, users can check-in at places, i.e. users can let others know of their whereabouts. Data about check-in activities have already been explored to understand user behaviours to provide personalised advertising and promotion of businesses [30, 74–76]. Another functionality common in GWS, LBS is place recommendation: nearby places are suggested to the user by matching the description of the places with the user needs or interests.

The performance of place recommendation techniques depends on the richness of the geographic metadata used. This geographic metadata includes places or points of interests (POIs), comments, ratings about places, abstract category tags and descriptive semantic tags. By descriptive tags, we mean any short keywords which are semantically related to a place. For example, it would be appropriate to tag a *Computer Science Building* with tags such as *Software*, *Engineering*, and *Programming*. A categorical tag such as *Academic Building* for a place is much more abstract and less informative. Even though users often use LBS for check-in activities, they rarely tag a place. Currently, most of the places described in the geospatial databases used by prominent GWS and LBS providers are poorly tagged. A study by Ye et al. [8] on one such service showed that 30% of the places do not contain any tags. Our analysis of more than 1 million places on Foursquare[1] suggested that only 7% of the places had any descriptive tags and only 21% of the places had any tips/comments in the form of short text snippets. We show the distribution of tags and tips in Figure 3.1.
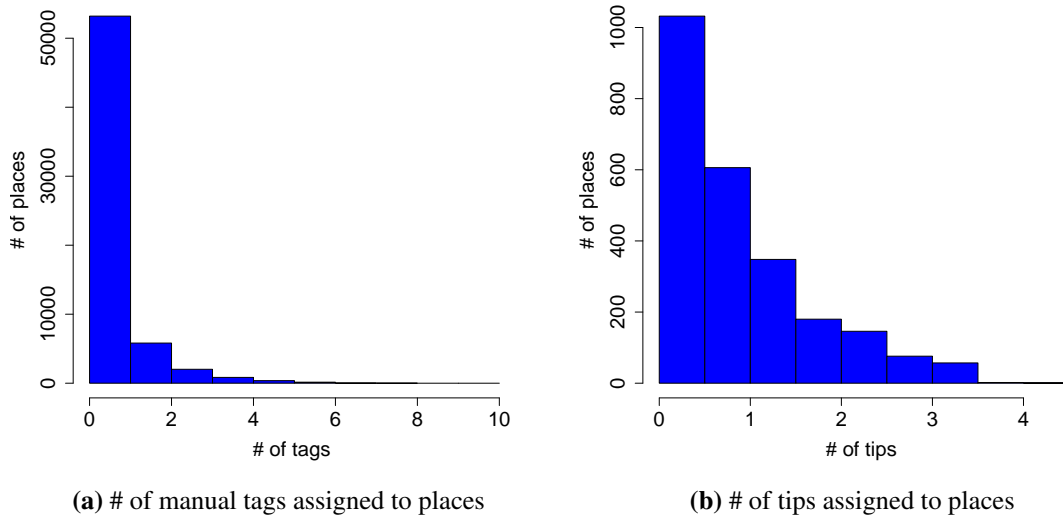
---

[1]https://foursquare.com/

(a) # of manual tags assigned to places   **(b)** # of tips assigned to places

Figure 3.1: Histograms of tags and tips at places

Geospatial application providers can not automatically generate tags since the available information about the places is often ambiguous or insufficient. However, users look for very specific information on the mobile phones based on their current physical location and social contexts. Therefore, it is crucial that techniques to automatically assign semantically related tags to places are developed so that search and recommendation of places can be more effective. In a different context, many complex problems related to information generation on the Web have been solved utilising the wisdom of the crowd [77–79]. For example, in [80, 81], Goodchild et al. and Sheth et al. discuss various ways in which explicit or implicit information provided by the users can be utilised to enrich information on the Web. Web users leave their footprint on the Web using resources such as online social networks (OSNs) and microblogging systems, which can be used to derive the user's preferences and interests. Many of the users of OSNs also use location based services to check-in into places. Based on the above observations, the common interests of a group of people that has visited a particular place can potentially provide further description for the place.

In this chapter, we describe how the two sources of information combined – user interest profiles on OSNs and check-in logs – can be utilised to derive tags for a place. We present an approach that automatically assigns semantic tags to places, based on interest profiles and check-in activities of users. We first extract semantic concepts

from the interest profiles of users available on OSNs. However, the interest profiles of users are often sparse and contain only a few keywords. In Section 3.1.3, we describe an interest expansion algorithm that discovers "hidden" interests by expanding the user interest profile in a controlled manner. The expansion algorithm is able to derive more concepts without deviating from the user interests. In Section 5.1.2, we provide a model to determine the probability that a particular semantic concept describes a place, based on keywords representing the interests of users and check-in activities of users at places. We consider the top-k probable semantic concepts for any given place and perform a hierarchical clustering on those concepts to derive the final set of tags. We give the details of this technique in Section 3.1.2. In summary, our approach is to use social interest profiles of a group of users to infer a collective set of interests of users that visit a particular place. Our approach exploits such collective interests to infer a set of semantic tags that are descriptive of that place. An abstract picture of our approach to automatic tagging of places is shown in Figure 3.2.

Figure 3.2: Collective interests of people checking in at places

We have evaluated our algorithm with real world datasets from popular social networking service, against a set of manually assigned tags. We have also studied the nature of tag probability distributions against the check-in activities by users in order to understand the quality of the top-probable tags and collective interests of people visiting places. The experimental results show that the automatically generated tags are similar to the manually assigned tags, and also that the sets of tags assigned to places are expected to stabilise as more unique users check-in at places. This indicates that top-probable tags derived can be consistently assigned to places irrespective of the number of people who have checked-in at those places. We give the details of

the experimental results in Section 5.2. In Section 3.3, we discuss and conclude our findings presented in this chapter. We now discuss some of the prior work that has been carried out to derive metadata for locations.

In recent years, there has been an increased interest in the area of analysis and enrichment of geographic data. The amount of volunteered geographic information (VGI) is rising, as more users are equipped with sophisticated mobile devices which enable them to actively contribute with geographic data. [82, 83] have studied various approaches to deriving and recommending tags to annotate geospatially annotated images based on various types of user data. The work by Haklay et al. [10] gives an example of how GPS traces and other geographic data provided by people can be used to create an accurate map of the world. In [4], Wang et al. discuss various approaches that can be adopted for manually tagging places using mobile phones. It shows, for instance, that users prefer more than one place annotation technique in location aware applications and the offline annotation scheme is the most preferred one. A semi-automatic approach to integrating information provided by users into the digital geographic gazetteers has been discussed in Kessler et al. [5].

All these works indicate that there is a need for obtaining and enriching geographic information and that the manual effort to generate such information is not enough. In Lian et al. [7], an automatic place naming technique based on user check-in activities is discussed. However, this deals with deriving only the names of the places while our approach provides descriptive tags for the places. Noulas et al. [16] provides a good example of the importance of semantic annotations, where they show that identification of user communities and comparison of urban neighbourhoods can be done using the annotations of places. In [8], Ye et al. find that significant amount of places lack even the abstract textual descriptions and hence focus on deriving the categorical tags for place categories such as *restaurant* and *cinema*. Our work, on the other hand, focuses on deriving more descriptive tags. To the best of our knowledge, assigning places with automatically derived semantic tags has not been studied yet. Such methodology is much needed as users rarely assign specific tags to places and rich information is needed for search and recommendation of places.

3.1 Tagging of Places with OSN and Check-ins Data

Online social networks enable users to express their social interests and other personal information via their user profiles. In addition, location based social networks let users

express their location information with check-in activities. In this section we describe how we use both the user interests listed in OSN profiles and the check-in activities of users to derive descriptive tags for places. We first present our probabilistic model for determining the probability that a given semantic tag describes a place, based on the interests of users that have visited the place. A hierarchical clustering technique is applied on the top-probable semantic concepts to remove possible 'noise' tags and derive the final semantic tags for places.

In the later sections of this chapter, we study the quality of such derived tags by analysing the co-occurrence of place names and derived tags on the Web. We further analyse the stability of the derived sets of tags as social interests of larger number of users is considered. Our analysis shows that highly relevant descriptive tags can be derived by analysing social and physical presence or **check-ins** data of users. We also infer with empirical data that most probable tags stabilize as data of larger number of users is exploited to derive place tags.

3.1.1 Probabilistic Model for Deriving Tags for a Place

Our probabilistic model considers the check-in activities of users and their interests to derive the most probable tags for a place. Let U denote the set of all users who check-in at places and let P denote the set of all places (or POIs) the users can check-in. A user check-in is modelled as a tuple of the form $\langle u, t, p \rangle$, where $u \in U$, $p \in P$ and t is the timestamp of the check-in activity. The set of all user check-ins is denoted by CH. From CH we can extract CH_{ip}, which is total the number of check-ins of user i user at place p, and CHU_p which is the set of users who have checked in at least once at place p. The set of concepts in the interest profile of user i is given by K_i.

When the ith user checks in at p, we consider each concept in K_i as candidate tag for p. We do so with the hypothesis that there is a possible semantic relationship between a place and any concept in the interest profile of the person checking in at that place. The check-in action by any users at p contributes to the expansion of the candidate tag set CT_p which is defined as $CT_p = \bigcup_i K_i$ where $i \in CHU_p$. Given the pth POI, the probability that p is checked in by i^{th} user is given by

$$Pr(U_{ip}) = \frac{CH_{ip}}{\sum_j CH_{jp}} \qquad (3.1)$$

where $i \in CHU_p$ and $\sum_j CH_{jp}$ is the total number of check-ins by all users at p.

The conditional probability that i^{th} user with n concepts in K_i attaches one of the concepts k_j as tag to a POI is given by

$$Pr(k_j|U_{ip}) = \frac{1}{n} \qquad \forall p \in P, k_j \in K_i \tag{3.2}$$

This is with the assumption that all concepts in a user interest profile equally represent the interests of a user. The total probability that the pth POI is attached with the concept k_j as a tag is given by

$$Pr(k_j) = \sum_i Pr(k_j|U_{ip})Pr(U_{ip}) \qquad \forall k_j \in CT_p \tag{3.3}$$

We call this the *Tag Probability* of the concept k_j. It is easy to see that $\sum_j P(T_p = k_j) = 1$ and $0 < P(T_p = k_j) \leq 1$ where $k_j \in CT_p$. This means that a categorical random variable T_p defines the probability distribution of the tags for the place p where the sample space $\Omega = CT_p$. We can see that a random variable T_p^n can be defined by considering the check-in activities of the first n unique users at place p (denoted by CT_p^n) with the sample space $\Omega = CT_p^n$. In our model to derive tag probabilities, concepts in the interest profile of a frequent visitor are considered as more probably related to the corresponding place. The most probable words in T_p^n capture common interests of users and are influenced by the frequency of check-ins of users. Note that number of tags in the sample space can be large and all are not descriptive of a place or POI. So, there is a need to consider top k probable tags for some value of k, and choose only relevant tags from that set. In the next subsection, we describe a mechanism to obtain few but highly relevant tags for a place.

3.1.2 Hierarchical Clustering of Top-probable Tags

In our approach to deriving semantic tags for places, though we can derive a set of top-probable tags for a place, not every tag is necessarily semantically related to that place. Also, there can be potentially multiple topics that semantically represent a place. This observation demands clustering of the tags so that we could obtain one or more "natural" clusters of tags to tag a place and discard unrelated tags which are noise. We describe

the advantages of clustering the top-probable tags in Section 3.2.2. Hierarchical clustering is one of the widely used clustering method for efficient clustering and Johnson et al. [84] list various techniques and advantages of hierarchical clustering. The work in Paolillo et al. [85] successfully employs hierarchical clustering to obtain clusters of interests from interest profiles of users without considering the geographical aspects of users.

We compute the semantic similarity between the tags and use the agglomerative nesting algorithm with the group average method [86] to obtain clusters of tags. The group average method is a very effective method for clustering documents and detailed results about the effectiveness are described in El et al. [87]. Determining the number of clusters given a set of elements is a well-known problem and various techniques for deriving the appropriate number of clusters have been proposed. In Langfelder et al. [88] a novel method for cutting the dendrogram obtained from hierarchical clustering to obtain clusters is discussed. We used this method to obtain the clusters of tags corresponding to each random variable for each place. Any tag that does not fall into the generated clusters is then discarded.

It has been noted that online social profiles contain very little text. So, it becomes necessary to expand such profiles so that larger but semantically similar profiles can be obtained to be useful in deriving place tags. In the next subsection, we describe an algorithm to semantically expand a sparse social interest profile. The algorithm expands a set of Wikipedia[2] concepts to obtain a larger number of but highly semantically related Wikipedia concepts. We further describe some of the reasons that motivated us to use only Wikipedia concepts as place tags while exploiting social data of OSN users. Since Wikipedia concepts will be used as potential tags for places in our current approach, we have used Wikipedia Link Vector Model (WLVM) [89] to obtain semantic similarity scores between tags. We use these scores to obtain clusters of tags for places.

3.1.3 Interest Profile Expansion Algorithm

Users describe themselves and their interests on online social networking profiles. Such profiles are a great source of information about the user, but they often contain only few short textual snippets or keywords. Such explicitly created profiles represent the interests of the users related to various aspects of their lives. Many of the field

---

[2]http://www.wikipedia.org/

values are textual descriptions such as *I love the smell of rain*, which are inherently ambiguous and complex to analyse. It has been found that user profiles in OSNs have very few fields under various categories such as *work*, *interests*, and *education* and have considerable textual descriptions which are complex to analyse [90, 91]. In [85, 92], various ways of representing and analysing user profiles have been discussed. We present an interest expansion algorithm that removes ambiguous concepts and expands an initial set of user interests. The expansion is done in a way to derive hidden related concepts, without deviation from the initial interests. These observations demand the disambiguation of keywords in user profiles and expansion of the sparse user profiles submitted by the users to get unambiguous and richer user interest profiles.

Our expansion algorithm uses Wikipedia to disambiguate and expand the interest profile of a user. Wikipedia is a vast repository of knowledge constantly updated and refined by a large user community. It has the advantage that all the concepts defined are rich in their article content with numerous links to related concepts. The concepts and the links between them form Wikipedia graph structure where concepts represent the nodes and links represent the edges. We use the term concept and node interchangeably in the work. In order to get a disambiguated user profile, we retain only those keywords which match to a single Wikipedia concept and discard remaining keywords so that a modified user profile contains unambiguous concepts. We disambiguated the interest profiles mentioned in Section 5.1.1 in this way and found that 20% of the keywords in user profiles matched to an exact Wikipedia concept. This showed that Wikipedia concepts can be used to represent social interest profiles of users.

Next we apply our user interest profile expansion algorithm to expand the disambiguated profile. The algorithm considers the fact that a Wikipedia concept can be associated with its *related* concepts based on the links to concepts in its content on Wikipedia. A Wikipedia concept is any entity which has an article body an has one more links to other articles. The algorithm also takes into account the fact that concepts with a large number of inlinks from other concepts tend to be more general as noted by Gabrilovich et al. [93] and hence does not include such concepts in the expansion. This ensures that general concepts such as *Education* and *United States* which have high indegree are not present in the expanded profile and hence not used as tags for places.

Algorithm 1 describes how the expansion is done. It considers each concept in the user profile and attempts to expand it in a depth first manner. The parameters R and R_{glob} control the expansion of any node by limiting the number of nodes

that can be expanded. The parameter $Indeg_{threshold}$ defines the maximum number of inlinks that a concept can have so that its not considered to be a general concept. The *distance* function computes the shortest distance between any two concepts which is the minimum number of links to be traversed from one concept to the other in Wikipedia graph structure. The set of neighbour nodes which would be expanded from a given node is decided by the proximity of those nodes to the nodes in W. The measure of proximity of a node u is stored in $r[u]$ as seen in the algorithm. For a given node, the algorithm only expands those nodes that are closest to the set of nodes in W. This ensures that only those nodes more related to the original interests of a user are expanded further.

A node v_i is expanded only if $\prod_{k=i-1}^{0} \frac{1}{outdegree(v_k)} \geq R_{glob}$ where i is the height of the node v_i in the expansion tree and $v_{i-1}, v_{i-2}, \ldots v_0$ represent the ancestors of v_i in the expansion tree. During the expansion j^{th} node v_{ij} at height i, at most N_{ij} neighbours are added to the expansion list which are at unit distance from v_{ij} in Wikipedia graph. At most k nodes are considered for expansion from any given node. So, the maximum number of nodes added due to the expansion of a node is $M_0 + M_1 + M2 \ldots + M_h$ or $O\left(\sum_{i=0}^{h} M_i\right)$, where $M_i = \sum_j N_{ij}$ and h is the maximum height possible for all the non-leaf nodes in the expansion tree. For any M_i, neighbours of at most k^i nodes are considered. The result from the interest profile expansion algorithm for a user i corresponds to the set K_i in the probabilistic model. In the next section, we evaluate how both approaches combined can provide meaningful descriptive tags for places.

3.2 Experimental Evaluation

We have performed an experimental evaluation in order to verify the effectiveness of our approach. We first describe the real world datasets used in the experiments, and then present the results of our evaluation. The evaluation is divided into different parts. We report on the expansion algorithm, the parameters used and the distribution of the profile sizes. We show how the assigned tags evolve with the increasing number of user check-in activities and how they compare to a set of manually assigned tags. Finally, we analyse the nature of the tag probability distributions which indicates that the set of automatically generated tags is expected to stabilise with the increasing number of unique user check-ins. The results show that though not many semantically related tags can be derived with check-ins by 1 to 3 unique visitors on an average, as more users check-in, we can derive more tags semantically related to places. We also found that despite the increasing number of unique users checking in at a place, the

---

Algorithm 1 Interest Profile Expansion

---

function EXPANDPROFILE(W)

    $U \leftarrow \phi$

    for all $c \in W$ **do**

        AddNode(c,1,W);

    end for

end function

function ADDNODE(v, R, W)

    if $R \geq R_{glob}$ **then**

        $N \leftarrow \{u \mid dist(u,v) = 1\}$

        for all $u \in N$ **do**

            if $indegree(u) < Indeg_{threshold}$ **then**

                for all $c \in W$ **do**

                    $r[u] \leftarrow r[u] + distance(c,u) + distance(u,c);$

                end for

                add(u,U);

            end if

        end for

        for all $t \in$ *TopKNeighbor(r)* **do**

            AddNode($t, R * 1/|outdegree(v)|, W$);

        end for

    else

        return;

    end if

end function

---

tags are still semantically close to the places. Moreover, the results show that both automatically generated tags and manual tags are equally "semantically close" to a particular place.

3.2.1 Dataset Description

We collected data from Foursquare[3] for over one million random places in UK, USA and Ireland between June and July 2012, to check how well the places are described. Only 7% of the places had any descriptive tags and only 21% of the places had any tips/comments in the form of short text snippets, which again confirmed the lack of rich description of places.

We then collected Facebook[4] and Foursquare user profiles of 104 volunteers residing in the city of Galway, Ireland. These were random users as we requested

---

[3]https://foursquare.com/
[4]http://www.facebook.com/

people to participate through various social media and announced prizes for their contribution. The social interests of the users were obtained from their Facebook profiles by extracting the text in the fields corresponding to hometown, interests, activities, education, work, and events. We have found that interest profiles were sparse in terms of the keywords and our observations are indeed similar to the figures stated in [90, 91]. The size of the user profiles in terms of number of keywords can be fit with a Poisson distribution using a *Maximum Likelihood Estimation* (MLE) ($n = 104$, $\lambda = 362.1$, $S.E = 1.9$) as shown in Figure 3.3(a). We have obtained check-in activities from both Foursquare and Facebook profiles of the volunteers. The check-in activity data contains 4476 records of check-ins of users which they had generated using their Facebook and Foursquare mobile applications. There are 1633 unique places where users had checked-in and 215 places where at least 2 users had checked-in.

3.2.2 Evaluation

Interest Profile Expansion Algorithm

In this subsection, we discuss the details of the interest profile expansion algorithm by discussing the values assigned to various variables in the expansion algorithm and the nature of expanded user profiles obtained by running the algorithm on user social profiles. For generating the values assigned to the different variables in the expansion algorithm we have proceed as follows. The fact that concepts with a large number of inlinks from other concepts tend to be more general [93] and hence the algorithm does not include such concepts in the expansion. We have analysed the graph structure of Wikipedia to understand the nature of inlinks among nodes. We have first sorted the concepts by the number of inlinks to them and manually inspected many of the top concepts. This has shown that indeed such concepts were very general in nature. Since we have not found any formal approaches to decide the generality of Wikipedia concepts, we discarded top 1% of the concepts and obtained the statistics for the inlinks of the remaining concepts. All the remaining concepts had very few inlinks ($n = 3537875$, $min = 0$, $max = 221$, $mean = 9.274$). Hence we set the value of $Indeg_{threshold}$ to 221 which ensured that nodes with more than 221 inlinks were not added during expansion. We set the expansion controller variable R_{glob} to $1/100$ which meant that a concept is expanded only if it has no more than 100 ancestors considered during the expansion. The expansion algorithm considerably enriched the user interest profiles with related concepts in Wikipedia. The expanded user interest profiles were significantly larger compared to their original size and we could fit the
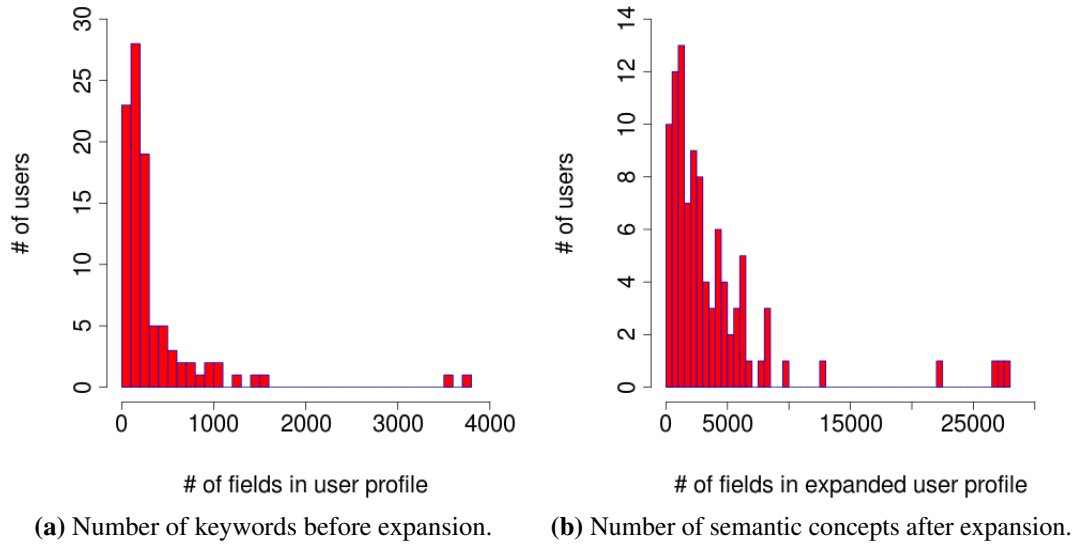
(a) Number of keywords before expansion.   **(b)** Number of semantic concepts after expansion.

Figure 3.3: Sizes of interest profiles before and after expansion.

size with Poisson distribution using MLE ($n = 104$, $\lambda = 3843.835$, $S.E = 6.295285$) as shown in Figure 3.3(b).

Automatic Semantic Tagging Results

For the automatic semantic tagging we have considered only those places which were checked-in by at least 2 users. For each place p, we have computed the random variable CT_p^n by incrementally considering the unique users who had checked-in at p. This process defined CHU_p number of random variables corresponding to tag probabilities for p. We then applied the hierarchical clustering method to obtain the clusters of tags corresponding to each random variable for each place.

In order to evaluate the quality of the derived tags, we have used a set of manual tags assigned by volunteers as ground truth. Another potential alternative we will consider in our future work will involve using the contents of appropriate web pages. Deriving a set of relevant tags for a place include implementation of a mapping method between any website and a place to use n-grams present in the web pages. In the current work, seven volunteers manually tagged the places they knew among the places in the collected check-in records. They tagged a total of 25 unique places with multiple tags (mean number of tags per place = 22.96). Manual inspection of automatically derived tags and manually assigned tags revealed that most of the tags in such clusters were highly related to the places under consideration, though users

research associate   coursework   graduate school   xkcd   techcrunch   social semantic web   research
assistant   dbpedia   college   graduate entry   semantic web services   sparql   mobile technology   microformat
   semantic web   foaf   rdf schema   researcher   rdfa   phd   research proposal   web science   sfi

(a) Manually assigned tags.

linked data   hcard   amazon mechanical turk   ntriples   defeasible reasoning   machinereadable medium   enterprise
information system   rdf schema   semantic web services   theme computing   google analytics   notation   semantic html
snomed ct   paraconsistent logic   foaf software   domain knowledge   description logic   sparql   business semantics
management   text mining   semantic publishing   entityattributevalue model   semantic web stack   semanticallyinterlinked
online communities   resource computer science   probabilistic logic   smartm   computational semantics   turtle syntax
timo honkela   metacrap   resource web   latent semantic analysis   semantic web   giant global graph   embedded rdf
semantic advertising   semantic sensor web   glossy display   website parse template   corporate semantic web   grddl
conrad wolfram   social semantic web   rule interchange format   principle of explosion   semantic computing   rdfa
nextbio   ontology learning

(b) A cluster of tags derived from top-probable tags after check-ins by 10 users.

Figure 3.4: Manual and derived tags assigned to Digital Enterprise Research Institute

had not tagged places with the derived tags. Figure 3.4 shows both the manual and automatically derived tags for Digital Enterprise Research Institute (DERI), a Semantic Web research institute, hosted at NUI Galway, Ireland. Frequent manual tags and the most probable tags are shown in larger fonts. We can see that though automatically derived tags are not exactly the same as the manual tags, they are good candidate tags for DERI. Similar observations can be made in Figure 3.5 James Hardiman Library at NUI Galway, Ireland.

nui galway students union   math   michael d higgins   sin newspaper   geology   library   science
festival   history   engineering   union of students in ireland   philosophy   study   news papers   student
union   coding   student   rag student society   nuig   arts   rag society

(a) Manually assigned tags.

nui galway students union   volvo ocean race   trinity news   sin newspaper   senior british open
championship   tyrone crystal   union of students in ireland   flint glass   rag student society   afca
national championship trophy   world junior championships in athletics   the university observer   lead
glass   going on   whitbread   the first post   edinburgh crystal   oxygenie student media award   philips
lumileds lighting company   charles bacik   the college view   ivana bacik   michael d higgins   the
university times   allireland   ronan ogara   waterford wedgwood   damien dempsey   royal naval sailing
association   geraldine kennedy   quantcast   irish hospitals sweepstake   kps capital partners
rouzbeh rashidi   jasper conran   suicide prevention   joseph mcgrath politician

(b) A cluster of tags derived from top-probable tags after check-ins by 3 users.

Figure 3.5: Manual and derived tags assigned to James Hardiman Library, NUI Galway.

Automatic Semantic Tagging Evaluation

In the previous subsection, we discussed that online social interest profiles of users can have arbitrarily long text sequences and hence described a method to derive Wikipedia concepts in order to obtain disambiguated interest profiles. We also obtained manually assigned tags from few volunteers for places described in the previous subsection. However, few volunteers and the handful of tags they contributed meant that robust

ground truth was not available for our evaluation. So, we had to evaluate the quality of manually assigned tags against corresponding places and then use subsequent inference as baseline to evaluate the quality of automatically derived tags. For a systematic evaluation of the generated tags, we have measured the Normalised Web Distance (NWD) described by Gracia et al. [94] between derived tags, manually assigned tags and the place names. We have then studied the statistical distribution of NWD scores of manually assigned tags and compared it NWD scores obtained by derived tags. We describe our findings in the next few paragraphs.

Normalised Web Distance has been extensively used to obtain the semantic relatedness between any two strings, where the extensive data on the Web is used. Formally, the NWD between any two strings x and y is given as

$$d_{nwd}(x,y) = \frac{max\{logf(x), logf(y)\} - logf(x,y)}{logN - min\{logf(x), logf(y)\}} \tag{3.4}$$

where $f(x)$ is the number of Web pages containing the string x, $f(y)$ is the number of Web pages containing the string y, $f(x,y)$ is the number of pages where both x and y appear, and N is the total number of pages indexed by a specific search engine.

We first analyse how different users visiting a place affect the set of generated tags. For each random variable T_p^n, we have computed the d_{nwd} between the top 350 automatically derived tags and place names using the index provided by *Yahoo*[5]. It is possible that some tags have an infinite NWD to a place, which were considered as invalid and discarded. Figure 3.6(a) shows the box plot of number of valid tags, i.e. tags with a finite NWD, over all places. Please note that, for instance, for the case of 6 users, only places which have at least 6 distinct users were considered. We can see that the more unique users check-in at places, the more valid tags are generated.

We then computed the values of d_{nwd} between place names and the manually assigned tags to compare the performance of our semantic tagging technique. The five-number summary of d_{nwd} between manual tags and place names is (min=0.0000, $Q1$=0.1216, $median$=0.3032, $Q3$=0.8732, max=1.9030) with $mean$=0.45730. The five-number summary of d_{nwd} between automatic tags and place names, considering all check-in activities, is (min=0.0000, $Q1$=0.2053, $median$=0.5340, $Q3$=1.0000, max=3.4930) with $mean$=0.5719. This shows that automatic tags exhibited d_{nwd}

---

[5]http://developer.yahoo.com/search/boss/
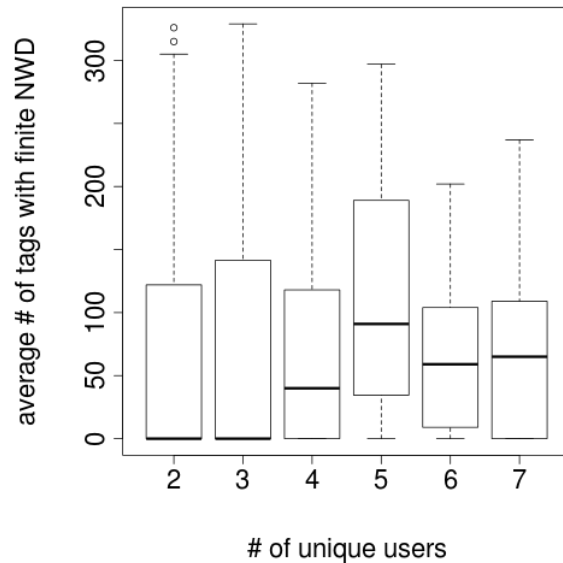
(a) Number of derived tags with finite NWD.
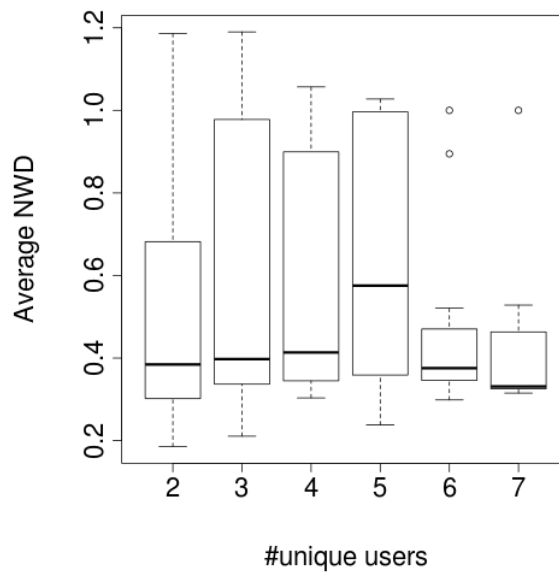
(b) Average NWD scores of tags.

Figure 3.6: Variation in the Normalised Web Distance scores against the number of unique users.

values comparable to those of the manual tags. The Welch's t-test showed that mean value of d_{nwd} for automatic tags is greater than that of manual tags with 95% confidence interval of (0.053, 0.144) where H_A is that true difference in means is not equal to 0. This means that on an average, the d_{nwd} scores obtained by automatically derived tags are not much higher than the ones obtained by the manual tags.

Figure 3.6(b) shows the average values of d_{nwd} for the valid tags obtained against the number of unique users. We can see that in spite of more unique users visiting a place, the average scores of d_{nwd} obtained by the tags remain close to the ones achieved by manually assigned tags.

We noted that we could derive an average of 158 tags for places with expanded user profiles whereas we could derive 51 tags with unexpanded profiles. We also observed that only 9% tags obtained from expanded interest profiles had infinite values of d_{nwd} against places whereas this was 17% for the unexpanded user profiles. This clearly indicated the advantages of carefully expanding the concepts in user profiles and using them as probable tags. Clustering the top-probable tags obtained from expanded user profiles showed that 30% of the tags belonged to some cluster and were related to each other and only 2% of the tags had infinite normalised web distance. 70% of the tags did not belong to any cluster and were not related to each other and 8% of such tags had infinite normalised web distance. This showed that clustering the tags fetched tags related to each other and to the place thereby removing any 'noise' tags among the top-probable ones.

In the previous paragraphs, we have shown that large number of semantically relevant tags can be derived for places. However, it is necessary to study two crucial factors that can affect the quality and quantity of derived tags.

- The number of users whose profiles are utilized for deriving tags

- The stability of the tags that are inferred as relevant after considering a specific number of social profiles

This demands the analysis of tag probability distributions. So, in the next few paragraphs, we discuss the nature of tag probability distributions for various places in the current data set.

Nature of the Tag Probability Distributions

We have studied the nature of the tag probability distributions of a place over the number of unique visitors of that place. We considered only those places which had been checked-in by at least 5 distinct users to study the variation in the tag probability distributions. We have computed the entropy [95] to analyse the information content or *randomness* of tag probability distributions, and we have used Jensen-Shannon divergence [96] to analyse the variations among tag probability distributions.   The
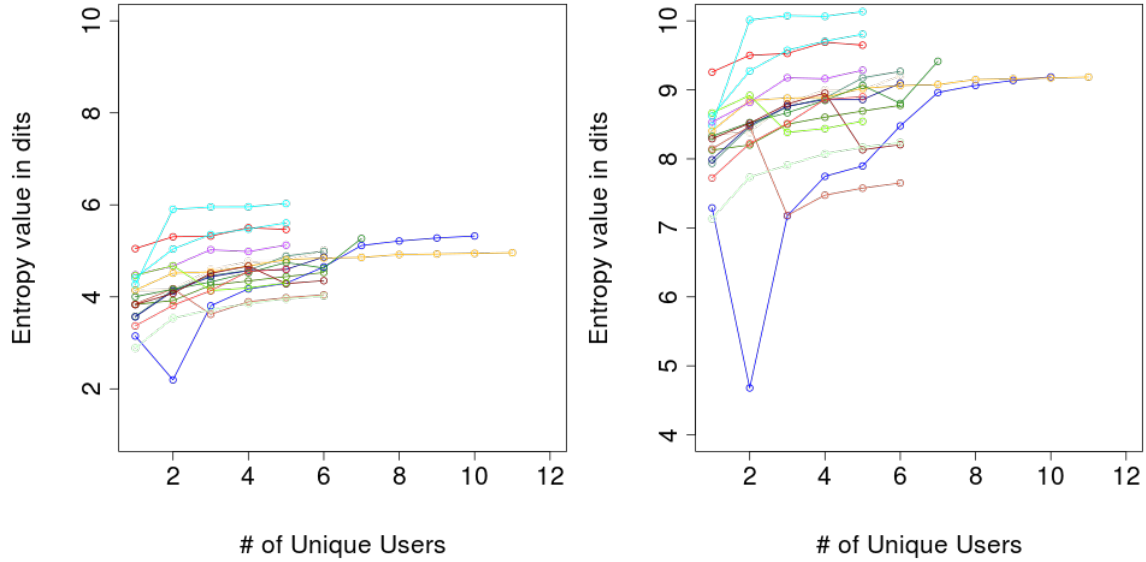
(a) Entropy values with unexpanded interest profiles.   **(b)** Entropy values with expanded interest profiles.

Figure 3.7: Entropy values observed over the tag probability distributions w.r.t. the number of unique visitors.
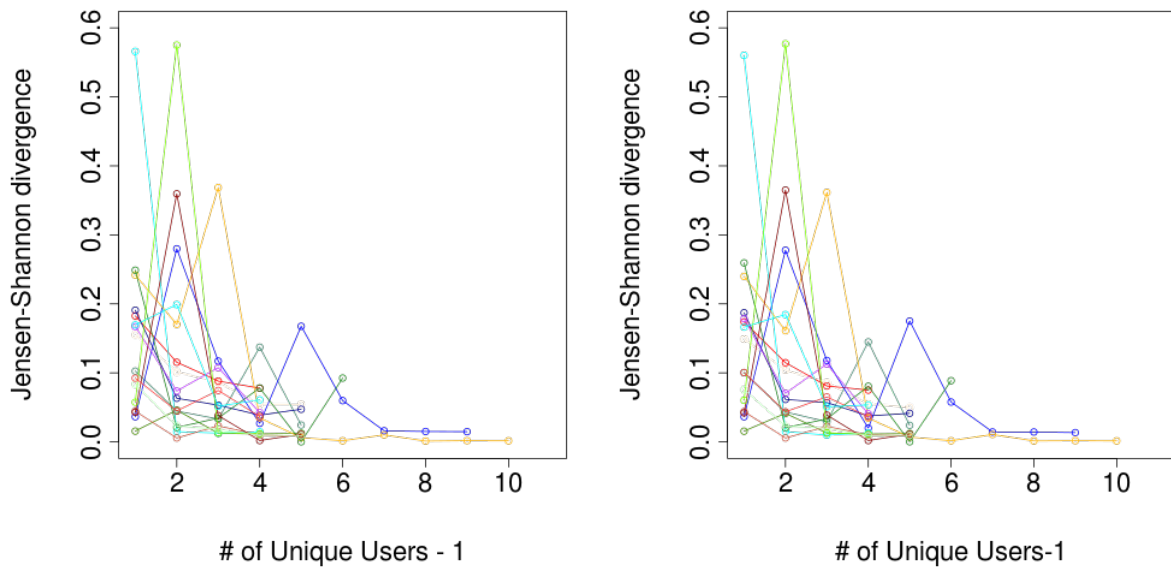
(a) Jensen-Shannon divergence values with unexpanded interest profiles.

(b) Jensen-Shannon divergence values with expanded interest profiles.

Figure 3.8: Jensen-Shannon divergence w.r.t. the number of unique visitors.

entropy E of a discrete random variable X is defined as

$$E(X) = -\sum_{1}^{n} Pr(x_i) log(Pr(x_i)) \tag{3.5}$$

where x can assume n discrete values with the probabilities $Pr(x_1), Pr(x_2), \ldots, Pr(x_n)$.

We depict the variation in entropy of T_p^n when unexpanded user profiles are considered in Figure 3.7(a). As a reminder, T_p^n is a random variable representing tag probability distributions for place p that is derived by considering social profiles on n users that have visited that place. Figure 3.7(b) shows the variation in entropy when expanded user profiles are considered. Number of unique users that have checked in for places varies. So, we see that entropy lines for some of the places stop at specific points on x-axis which represents unique number of users that have checked in at a place. We see that the increase in the entropy values is lesser after more unique users check-in. This indicates that the information content of T_p^n does not increase in spite of increased sample space and stabilises with the number of unique users visiting place p. It also implies that some of the semantic tags become more probable and thereby reduce the entropy in spite of increased sample space, CT_p^n defined in subsection 5.1.2.

Given any two probability distributions P and Q, the Kullback-Leibler divergence can be used to measure the statistical dependence between them. Since this is an asymmetric measure and can have infinite values, the Jensen-Shannon divergence can be used to measure the dependencies between P and Q which gives finite and symmetric values of dependencies. The Jensen-Shannon divergence between two discrete random variables P and Q is defined as

$$JSD(P||Q) = \frac{KLD(P||M)}{2} + \frac{KLD(Q||M)}{2} \qquad (3.6)$$

where KLD is the Kullback-Leibler divergence and $M = \frac{P+Q}{2}$ is the mean distribution. The Kullback-Leibler divergence between any two discrete random variables P and Q is defined as

$$KLD(P||Q) = \sum_i P(i) ln \frac{P(i)}{Q(i)} \qquad (3.7)$$

and is defined for any non-zero values of $P(i)$.

If two random variables are highly dependent, the Jensen-Shannon divergence value between them tends to be small.

We computed the Jensen-Shannon divergence between T_p^n and T_p^{n+1}. We show how the divergence value diminishes based on the number of unique users in Figure 3.8(b) when expanded user profiles are considered. Interestingly, the divergence values obtained for the random variables when expanded profiles were used are very similar to the ones corresponding to the unexpanded profiles and are shown in Figure 3.8(a). This indicated that in spite of considering various interests of users to derive tag probability distributions of a place, such distributions showed high dependence as interests of more users were considered. This means that few of the tags will consistently emerge as the most probable tags when more social profiles are considered.

3.3 Discussion

In this chapter, we have presented an algorithm to automatically derive descriptive semantic tags for places, based on users' interests found in online social network profiles and their check-in activities. We have shown that highly semantically relevant tags for places can be derived using these data. Specifically, we derived from each user a set of concepts based on the user interests, using our interest profile expansion algorithm. The sets are used in our probabilistic model together with the hierarchical clustering techniques to derive a set of tags for a place, based on the users that have visited the place. We performed an experimental evaluation that shows that not only we are able to automatically derive meaningful tags for different places, but also that the sets of tags assigned to places are expected to stabilise with the increasing number of user check-ins. In the future work, we will obtain larger datasets to validate our findings rigorously. We will also consider other online sources of user data, such as Twitter which represent the real time interests of people and interactions at places. Determining the appropriate number of tags to be annotated for a place based on the nature of the place is another crucial future work involved.

In this chapter, we used the check-ins data and social profiles of users for deriving place tags. However, some of the data sources might lack these types of physical (check-ins) and virtual (social interest profiles) data of users. But, many data sources capture time series data of approximate number of people at places using physical sensors such as motion detectors and virtual sensors such as check-in counts on online networks. In the next chapter, we describe a highly accurate probabilistic model to infer about events organized at places using time series of count data representing the number of people at a place.

Chapter 4

Event Detection at Places Based on User Presence

The recent emergence of the Internet of Things (IoT) demonstrates that large amount of contextual data about places can be captured by deploying sensors at locations. Time series data generated by sensors and Web users can be a potential data source to infer environmental and social contexts at places. On the other hand, events conducted at places signify the importance of a place and constitute it's metadata. So, inferring about events conducted at places enables us to automatically annotate places with event metadata. Specifically, detecting outliers that have unusually high values in a time series data is crucial for inferring about any events in the real world. In this work, we describe an infinite Poisson mixture model to detect events by identifying outliers in time series of count data. This unsupervised technique estimates the probability densities of count data which have an unknown Poisson mixture while it simultaneously detects outliers in the data. The advantage of our model is that outliers are mapped to mixture components discovered by infinite mixture model and thus inference can be drawn on the different 'types' of outliers and their proportions in the data. This lets us identify and categorize events based on magnitude of outlier data. We have analysed the performance of our model against a well-known event detection technique based on Markov modulated Poisson process (MMPP) using synthetic and real world data. Results show that our approach to detecting events is more appropriate in analysing periodic count data as compared to the MMPP baseline. The experiments demonstrate that the presented model provides robust, detailed, and interpretable results for the analysis of outliers to detect events.

In recent years, there has been a surge in the real world data generated by sensors on mobile phones as well as sensors embedded into various physical infrastructures. Data from GPS sensors, accelerometers on mobile phones capture fine-grained activities of individual users. On the other hand, sensors such as RFID tags, traffic polling systems, motion detectors for buildings etc. generate data, based on activities of population. These huge amounts of real world data have been extensively used by researchers to understand specific patterns of human activity. Particularly, analysing unusual trends in behaviour exhibited by users in the context of hourly, daily and weekly periodic variations is crucial to gain useful insights about real world situations. For example, a popular cultural event held in a city would be reflected in the amount of vehicles entering the toll gate of a city. Similarly, a minor increase in the number of vehicles entering a city can be due to a less famous event. These events in the real world can be mapped to unusually high count values in the corresponding time series data of user visits. Detecting such events in an unsupervised manner demands detecting outliers in time series of periodic count data. Also, it is crucial that outliers are classified into various categories based on their values. This lets us identify and categorize events based on the magnitude of increased counts due to them.

Modelling univariate data and detecting outliers has been extensively studied in the field of statistics and machine learning. In [97], Ihler et al. demonstrate the need for careful analysis of time series data when such data is generated by user activities is shown and a model based on non-homogeneous Markov Modulated Poisson Processes (MMPP) has been used for analysis. Similar models based on different versions of MMPP have been used in [15, 98, 99]. The motivation behind MMPP based outlier detection models is the fact that univariate time series data has no clear boundary to differentiate *abnormal* high count value outliers from the *normal* periodic counts. Any MMPP based outlier detection model has the assumption that most of the periodic count data can be modelled by a Poisson distribution with fixed rate parameter. This assumption about the nature of human generated periodic count data does not always hold good as they exhibit multimodality with unknown number of modes. We demonstrate this in Section 4.2 with two datasets. Detecting outliers in multimodal data which is represented with a probabilistic mixture model has been widely studied in the field of statistics. Yamanishi et al. [100] discuss the advantages of detecting outliers with finite mixture model. Specifically, the authors represent data with finite mixtures of Gaussian distributions. These finite parametric models require prior information about the data and hence are less flexible in accurately modelling the data.

On the other hand, nonparametric Bayesian models have high accuracy as they flexibly determine the *right* number of mixture components and mixture weights while estimating the probability density of data. So, they have been used effectively in various research works [65, 101, 102]. To the best of our knowledge, in spite of the flexibility of nonparametric Bayesian models, there has been no study on utilising them for detecting outliers in count data. Motivated by these observations, we have developed an infinite Poisson mixture model to capture the generative mechanism for count data and to identify outliers based on mixture components that are inferred. We have developed an appropriate version of Gibbs sampler, a Markov Chain Monte Carlo (MCMC) algorithm to perform the probabilistic inference. The research contribution made in this chapter and its advantages can be summarized as follows.

- We develop an infinite Poisson mixture model for estimating outliers in a count dataset based on the underlying mixture components. We also develop a Gibbs sampler algorithm for parameter learning and inference.

- The proposed model has some major advantages over MMPP based baseline technique and other traditional outlier detection techniques. It is fully unsupervised and performs accurately even in cases where data has varying degrees of multimodality. It identifies outliers in a dataset and also categorizes them into a previously unknown number of outlier categories. Since outlier and non-outlier data are represented by probability densities, sophisticated probabilistic queries can be made. Finally, any prior information about outliers in a dataset can be incorporated into the model by appropriately choosing prior distribution parameters.

- We apply our outlier detection technique on a real world dataset to detect events in an unsupervised manner. We show that our technique has better accuracy and provides robust analysis.

The rest of the chapter is organized as follows. We first discuss the state of the art work to detect outliers in count data sets. In Section 4.1, we discuss our probabilistic model for detecting outliers in any count dataset. The model detects outliers of unusually high count in a dataset by discovering mixture components of an unknown Poisson mixture density. In Section 4.2, we analyse the performance of the proposed model against a state-of-the-art version of MMPP-based outlier detection technique using synthetic and real world data. In the last section of the chapter, we state our conclusions and discuss the future work.

Bayesian parametric framework for outlier detection has been studied extensively in statistics and machine learning [103–106]. Particularly, a non-homogeneous Markov Modulated Poisson Process (MMPP) based outlier detection techniques have been developed for situations where events need to be detected based on data generated by periodic human activities. In this approach, periodic count data is modelled by Poisson process regulated by a Markov chain with a fixed number of states [97]. Effectiveness of the approach has been studied on vehicular traffic data and user presence data to detect events. In [15], Pozdnoukhov et al. model the number of tweets[1] generated by users and exploit MMPP to quantify abnormal volume of tweets. Modelling and predicting voice packet data with a version of MMPP has been studied by Heffes et al. [98]. Scott et al. [99] study the identification network intrusion using MMPP. These works have analysed how variations of MMPP model can be used to detect outliers and thus unusual bursts in human activity which correspond to events. Traditional models such as MMPP based techniques have assumed fixed number of Poisson mixture components in time series data for estimating outliers. This assumption can lead to either overfitting or underfitting of data depending on number of mixture components used in the model.

On the other hand, Yamanishi et al. [100] discuss the advantages of detecting outliers with finite mixture model. Specifically, the authors represent data using a finite mixture of Gaussian distributions as follows and show that outliers can be detected.

$$f(x) = (1 - \epsilon)f_n(x) + \epsilon f_0(x) \tag{4.1}$$

$$f_n(x) = \sum_{i=1}^{K-M} \pi_i \phi(x|\mu_i, \sigma_i^2) \tag{4.2}$$

$$f_0(x) = \sum_{j=K-M+1}^{K} \pi_j \phi(x|\mu_j, \sigma_j^2) \tag{4.3}$$

Here, ϵ is a small positive fraction which represents the proportion of outliers. $f_n(x)$ is the probability density of data that are non-outliers and $f_0(x)$ is probability density of outlier data. However, the numbers of mixture components representing non-outlier and outlier data are fixed in this model using K. This is a limited assumption as number of mixture components in a dataset cannot be known a priori. So, outlier detection solutions based on previously described finite mixture models face the problem of either underfitting, if few mixture components are assumed or overfitting otherwise. This potentially leads to less accuracy in detecting outliers. Interestingly, the work in

---

[1]https://twitter.com/

[69] considers real world data and shows the need to use an infinite mixture model to fit count data more accurately compared to a parametric model. Similar results have been presented in [65, 101] and numerous other works. These observations motivate us to use an infinite mixture model for outlier detection in any univariate dataset.

4.1 Generative Model for Outlier Detection in Poisson Mixture Data

In this section, we discuss our generative model that can detect outliers by nonparametrically fitting a count dataset and deriving the modes that generate anomalies of atypically high values. The model is based on the standard results of dirichlet process mixture model [107, 108] for nonparametric density estimation of a given dataset. In Figure 4.1, we show the generative model in plate notation. Note that we use the term mixture component and cluster interchangeably in the following discussion. The random variables in this nonparametric Bayesian model are defined as follows.

$$p|a,b \sim Beta(a,b) \quad e_i|p \sim Bernoulli(p)$$
$$\pi|\alpha \sim GEM(1,\alpha) \quad z_i|\pi \sim Discrete(\pi)$$
$$\theta_k|\lambda \sim G_0(\lambda) \quad\quad\quad \theta_l'|\lambda' \sim G_0(\lambda')$$

$$x_i|e_i, z_i, \{\theta_k\}_{k=1}^{\infty}, \{\theta_l'\}_{l=1}^{\infty} \sim e_i F(\theta_{z_i}') + (1-e_i)F(\theta_{z_i})$$

$$(4.4)$$

Here, we consider N items $\mathbf{x} = \{x_1, x_2, \ldots, x_N\}$ which are discrete count data from an unknown Poisson mixture. p is a random variable which has a Beta prior distribution with a, b as hyperparameters. p represents the probability that a count value in the dataset is an outlier with unusually high value. e_i is a Bernoulli random variable which would be 1 when the data x_i is an outlier. π is a sample from a stick breaking process [67] denoted by GEM with the parameter α. It represents the discrete distribution over the cluster indices of the parameter vectors that exist for the mixture component of the dataset. z_i represents the index of cluster chosen for the data x_i. θ_l' represents a sample from the base distribution with parameters λ' and θ_k is a sample from the distribution parametrized by λ. θ_l', θ_k represent the Poisson rate parameter for outlier and non-outlier clusters respectively. Since we need to model an unknown Poisson mixture data, we have used Gamma distribution as the base distribution G_0 which is the conjugate prior distribution for Poisson distribution. Thus, λ' and λ are
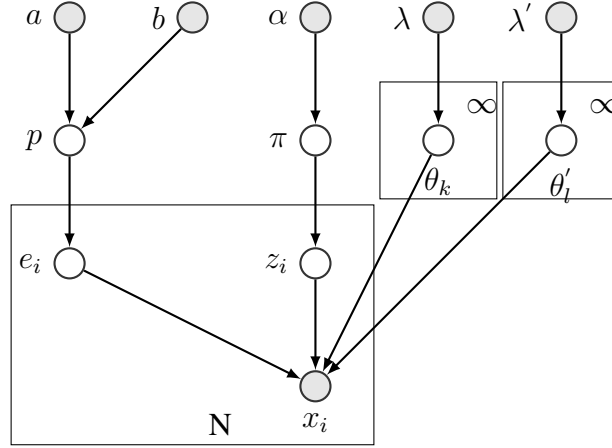
Figure 4.1: Generative model for detecting the outliers in periodic count dataset

hyper parameter vectors defined as $\lambda' = (shape', rate')$ and $\lambda = (shape, rate)$ for θ'_l, θ_k respectively. The data x_i have the Poisson likelihood given by the last statement in the Equation (4.3). x_i would be generated by a Poisson distribution with rate parameter θ'_l if it is an outlier data and by a Poisson distribution with rate parameter θ_k otherwise. The data $\mathbf{x}$ is modelled with two infinite mixtures as we need to model an unknown number of non-outlier and outlier clusters present in the data. In Figure 4.1, we show the Bayesian network for the generative model described. Here, the hyper parameters whose values are known are marked dark and random variables with unknown distribution are marked white. The parameter learning and inference for this network involves obtaining the joint distribution of the random variables. We use one of the widely adopted Gibbs sampling, a Markov Chain Monte Carlo (MCMC) algorithm which requires the full conditional distributions for every random variable involved in order to obtain samples from the joint posterior distribution. We now show the conditional distributions derived for each of the random variables.

$$
\begin{aligned}
&Pr(e_i = 1 | x_i, \mathbf{e}_{-i}, a, b, z_i, \{\theta_k\}_{k=1}^{\infty}, \{\theta'_l\}_{l=1}^{\infty}) \\
&= Pr(e_1 = 1 | \mathbf{e}_{-i}, z_i, \{\theta_k\}_{k=1}^{\infty}, \{\theta'_l\}_{l=1}^{\infty}) \\
&\quad Pr(x_i | e_i = 1, z_i, \{\theta_k\}_{k=1}^{\infty}, \{\theta'_l\}_{l=1}^{\infty}) \\
&= Pr(e_i = 1 | \mathbf{e}_{-i}, a, b) F(x_i | \theta'_{z_i})
\end{aligned}
\tag{4.5}
$$

Here $\mathbf{e}_{-i}$ denotes the set with all e_j, $j \neq i$. We use similar notation for $\mathbf{x}_{-i}$, $\mathbf{z}_{-i}$. $\mathbf{x}_k$ denotes the set of data items belonging to k^{th} cluster. $\mathbf{e}$ denotes the set with all e_j. We

can derive the conditional distribution of **e** as follows.

$$Pr(p|a,b) = p^{a-1}(1-p)^{b-1}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$$

$$Pr(e_1,\ldots,e_n|p) = p^{\sum e_i}(1-p)^{n-\sum e_i}$$

$$Pr(e_1,\ldots,e_n|a,b)$$

$$= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int p^{(\sum e_i+a-1)}(1-p)^{(n-\sum e_i+b-1)}dp$$

$$= \frac{\Gamma(a+b)\Gamma(\sum e_i+a)\Gamma(n-\sum e_i+b)}{\Gamma(a)\Gamma(b)\Gamma(n+a+b)}$$

(4.6)

$$Pr(e_i=1|\mathbf{e}_{-i},a,b) = \frac{Pr(\mathbf{e}|a,b)}{Pr(\mathbf{e}_{-i}|a,b)}$$

$$= \frac{\Gamma(a+\sum \mathbf{e}_i)\Gamma(n-1+a+b)}{\Gamma(a+\sum \mathbf{e}_{-i})\Gamma(n+a+b)}$$

$$= \frac{a+\sum \mathbf{e}_{-i}}{a+b+n-1}$$

(4.7)

In a similar fashion to Equation (4.7) we can derive that

$$Pr(e_i=0|\mathbf{e}_{-i},a,b) = \frac{b+n-\sum \mathbf{e}_{-i}-1}{a+b+n-1}$$

(4.8)

$$Pr(e_i=0|x_i,\mathbf{e}_{-i},a,b,z_i,\{\theta_k\}_{k=1}^\infty,\{\theta_l'\}_{l=1}^\infty)$$

$$=Pr(e_i=0|\mathbf{e}_{-i},a,b)F(x_i|\theta_{z_i})$$

(4.9)

similar to Equation (4.5). We use the Chinese Restaurant Process proposed by Aldous et al. [109] as the basis for deriving the clustering property of the dataset. The probability that a data item falls into one of the active cluster j with parameter ϕ is derived as follows. Note that ϕ is equal to an existing value of θ_k if the cluster j is marked as non-outlier cluster and θ_k' otherwise.

$$Pr(z_i=j|\mathbf{x},\alpha,e_i=1,\mathbf{z}_{-i},\{\theta_k\}_{k=1}^\infty,\{\theta_l'\}_{l=1}^\infty)$$

$$= Pr(z_i=j|x_i,\alpha,\mathbf{z}_{-i},\phi)$$

$$= Pr(z_i=j|\alpha,\mathbf{z}_{-i},\phi)Pr(x_i|z_i=k,\alpha,\mathbf{z}_{-i},\phi)$$

$$= Pr(z_i=j|\alpha,\mathbf{z}_{-i})Pr(x_i|\phi)$$

$$= \frac{n_{j,-i}}{n+\alpha-1}F(x_i|\phi)$$

(4.10)

When the data belongs to an outlier cluster with $e_i = 1$, the probability that z_i gets a new value of cluster index $K + 1$ where already K clusters exist is given by

$$
\begin{aligned}
&Pr(z_i = K + 1 | \mathbf{x}, \alpha, e_i = 1, \mathbf{z}_{-i}, \lambda, \lambda') \\
&= Pr(z_i = K + 1 | x_i, \alpha, \mathbf{z}_{-i}, \lambda') \\
&= Pr(z_i = K + 1 | \alpha, \mathbf{z}_{-i}, \lambda') Pr(x_i | z_i = K + 1, \alpha, \mathbf{z}_{-i}, \lambda') \\
&= Pr(z_i = K + 1 | \alpha, \mathbf{z}_{-i}) Pr(x_i | \lambda') \\
&= \frac{\alpha}{n + \alpha - 1} \int F(x_i | \theta') G_0(\theta' | \lambda') d\theta'
\end{aligned}
\tag{4.11}
$$

The conditional distributions for the parameter vectors of the data can be determined as follows.

$$
\begin{aligned}
Pr(\theta_l' | \mathbf{x}, \mathbf{e}, \mathbf{z}, \lambda') &= Pr(\theta_l' | \mathbf{x}_l, \lambda') \\
&\propto G_0(\theta_l' | \lambda') \mathbf{L}(\mathbf{x}_l | \theta_l')
\end{aligned}
\tag{4.12}
$$

If a new cluster is formed for generating outlier data x_i, the conditional distribution for the parameter of that cluster θ_l' is derived as follows.

$$
Pr(\theta_l' | x_i) = \frac{G_0(\theta_l') F(x_i | \theta_l')}{\int G_0(\theta') F(x_i | \theta')}
\tag{4.13}
$$

We can obtain the conditional distributions for the cluster indices and parameters when $e_i{=}0$ in a similar fashion described in Equations (4.9) - (4.13).

The Gibbs sampler is used to derive the posterior samples of random variables in question using the full conditional distributions described above. Specifically, at each MCMC sample, we infer on the number of mixture components or clusters in the data and their weights and parameter values. Additionally, we label each of the inferred clusters as outlier or non-outlier cluster and the data belonging to those clusters as outlier data and non-outlier data respectively. We give the details of the direct Gibbs sampling in Algorithm 2 that we use for the inference. Here, we initialise two clusters where one represents the non-outlier data and the other represents outlier data. The Poisson rate parameters for these two clusters are initialised using the minimum and maximum values of the dataset. All data items are initially assigned to non-outlier cluster and are marked as non-outlier data items. The hyper-parameters of Gamma distribution are initialised using the simple statistics of the data, namely the maximum and minimum. A Gamma distribution with shape parameter k and rate parameter 1 has a mean and variance value as k. The restriction on the Gamma prior distribution for

drawing the Poisson rate parameters for outlier clusters thus results in the formation of outlier clusters with Poisson rate parameters biased towards the maximum value in the dataset. Similarly, Poisson rate parameters of non-outlier clusters are biased towards the minimum value in the dataset.

At each iteration of the algorithm, a data item is labelled as an outlier or non-outlier. The data item is then assigned to one of existing clusters or assigned to a new cluster with certain probabilities. If data item is assigned to a new cluster, the new cluster is marked as an outlier cluster if the assigned data item has been marked as outlier and as non-outlier cluster otherwise. The cluster parameters are updated for outlier clusters and non-outlier clusters using the samples from corresponding base distributions. After all the iterations, we use all samples to compute the number of clusters in the data using the posterior mode of number of active clusters in samples. The number of clusters is used to select matching samples which we use to compute outlier probability, cluster parameter values, cluster weights. We use the same samples to compute the posterior mode of the cluster outlier labels to identify clusters as outlier and non-outlier clusters. We then assign each data item to a cluster where it has the maximum probability density.

4.2 Experimental Evaluation

We have evaluated the performance of the presented outlier detection technique using synthetic and real world datasets. We first compared the effectiveness of our outlier detection technique against an appropriate version of MMPP based technique in identifying outliers and outlier probabilities for synthetic datasets. Then, we evaluated the performance in identifying real world events using the *buildings* event dataset mentioned in [97] by Ihler et al. Results show that the proposed technique is more accurate and does not require manual tuning of parameters as opposed to MMPP baseline.

4.2.1 Synthetic Datasets

We have considered four types of Poisson mixture data varying in the number of mixture components or clusters and dispersion. These have the typical distribution that time series data generated by periodic human activities have, as described in [15, 97]. We have used variance to mean ratio or dispersion index $D = \frac{\sigma^2}{\mu}$ to measure the dispersion in data. A statistical description of the datasets is as follows:

Algorithm 2 Gibbs Sampling for outlier detecting infinite mixture model

---

1: **function** INITIALISEPARAMETERS($\mathbf{x}$)
2:     $\mathbf{e} \leftarrow 0$
3:     $theta'_1 \leftarrow maximum(\mathbf{x})$
4:     $theta_1 \leftarrow minimum(\mathbf{x})$
5:     $(rate', rate) \leftarrow (1, 1)$
6:     $(shape', shape) \leftarrow (max(\mathbf{x}), min(\mathbf{x}) + 1)$
7: **end function**
8: **function** GIBBSSAMPLER($\mathbf{x}$,e,$\mathbf{z}$,θ)
9:     IntialiseParameters($\mathbf{x}$)
10:    At any iteration t with $t > 1$ of the sampling
11:    For $i = 1, \ldots, n$
12:    **if** x_i is the single element in its cluster **then** remove the cluster, it's parameter and
       decrement C by 1
13:    **end if**
14:    Sample e_i with the following probabilities

> Choose a non-outlier cluster parameter$\theta_{i(t-1)}$ where
> x_i has the maximum density
> $$Pr(e_{it} = 1) \propto \frac{a + \sum \mathbf{e}_{-i}}{a + b + n - 1} F(x_i|\theta_{i(t-1)})$$
> Choose an outlier cluster $\theta'_{i(t-1)}$ where
> x_i has the maximum density
> $$Pr(e_{it} = 0) \propto \frac{b + n - \sum \mathbf{e}_{-i} - 1}{a + b + n - 1} F(x_i|\theta'_{i(t-1)})$$

15:    Draw a sample for the cluster index with the probabilities as follows

$$Pr(z_{it} = k, k <= C) \propto \frac{n_{k,-i}}{n + \alpha - 1} F(x_i|\theta_{k(t-1)})$$

16:    **if** e_{it} is 1 **then**

$$Pr(z_{it} = k, k = C + 1) \propto \frac{\alpha}{n + \alpha - 1} \int F(x_i|\theta) G_0(\theta|\lambda') d\theta$$

17:    **else**
$$Pr(z_{it} = k, k = C + 1) \propto \frac{\alpha}{n + \alpha - 1} \int F(x_i|\theta) G_0(\theta|\lambda) d\theta$$

18:    **end if**
19:    **if** a new cluster is formed **then**

$$C = C + 1$$

20:        and mark the new cluster as an outlier cluster if e_{it} is 1 or 0 otherwise
21:    **end if**

22:    For all clusters marked as outlier clusters, sample the cluster parameters from the posterior distributions

$$Pr(\theta_{kt}) \propto G_0(\theta_{k(t-1)}|\lambda^{'})F(\mathbf{x}_{k(t-1)}|\theta_{k(t-1)})$$

23:    For all clusters marked as non-outlier clusters, sample the cluster parameters from the posterior distributions

$$Pr(\theta_{kt}) \propto G_0(\theta_{k(t-1)}|\lambda)F(\mathbf{x}_{k(t-1)}|\theta_{k(t-1)})$$

24: **end function**

Type	Minimum	Q1	Median	Mean	Q3	Maximum
1	4.0	24	125.5	193.3	329.0	579.0
2	4.0	19.2	38.5	43.6	64.7	98.0
3	16.0	28.0	32.0	48.9	38.0	160.0
4	17.0	28.0	32.0	36.5	40.0	76.0

Table 4.1: Statistical summary of the 4 types of synthetic data

1. dataset 1 - *Data with multiple clusters and large dispersion*: Data from a Poisson mixture of 5 components with rate parameters (10,30,130,320,520) and mixture weights of (0.2,0.2,0.25,0.15,0.2).

2. dataset 2 - *Data with multiple clusters and small dispersion*: Data from a Poisson mixture of 5 components with mixture weights (0.13,0.18,0.2,0.24,0.25) with rate parameters (10,22,35,58,80).

3. dataset 3 - *Data with two clusters and large dispersion*: dataset contains two distinct Poisson mixture components. The mixture weights are (0.75, 0.25) with the rate parameters (30,130).

4. dataset 4 - *Data with two clusters and small dispersion*: dataset has been generated from a mixture of two Poisson distributions with rate parameters(30,60) and mixture weights (0.75, 0.25).

The statistical summary of the datasets is given in Table 4.1. There are 250 data items for each type of dataset. We can see from the table that these data have varying dispersion and number of clusters among them. We have used such diverse data set for rigorous validation.

We have run our algorithm and obtained outlier probability for the datasets and outlier labels for each datum. We have used the uninformative priors for Beta distribution as ($a = 1, b = 1$). These values ensure that the algorithm has weak and uniform prior information regarding outlier probability. In case there is strong prior information about outlier probability, that information can be incorporated by appropriately altering the values of a and b. Since the number of clusters formed is sensitive to α [110], we have used a weak informative prior value of $\alpha = 0.01$. We experimented with various types of data using the presented version of Gibbs sampler algorithm. The sampler converges within the first few tens of iterations for datasets with large dispersion. In contrast, the sampler converged within first few hundred iterations for datasets with small dispersion. So, we have run the algorithm to generate 2000 samples of random variables for each dataset. We have used a burn-in period of 400 iterations and discarded those samples and obtained i.i.d samples at a lag of 10 iterations to obtain posterior predictive density of data and posterior outlier probability. The posterior predictive density is defined as

$$Pr(\widetilde{x}|\mathbf{x}, \alpha) = \sum_{\theta} Pr(\widetilde{x}|\theta) Pr(\theta|\mathbf{x}, \alpha) \tag{4.14}$$

Here, $\widetilde{x}$ is the data element for which density is to be predicted using the observed data $\mathbf{x}$, hyperparameters α and parameters θ. We use the samples of the parameters θ_k', θ_k and mixture weight samples to compute posterior predictive density of the data. The posterior outlier probability at any iteration t is calculated with

$$F(p) \sim Beta(a + n_t', b + n_t) \tag{4.15}$$

Here n_t' is the number of data items assigned to clusters marked as outlier clusters and n_t is the number of data items assigned to clusters marked as non-outlier clusters. At every iteration of the algorithm, we compute the posterior mean of the outlier probability which is the considered the outlier probability of the entire dataset for that iteration. The results of running Algorithm 2 are shown in Figures 4.2-4.4. Figure 4.2 shows the posterior predictive density that is derived for each of the datasets from the samples of Gibbs sampler. Figure 4.3 shows the number of clusters formed for each dataset over the samples. We show the posterior mean of outlier probability for each of the datasets in Figure 4.4. Here, the horizontal lines show the cumulative mixture weights of the mixture components sorted by their Poisson rate parameter.
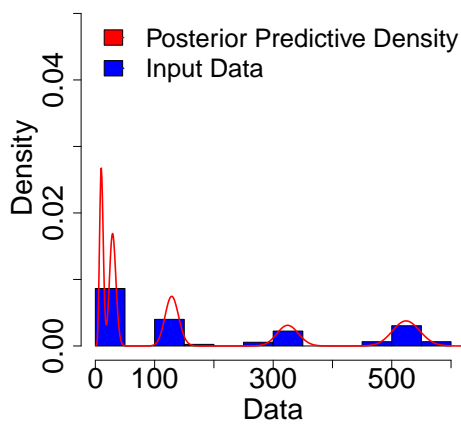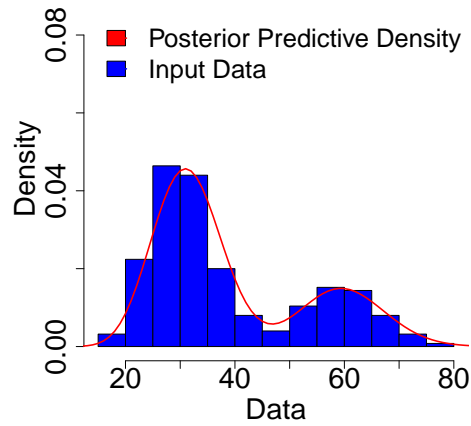
(a) Posterior predictive density for dataset 1
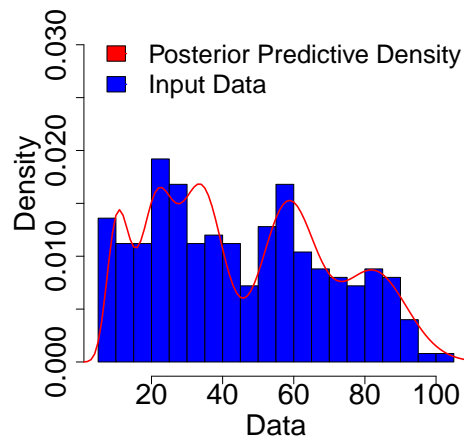
(b) Posterior predictive density for dataset 2
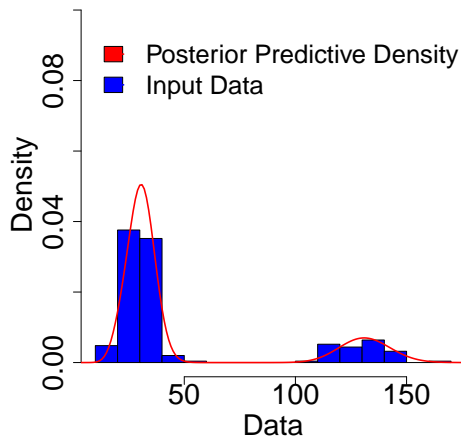
(c) Posterior predictive density for dataset 3

(d) Posterior predictive density for dataset 4

Figure 4.2: Posterior predictive density against data

The algorithm identifies appropriate number of clusters in the dataset within first few iterations as seen in Figure 4.3.

In order to compare the performance of our algorithm, we have used MMPP model discussed in [97] which identifies real world events based on outliers. This model considers that any count value in a periodic count dataset can be described as

$$N(t) = N_0(t) + N_E(t) \tag{4.16}$$

where $N_0(t)$ represents counts due to periodic activity. $N_E(t)$ represents any additional counts that are due to an event and hence represent an outlier.

The counts observed due to periodic activities are assumed to follow a Poisson distribution with rate parameter $\lambda(t)$ dependent on time. In order to account for the effect of day and hour on the periodic counts, the time dependent rate parameter is calculated as $\lambda(t) = \lambda_0 \delta_{d(t)} \phi_{d(t),h(t)}$. Here $\delta_{d(t)}$ controls the day effect and $\phi_{d(t),h(t)}$ controls the hour effect for a day.

The presence of an event results in increased counts in addition to periodic counts which is captured in $z(t)$ and $N_E(t)$ defined as

$$N_E(t) = \begin{cases} 0, & \text{if } z(t) = 0 \\ P(N, \gamma(t)), & \text{if } z(t) = 1 \end{cases} \tag{4.17}$$

Here $z(t)$ is controlled by Markovian transition probability matrix defined as

$$\begin{pmatrix} 1 - z_0 & z_1 \\ z_0 & 1 - z_1 \end{pmatrix} \tag{4.18}$$

and z_0 and z_1 have prior distributions and their values determine the number of events detected by the system as outliers with increased count values. The model also contains other appropriate prior distributions and hyperparameters needed for the Bayesian inference. We have obtained MCMC samples through Gibbs sampling of the random variables described above. The posterior mode of $z(t)$ in the samples for any count value determines whether that count value is an outlier and likely represents an event. The posterior mean computed with $z(t)$ for all the data can be used to detect outlier probability at each sample. We use this outlier probability to analyse the effectiveness of above described technique. In Table 4.2, we summarize the cumulative

mixture weights and outlier probabilities inferred by both algorithms to analyse their effectiveness. In order to analyse the performance of the algorithm in detecting outlier probabilities, consider dataset 3 which has two distinct clusters. Mixture weights for the clusters are (0.75, 0.25) with Poisson rate parameters (30, 130) respectively. Data items belonging to the second cluster represent outliers with high count values. In Figure 4.4c, it is shown that the algorithm appropriately identifies the mixture weight of the second cluster as proportion of outliers. In dataset 2, there is very low dispersion in spite of generating the data from 5 distinct clusters. We can see that the algorithm identifies the combined weights of the last three clusters as outlier probability. In Figure 4.4, the four plots show that the algorithm identifies sum of the mixture weights of high value rate parameter clusters, as outlier probability. This approach to finding outliers is more informative as outliers are categorized with different rate parameters and the proportion of each category of outliers can be known. Since we assign each datum to a single cluster where it has the maximum density, all the data items in a given cluster are either marked as outliers or non-outliers. Also, any prior information regarding outlier probability for the dataset can be easily incorporated using suitable Beta prior distribution.

In Figure 4.5, we show the outlier probabilities derived for synthetic datasets. Here the horizontal lines show the cumulative mixture weights of mixture components sorted by their Poisson rate parameter. The third row of the table shows that both models identify mixture weight of the cluster with large Poisson parameter as outlier probability for dataset 3. We can see these inferences in Figures 4.4c and 4.5c. Note that dataset 3 has a huge dispersion with clear bimodality. Since MMPP based model specifically distinguishes between normal and outlier dataset with its parameters, it successfully infers about the proportion of outlier data. However, in Figures 4.5a, 4.5b and 4.5d, we can see that MMPP based algorithm fails to identify the outlier probabilities correctly for datasets 1, 2 and 4.

For each of these datasets, it labels some of the data belonging to a cluster as outliers and the remaining in the same cluster as non-outliers. For example, consider the posterior mean of the outlier probabilities identified by the algorithm for dataset 2 is 0.62. This implies that the algorithm identifies some of the data points belonging to cluster with rate parameter 35 as outliers and rest of the data of same cluster as non-outliers. Similarly, the algorithm infers that the outlier probability for dataset 4 is 0.38 which implies that 13% of the data belonging to non-outlier cluster with rate
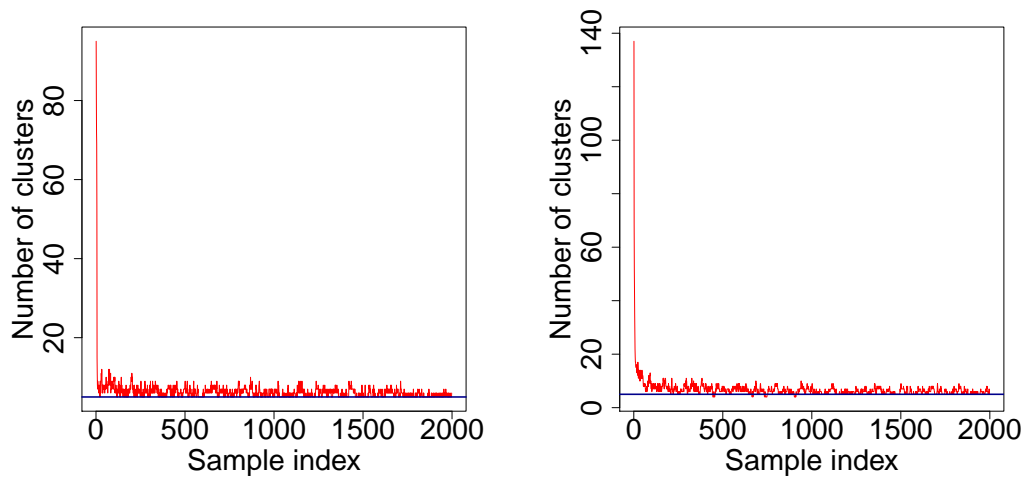
(a) Samples of number of clusters for dataset 1

(b) Samples of number of clusters for dataset 2
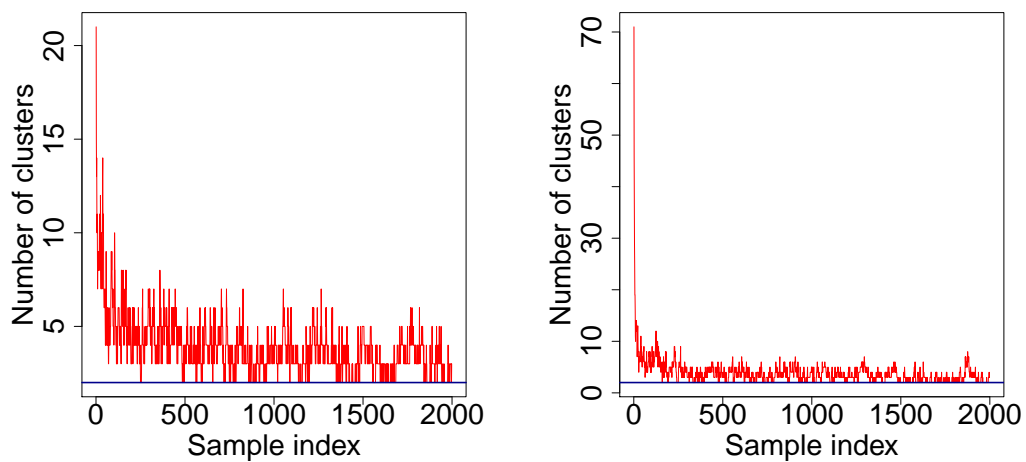
(c) Samples of number of clusters for dataset 3

(d) Samples of number of clusters for dataset 4

Figure 4.3: Samples of number of clusters obtained during MCMC sampling
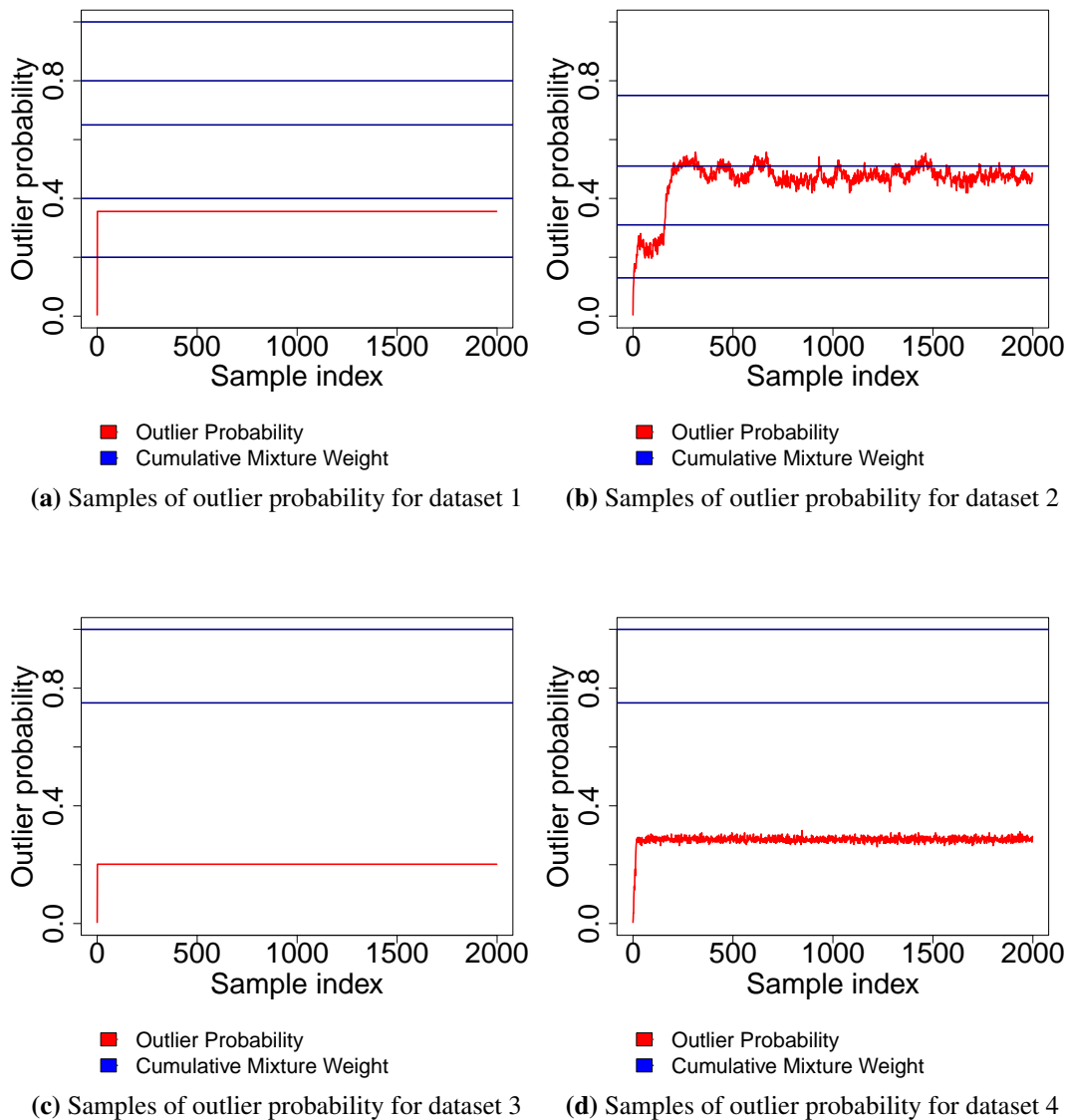
(a) Samples of outlier probability for dataset 1

(b) Samples of outlier probability for dataset 2

(c) Samples of outlier probability for dataset 3

(d) Samples of outlier probability for dataset 4

Figure 4.4: Samples of outlier probability obtained from the algorithm

Dataset	Cumulative Mixture Weights	Presented Model	MMPP model
1	(0.2,0.4,**0.65**,0.8,1)	**0.35**	0.57
2	(0.13,0.31,**0.51**,0.75,1)	**0.48**	0.62
3	(0.75,**1.0**)	**0.22**	**0.24**
4	(0.75,**1.0**)	**0.28**	0.38

Table 4.2: Comparison of outlier probabilities inferred by the models

parameter 30 are misclassified as outliers. These results clearly show that MMPP based outlier detection is not effective for datasets which have no strong bimodality.

On the other hand, the proposed model correctly identifies outlier probabilities based on mixture weights of different clusters in datasets as seen in Table 4.2. For example, in the first row, we can see that inferred outlier probability is 0.35 for the first dataset. This means that our model identifies the sum of mixture of weights of last two clusters as outlier probability. So, it infers that last two clusters are outlier clusters and data belonging to them are outliers. This also lets us make probabilistic queries about outliers belonging to outlier clusters. Similar observations can be made for the remaining datasets as shown in the table. We now explain the advantages of our technique in detecting real world events against MMPP based event detection technique.

4.2.2 Buildings Dataset

Now, we analyse the performance of our algorithm in detecting events using human generated periodic count data. We have used the *buildings* dataset mentioned in [97] which consists of count data of people's movements recorded every 30 minutes at the Calit2 institute building in the University of California, Irvine campus for a duration of 3 months. The data was recorded by optical detectors which count the number of people entering and exiting the building. Additionally, there are details of *events* that took place in the building. The time series data is effected by events which are aperiodic activities held in the building. There were 89 hours during which events have taken place in the building. Note that all the hours during which events have occurred might not have increased counts of people moving in the building. Few events in the building were unscheduled or unofficial and hence were not recorded. Some of these events can be directly seen as increased count values which are *outliers* in the time
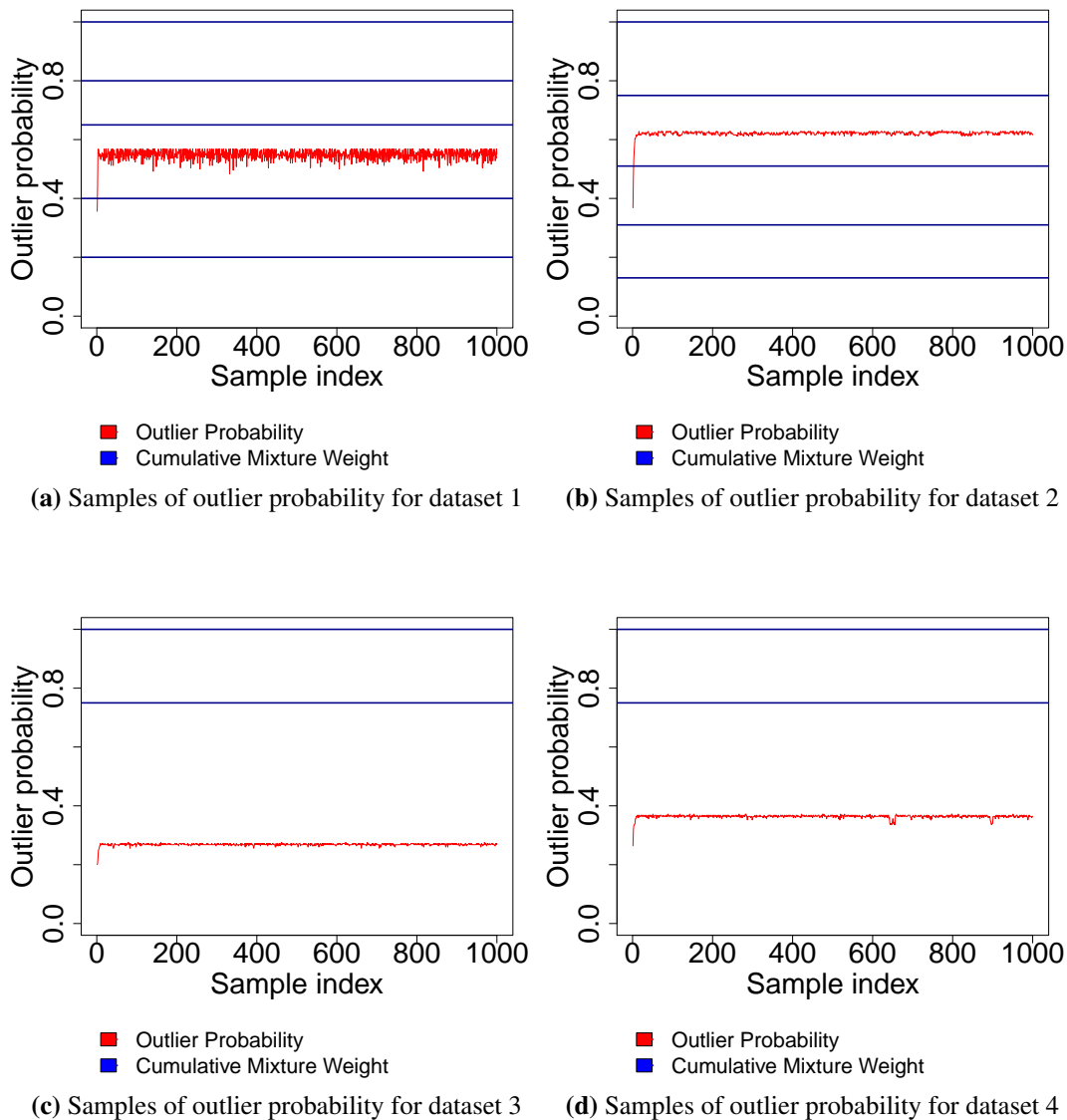
(a) Samples of outlier probability for dataset 1

(b) Samples of outlier probability for dataset 2

(c) Samples of outlier probability for dataset 3

(d) Samples of outlier probability for dataset 4

Figure 4.5: Samples of outlier probability obtained using MMPP model
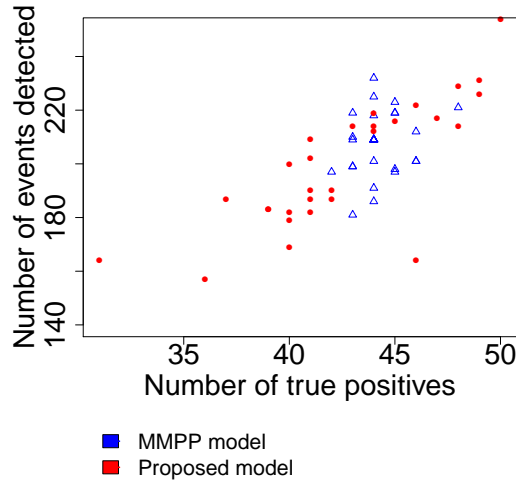
Figure 4.6: Performance evaluation for event detection

series data. We have used the time series data of hourly counts of people entering the building and real world events in the building to evaluate the two models. Detecting outliers in such data to infer about any real world events is challenging.

We have run our algorithm on count datasets corresponding to each hour of a day and detected outliers. We have used outliers to detect hours during which events were held in the building. We inferred Poisson rate parameters of mixture components for hourly dataset. So, the various *rates* at which people arrive at any given hour and unusual rates among them are known. For example, five different rates at which people arrive in the building at any hour shows the variability in the arrival rates. If two rate parameters represent outlier clusters, then count values belonging to these two clusters potentially represent events in the building. In order to analyse the performance of the MMPP based event detection technique, we set up the transition probability matrix over $z(t)$ with 30 sets of random weights to detect the same number of event hours as detected by our model. The performance of both the techniques are shown in Figure 4.6. Our technique had a true positive rate of 28% whereas the average true positive rate for the MMPP model was 21%. Another advantage of our technique is that the user need not have any prior knowledge about what constitutes an event. In contrast, such information is required in order to detect events using MMPP model as different weights on the transition probability matrix over $z(t)$ detect different number of events.

4.2.3 Check-in Counts Dataset

Foursquare is a famous location-based social network (LBSN) which lets its users broadcast their presence at places to friends or public by *checking-in* at places. LBSN check-ins data for a place can be summarized as count data which is the total number of check-ins within any hour for any given day. Such a total count is an approximate measure of how many people have visited a place at the given hour. By aggregating such hourly count data, we can get time series data indexed by days for any place. LBSNs also let the users to report any events that happen at places. Usually the check-ins count is high during the day till midnight and low for the rest of the hours [32, 75]. Also, the check-ins count is affected by the type of the place and would be low for a place like hospital compared to a busy airport as observed by Ye et al. [8]. Noulas et al. [111] show that check-in counts also vary based on the where the place is situated geographically. It is known that such total number of check-ins or visits of people vary based on the real life events at the corresponding place and can cause unusual bursts in the count data [2, 112]. The above observations reveal that count data based on user check-ins at places varies based on the time, events and neighbourhoods of places. In this work, we have addressed the problem of modelling count data to detect outliers and infer events. Now, we analyse the nature of such data and show the necessity of applying the proposed model.

Foursquare supports developers to download place specific information through open APIs. One such API[2] gives the count of Foursquare users present at a place at any time based on the check-ins by the users. We crawled this count data for each hour of the day for 100,000 random places in USA and UK over a period of 3 months. We have discarded all the places belonging to the *Home* category and also the places with no check-ins. The statistical summary for the mean of check-in counts is (Min=0.015, $Q1$=0.071, $Mean$=0.399, $Q3$=0.368, Max=149.5). We computed variance to mean ratio for each of the places for all hours of the day and obtained the summary as (Min=0, $Q1$=0.939, $Mean$=1.248, $Q3$=1.183, Max= 617.6). This clearly shows that check-in counts at places have varying degrees of Poisson over dispersion and under dispersion. We can note this dispersion by observing the multimodality of the data in the Figure 4.8. Thus, it showed that check-in counts at a place cannot be accurately modelled using a single Poisson random variable or a specific finite mixture of Poisson random variables and demands a flexible approach to outlier detection. We have considered user check-in datasets from two populous airports in USA namely San
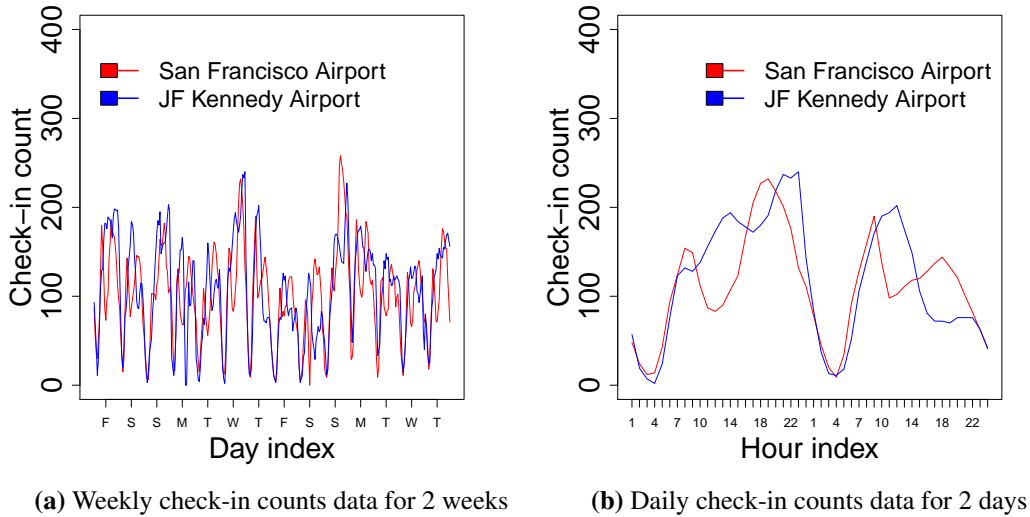
---

[2]https://developer.foursquare.com/docs/venues/herenow

(a) Weekly check-in counts data for 2 weeks

(b) Daily check-in counts data for 2 days

Figure 4.7: Periodic human activities at airports expressed via mobile check-ins

Francisco International Airport (SFO), John F. Kennedy International Airport (JFK) for analysing the performance of our algorithm on real world time series data. We crawled the time series data of check-in counts for these two airports for a period of 6 weeks. So we got time series of count data for each hour of the day which meant that the length of time series data is 42 for each hour for both the airports. We show the weekly and daily periodic check-in activities that mobile device users perform in Figures 4.7a and 4.7b.

We can see the periodic pattern of human activities expressed in the form of check-ins and the effect of hour of the day on check-in counts. Since, there are flights throughout the week, there is no effect of the day of the week in the daily check-in patterns. Detecting outliers in this time series data with parametric approach to identifying mixture components can be limiting. We have run the Gibbs sampler algorithm on the hourly time series for both the airports for each hour of the day.

The statistical summary of dispersion index for the 24 time series datasets is showing that there is presence of varying number of mixture components in the data. We found that there are varying numbers of mixture components for various hours of the day. Solutions to detecting outliers based on parametric mixture models such as MMPP would have limited performance in analysing these kinds of multimodal time series data. In order to analyse the performance of our algorithm, we will discuss the results for 4th and 9th hour time series data for SFO airport. In Figures 4.8a, 4.8b,
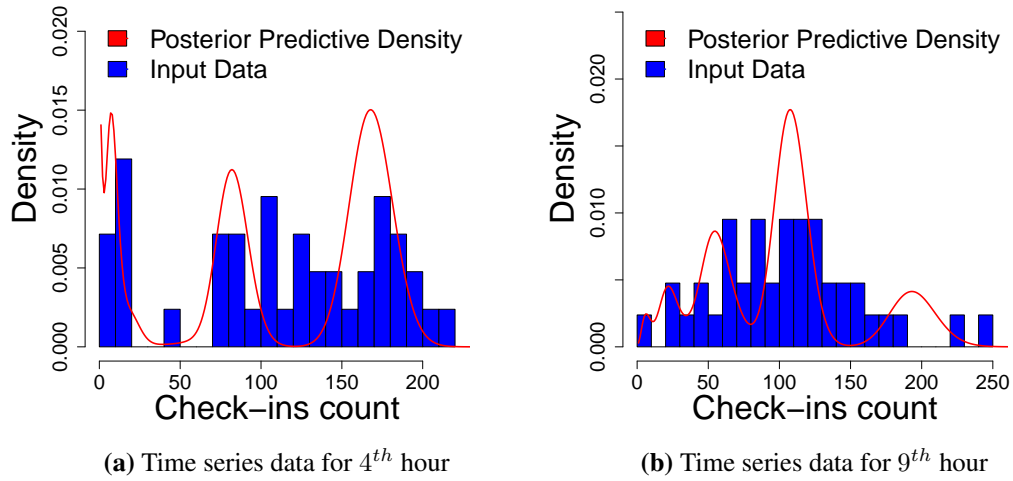
(a) Time series data for 4th hour

(b) Time series data for 9th hour

Figure 4.8: Posterior predictive density for time series data

(a) Samples of number of clusters for 4th hour time series data

(b) Samples of number of clusters for 9th hour time series data
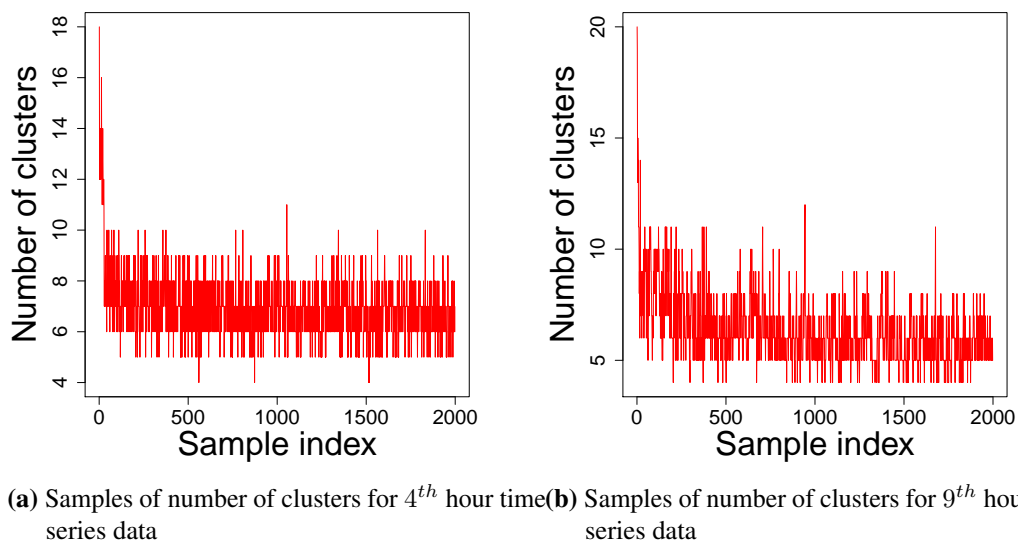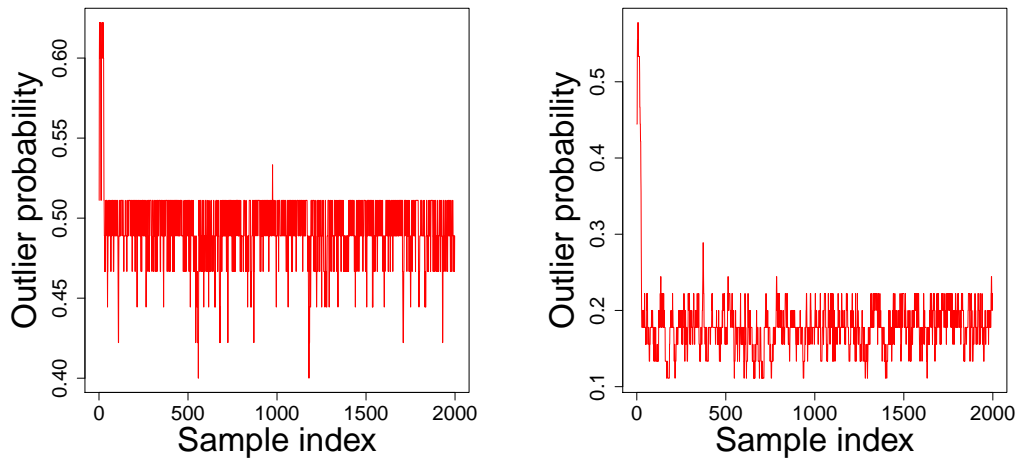
Figure 4.9: Samples of number of clusters obtained for the Airports time series data

(a) Outlier probability samples for 4th hour time series data

(b) Outlier probability samples for 9th hour time series data

Figure 4.10: Samples of outlier probability obtained for time series data of Airports

we show the check-in counts data and posterior predictive density derived from the algorithm. In Figures 4.9a, 4.9b, we show the number of clusters found for these data and the outlier probabilities are shown in 4.10a, 4.10b. We can see from Figure 4.9 that there are varying numbers of mixture components.

4.3 Discussion

In this chapter, we have presented a nonparametric Bayesian model to detect outliers in the periodic count data sets. We have used this model to describe the generative process behind an unknown Poisson mixture of probability densities and developed an appropriate version of the Gibbs sampling algorithm for probabilistic inference. This algorithm identifies the outliers based on the mixture components revealed by probability density of the data. We have first analysed the performance of the algorithm on a synthetic dataset varying in the number of mixture components and data dispersion and compared it with the performance of MMPP based outlier detection technique. We have shown that the algorithm effectively identifies outliers and outlier probabilities for the data sets even in cases where MMPP based technique has limited performance. Then, we have used the people's movement and events data for a university building and shown that our technique is more effective than MMPP based technique in detecting events based on outliers in the periodic count data. Finally, we have analysed the

hourly time series of count data for a large number of venues on Foursquare, a location-based social network. Significant Poisson over dispersion and under dispersion of this data has shown that parametric mixture models to identify outliers can be of limited performance and other solutions are required for the problem. We have analysed the hourly time series of user presence for two major airports and found that such data has varying number of mixture components. We have demonstrated the performance of our algorithm by deriving outliers and outlier probability for these data sets.

Our experiments on synthetic and real world data reveal that events organized at places can be effectively detected by analysing time series data of amount of crowd at a place. So, event metadata for places can be generated in spite of users not actively annotating places with event information. This further motivates us to research on ways to utilize any event related data at places to derive more metadata for places. In the next chapter, we analyse textual data generated regarding events organized at places and show that descriptive semantic tags can be derived for places.

Chapter 5

Semantic Tagging of Places with Real World Event Data

In the previous chapter, we defined and implemented a probabilistic model for detecting events at locations by analysing the time series count data about the number of people checking-in at a place. Interestingly, it has been noted that users generate large amounts of textual data among others while participating in events conducted at places. In this chapter, we analyse textual data generated during events and show that such data can be useful in deriving tags for places. Specifically, in order to tackle the issue of lack of manually annotated place tags and automatic generation descriptive tags for places, we propose a solution that utilizes data about a set of events that happen in a specific place and use it to extract meaningful descriptive tags for that place. We use data about events held in places from Meetup, a famous event-based social network, and we apply Latent Dirichlet Allocation (LDA) to derive sets of probable descriptive tags for each place. In order to evaluate our approach, we measure semantic relatedness between our derived tags and the manually assigned tags for the same places that are present in Foursquare, a location-based social network. Results on semantic relatedness score show that event data can be used to derive semantically relevant tags for places, and also provide an indication of where is the cut-off point (threshold) that identifies the best tags among the ones that are found using LDA. In the next few paragraphs, we motivate the need for tagging places and how event related textual data can be useful in achieving this task.

As evident in the discussions in Chapter 3, major geospatial databases that contain data about places suffer from the lack of descriptive tags for places. Two major causes for this problem are that writing them is a time-consuming process and only a few

users do it despite having knowledge about places. In order to tackle this issue and automatically generate descriptive tags for places, we propose a solution that utilizes data about a set of events that happen in a specific place and use it to extract meaningful descriptive tags for that place. We use data about events held at places on Meetup, a well-known event based social network and apply Latent Dirichlet Allocation (LDA) to derive sets of probable descriptive tags for any place. In order to evaluate our approach, we measure semantic relatedness between tags derived for places on Meetup and manually assigned tags from Foursquare, a location based service. Results show that event data can be used to derive semantically relevant place tags. This shows that location based services can benefit from capturing data about events to derive place tags.

Descriptive tags or short text snippets about places are crucial in providing fine grained and accurate place search and recommendation results. By descriptive tags, we mean any short keywords which are semantically related to a place. The richer the place tag sets, the better the search and recommendation results. These tags can be used to support keyword search queries about places. For example, a set of descriptive tags such as *computer science*, *software development*, *students*, *higher education* etc. can be used for more effective retrieval of a computer science research center than when it is annotated with an abstract category tag such as *education*. Note that a place can have many topics associated with it, with each topic having its own set of tags. For example, a new software company building is associated with topics like *start up* and *software* with each topic having its own descriptive tags. The tags *entrepreneurship*, *incubator* describe the topic *start up* whereas *software development*, *mobile application* describe *software*. It is essential that any effort to annotate places, either manually or automatically, takes this observation into consideration. Please note that we interchangeably use the term place for Point of Interest (PoI) which is a physically fixed point location on the surface of earth.

The annotation of places with descriptive tags is crucial for effective retrieval of relevant places, but only those people who have sufficient knowledge about a place have the ability to annotate that place with appropriate tags. Hence, major location based service providers like Foursquare[1], Facebook Places[2], Google Maps[3], OpenstreetMap[4] and others heavily rely on information voluntarily provided by users to

---

[1] https://foursquare.com/
[2] https://developers.facebook.com/docs/reference/fql/place/
[3] https://www.google.ie/maps/preview
[4] www.openstreetmap.org/

populate their geospatial databases. The analysis of places on a famous location based service provider showed that most of the places lack manually annotated categorical tags and descriptive tags [8, 9, 26]. Though many people physically visit places or have knowledge about places, they do not systematically share such knowledge by annotating places with tags, a time consuming and knowledge intensive task. This has resulted in poor metadata on the geospatial databases provided by major players such as Foursquare, Google Maps and Openstreetmap. A study by Le et al. [113] found that only 10% of the places among the 1 million places on Eventfull[5], an event based social network for event discovery, had any tags. Similarly, Hegde et al. [26] found that only 7% of the places on Foursquare, a location based service, had descriptive tags among more than a million places.

Though users do not actively annotate places with metadata, they do generate data about places when they participate in various events conducted at those places. This data can potentially contain rich information about the places which can be used to derive place tags. The motivation for this hypothesis is that many of the events conducted at a place will share common topics or themes and are attended by a group of users some of whom share one or more common interests. So, textual descriptions about groups that have conducted events at a place along with user comments and interest profiles are potentially descriptive of that place. Recently, event based social networks such as Meetup[6], Plancast[7], Eventfull[8] have become popular [43] and capture vast amount of data about events held at places. Users with similar interests can join groups and groups organize events at places for the users on these networks. Groups have their own profiles where a group profile describes the purpose and motivation of the group. A user has a social profile possibly for each of the group she belongs to and describes her social interests with a group specific profile. Users can express their views about the events via comments, ratings and recommendations. These rich data generated about events conducted at places can potentially be used for deriving tags for the corresponding places. For example, data about events conducted at a *fitness club* by *women's sport*, *nutrition*, *fitness* groups attended by users with possible interest in *sports*, *diet*, *fitness* among other interests and their comments about events can help to find that *exercise*, *nutrition*, *outdoor* are candidate descriptive tags for that place.

The main questions we address in this chapter are as follows:

---

[5]http://eventful.com/

[6]http://www.meetup.com/

[7]http://plancast.com/

[8]http://eventful.com/

1. How can we use textual data about events conducted at a place to derive descriptive tags for that place?

2. What is the threshold score of relevance to find the best descriptive tags for a place, using a quality metric?

3. How does the quality of derived tags vary based on the category of a place?

We describe an approach that uses event data to derive descriptive tags for places. We have used textual data about events namely profiles of users attending an event, profile of social group organizing the event and user comments on Meetup, an event based social network. In order to derive place tags corresponding to various topics describing a place, we have applied Latent Dirichlet Allocation (LDA) model on these textual data to derive probable topics and probable words for those topics and used the words as place tags. We have simulated probable tags for places using the parameters inferred in LDA model and compared them against a ground truth of manual tags using a semantic relatedness measure. We have used these semantic relatedness scores to infer the threshold score above which a tag can be considered semantically relevant for a place. We have then studied the effect of topic ranking and word ranking on the relevance of derived tags by analysing semantic relatedness between derived tags and manual tags. The experimental results show that relevant tags can be automatically derived for places using our approach based on user text generated for events at those places. Currently, most of the location based services like Foursquare, Google Maps, Openstreetmap do not capture any event data about places in their geospatial databases. We show that these services can benefit from obtaining and analysing place related event data to derive place tags.

The rest of the chapter is organised as follows. In Section 5.1, we describe how the text generated during events at places can be utilized to derive most probable topics and tags for places. We discuss the experimental results and nature of derived tags in Section 5.2. In Section 5.3 we conclude the chapter and discuss our future work. Now, we discuss some of the state of the art techniques in location metadata generation.

The users on the Web are actively producing large quantities of volunteered geographic information (VGI). In spite of that, most places or Points of Interests (POIs) lack rich metadata in the form of tags [8, 9, 26, 113]. There have been various solutions proposed to overcome the problem of lack of rich metadata in geospatial databases. Majority of them consider *wisdom of the crowd* at places or the collective user data associated with places in order to generate geospatial metadata. Rattenbury et al. [114]

use geotagged image data to categorize tags into place tags or event tags but does not derive tags for places. [82, 83] have studied various approaches to deriving and recommending tags to annotate images based on various types of user data. Haklay et al. [10] use GPS traces and other geographic data provided by volunteers to create an accurate geographic map of the world. A semi-automatic approach to integrating information provided by users into the digital geographic gazetteers has been discussed in [5] by Kessler et al. In [15], the authors, Pozdnoukhov et al. study identifying events at various geographic regions and derive topics for those geographic regions using a Twitter[9] data set.

There have been significant efforts in deriving tags for places rather than whole geographic region. In [8], Ye et al. focus on deriving categorical tags for place categories such as *restaurant* and *cinema* based on the *check-ins* or place visits of users. Hegde et al. [26] find that social interest profiles along with the place visiting behaviour of users can be used to derive descriptive tags for places visited by users. These two approaches require data about check-ins done by the users in order to derive tags. This can be quite restrictive when users are unwilling to express or share their check-in information completely. Biancalana et al. [115] use content from location-based service to derive relevant text snippets that can potentially be used as tags points of interest. In [9], Mansour et al. model any place based on the way the place is mentioned in tweets to generate keywords for places that closely align with query terms of place search queries by users.

In a different context, data about real world events expressed on online social networks has been effectively used for various information retrieval tasks. In [116], Yin et al. use event and location data of users to recommend locations and events based on the user preferences. Liu et al. [117] show how event related data can be used to recommend tags for online social groups and suggest venues for conducting events. Qiao et al. [118] use social relationships of users, ratings for events along with geographical characteristics of events to recommend events. Le et al. [113] have worked on deriving tags for geographic regions and locations using textual descriptions of events. But, they assume that concept hierarchies exist for event descriptions text and use temporal profiles of events to derive tags. Also, they do not consider the fact that multiple topics can be associated with a place and separate tags for those topics have to be derived. However, there has been no work on using unstructured text data

---

[9]https://twitter.com/

about events to derive tags corresponding to possibly multiple topics for a place, to the best of our knowledge.

5.1 Tagging of Places with EBSN Data

Event based social networks (EBSN) are a major category of online social networks subscribed by users. They enable a group of users with similar interests to conduct events and physically meet at places to participate in events. Any event conducted at a place and represented on an EBSN has extensive data such as profile of the social group conducting the event, interest profiles of users participating in the event, comments and ratings by users about the event, duration of the event etc. It is easy to realize that many of the events conducted at a place will share common topics or themes. For example, it is more likely that science related events are conducted at a *science museum* and art related events are conducted at a *theatre*. Data about all the events conducted at the science museum can potentially be used to derive tags such as *science*, *experiments*, *museum* etc. Since there can be many topics associated with a place, it is essential that most probable tags are derived for each of the topics. For example, a computer science research centre can have 2 prominent topics namely *Software* which is described by tags such as *Programming*, *Computer Code*, *Debugging* etc. and topic *Research* which is described by tags such as *PhD*, *Conference*, *Research Article* etc. Any approach to derive tags for places needs to consider tags for all relevant topics rather than tags for the most relevant topic of a place. In this section we describe how tags can be derived for places using text generated about the events at those places.

Overview of Data Analysis

We have used openly available event data on Meetup, a famous EBSN, for our experiments and data analysis. We approach the problem of deriving place tags by obtaining most probable words that can semantically represent a place using textual data about events at places. Since a place can have multiple topics describing it, we need to derive clusters of tags where each cluster represents a topic for a place. In Chapter 3, we used Wikipedia Link Vector Model (WLVM) to derive similarity scores between derived tags and cluster them using hierarchical clustering. We used this approach as we had retained and expanded only Wikipedia concepts present in social profiles of online social networks users. The motivation for using Wikipedia concepts for processing profiles on a specific online social network -Facebook- has been described in the same chapter. However, Meetup online platform allows users to provide only short text snippets to describe user interests and group interests. Such text snippets

can be clustered based on their co-occurrence using methods such as Latent Dirichlet Allocation (LDA) [51], without the need for any additional processing to derive cluster scores. So, we have derived probable tags corresponding to multiple topics using LDA [51]. This probabilistic model has been widely used for unsupervised clustering of text documents by deriving latent topic distributions and word distributions for topics.

We have analysed the relevance of derived place tags by comparing them with manual tags. Comparing pairs of textual data to measure the relatedness between them has been well studied in the field of computational linguistics. Solutions to obtain distance and similarity scores between textual data have been proposed in [94, 119, 120] and many others. However, semantic relatedness measure based on Explicit Semantic Relatedness (ESA) derived in Gabrilovich et al. [121] has become widely used due its simplicity and applicability to any type of text. So, we have used semantic relatedness scores based on ESA for our analysis. Note that we cannot use WLVM technique used in Chapter 3, as we do not further process event related data to exclusively obtain Wikipedia concepts. We have used a simulation algorithm to understand semantic relatedness scores exhibited by derived tags. We have run the algorithm to obtain the distribution of semantic relatedness scores between derived tags and manual tags which form ground truth. We found that only a minor proportion of derived tags from LDA have high degree of semantic relevance to manually annotated place tags. This motivated us to further analyse simulated semantic relatedness score to find out threshold score of relevance that can be used to obtain only relevant tags among derived tags for any place. We have used threshold score to analyse the effect of topic rank and word rank on relevance of a derived place tags. We have found that top 5 words from top 3 topics for a place achieve semantic relatedness scores higher than threshold score and can be used as place tags. We have then analysed the quality of derived tags for different categories of places. Further analysis of derived tags shows that derived tags have high relevance for places belonging to all categories except for *food*, *hotel* and *transport* categories which are generic place categories.

5.1.1 Datasets Description

We have used openly available event data on Meetup[10], a popular event based social network, for our analysis. We chose 1400 places randomly in USA, UK and crawled event data for these places over a period of 6 months. We retained event details such as name, latitude, longitude of place of event along with textual data of the profile of

---

[10]http://www.meetup.com/

social group conducting an event, user profiles of event participants and user comments. We represent each place with what we call a *place document* which is a document whose text is generated from all the events that are organized at a place. This document conceptually represents all relevant interaction and interests of people who have visited the place and participated in any event. In order to generate a place document for a place, we used aggregated textual data corresponding to group profile, user profile and comments for all the events that are conducted at that place on Meetup. We generated the place documents by concatenating textual data of all the events conducted at any place.

In order to analyse effectiveness of our technique, we used Foursquare tags as ground truth since they are manually annotated by users on Foursquare. Note that most places on Foursquare do not have descriptive tags. So, finding a corresponding place entry for a Meetup place was a challenge. We crawled places in 200 meters radius of places on Meetup to get the possible matching places on Foursquare, using latitude and longitude values of Meetup places. Further, we matched the names of Meetup places against the names of Foursquare places using Levenshtein distance proposed by Sankoff et al. [122] and substring match percentage to filter out irrelevant places. Then we manually found exact matches of from the relevant list of places. Finally, 356 place pairs were found of matching Meetup and Foursquare places, which showed that 25% of the Meetup places in our data set had a matching Foursquare place. The metadata crawled for these Foursquare places includes tags, place name, latitude, longitude, category among others. We have put all these data on a repository for use by interested readers[11].

5.1.2 Deriving Tags with Latent Dirichlet Allocation

In order to retrieve probable tags corresponding to multiple topics related to a place, we have applied LDA on a collection of place documents. The primary version of latent dirichlet allocation proposed by Blei et al. [51] is an unsupervised text clustering technique by inferring the latent topic and word distributions for a set of documents in a document collection. Please refer to 2.3.2 for detailed discussion about LDA. As a reminder, the variables involved in the LDA model are as follows.

1. N is the number of documents in a collection

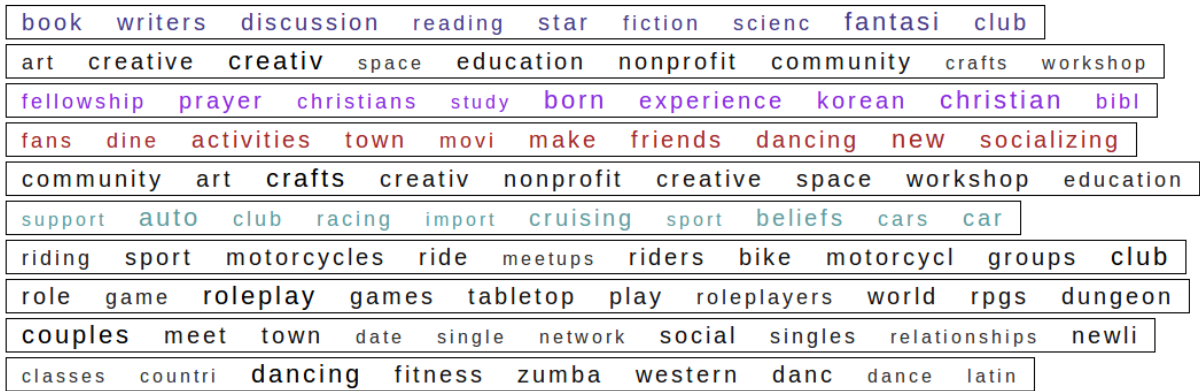2. K is the number of latent topics

---

[11]https://www.dropbox.com/sh/i6y8k95me1qsmdo/AAAJM68Qmwb-OAvRBubYUulza?dl=0

3. M is the number of words in a given document

4. w_{ij} is the jth word in i^{th} document

5. π_i is the topic distribution for i^{th} document

6. θ_k is the word distribution for k^{th} topic

Here, π_i indicates the topic distribution for i^{th} document. It is a discrete distribution over topic indices for the documents. θ_k are parameters of a discrete distribution with the sample space comprising of all the words in the document collection. The prior distribution for θ_k is defined by Dirichlet distribution with parameters λ. z_{ij} is the latent topic indicator for word w_{ij}. This model can be used to infer per document topic distributions π_i and per topic word distributions θ_k.

We have used Gibbs sampling mentioned by Griffiths et al. [58] on the collection of place documents to derive the latent topic and word distributions over those topics for an unknown number of topics over the unique words in the document collection. In order to compute the appropriate number of topics for the document collection, we have computed log-likelihood measure over $P(\mathbf{w}|K)$ where $\mathbf{w}$ is the set of all words in place documents and K is the number of topics. In Figure 5.2(a), we show the log-likelihood score against the number of topics. Based on the log-likelihood scores obtained, we have considered 100 topics for modelling the collection of place documents with LDA model. We show the top 10 words of 20 random topics in Figure 5.5.

Deriving most probable topics and most probable words for those topics by applying LDA on place document collection provides us the most probable topics and tags for places. It is possible that all topics derived might not be relevant for a place. Similarly, among some of the relevant topics for a place, all words might be not equally be descriptive of topics. So, the challenge here is to infer the statistical nature of the topics and tags that can be used as relevant place tags. So, we have used the ground truth of manually annotated tags on Foursquare to analyse the relevance of derived topics and tags. We analyse the nature of the derived tags by computing the semantic relatedness scores between manually assigned tags using Explicit Semantic Relatedness (ESA) measure proposed by Gabrilovich et al. [121]. Semantic relatedness between two text snippets is a measure of how strongly the text snippets are related to each other and this notion is more general notion than semantic similarity measures. ESA is a technique for the vector representation of a word using the text documents present in

book   writers   discussion   reading   star   fiction   scienc   fantasi   club

art   creative   creativ   space   education   nonprofit   community   crafts   workshop

fellowship   prayer   christians   study   born   experience   korean   christian   bibl

fans   dine   activities   town   movi   make   friends   dancing   new   socializing

community   art   crafts   creativ   nonprofit   creative   space   workshop   education

support   auto   club   racing   import   cruising   sport   beliefs   cars   car

riding   sport   motorcycles   ride   meetups   riders   bike   motorcycl   groups   club

role   game   roleplay   games   tabletop   play   roleplayers   world   rpgs   dungeon

couples   meet   town   date   single   network   social   singles   relationships   newli

classes   countri   dancing   fitness   zumba   western   danc   dance   latin

(a) Topic Set 1

poker   can   ride   hike   want   ski   skiing   outdoor   texa   will

lgbt   professionals   group   couples   lesbian   friends   gay   social

self   deutsch   best   empower   medicine   laughter   german   fun   exploration

indi   film   espanol   movi   scrabbl   spanish   movies   fans   films

sculpting   carving   massag   nlp   bead   sculpt   european   chi   therapy

light   consciousness   healers   healing   heal   crystal   worker   reiki   energi

coe   fitness   times   social   frisbee   frisbe   ultim   make   fun   pickup

runners   run   marathon   runs   distanc   train   new   trail   workout   running

dancing   danc   bachata   salsa   lessons   merengue   dance   ballroom   latin

group   cardio   peac   kickboxing   inner   extrem   fitness   exercise   workout   peace

(b) Topic Set 2

Figure 5.1: Top 10 words for 20 random topic indices

a knowledge base such as Wikipedia[12]. The results from our experiments show that top 3 topics and top 5 words of those topics are highly semantically related to the manually assigned tags on average. Now we describe how the topic distributions and word distributions inferred from LDA for place documents can be used to obtain only those tags that are relevant for any given place.

In order to analyse the relevance of derived tags, we have analysed semantic relationship between probable tags derived from LDA and manual tags. In Algorithm 3, we have simulated semantic relatedness scores for derived place tags against corresponding manual tags. We use this algorithm to obtain the distribution of semantic relatedness scores between derived tags and manual tags which is further used to identify semantically relevant tags from the set of derived tags. In this algorithm, we draw topics and words with probabilities proportional to the inferred distributions for each of the place documents. Here P is the set of all places where at least one event

---

[12]http://www.wikipedia.org/

Algorithm 3 Simulation of Semantic Relatedness Scores


1: **function** GETAVERAGESEMANTICRELATEDNESS(P)
2:     **for all** $p \in P$ **do**
3:         **for all** $m \in M_p$ **do**
4:             $relVector_p \leftarrow SimulateRelatedness(m, \theta_p)$
5:         **end for**
6:     **end for**
7: **end function**
8: **function** SIMULATERELATEDNESS(m, θ_p)
9:     **while** $iter \leq 10000$ **do**
10:         Draw topic t as $t|\theta_p \sim$ Discrete(θ_p)
11:         Draw word w as $w|\phi_t \sim$ Discrete(ϕ_t)
12:         $relVector_p \leftarrow$ SemanticRelatedness(w, m)
13:         $iter \leftarrow iter + 1$
14:     **end while**
15:     **return** $relVector_p$
16: **end function**
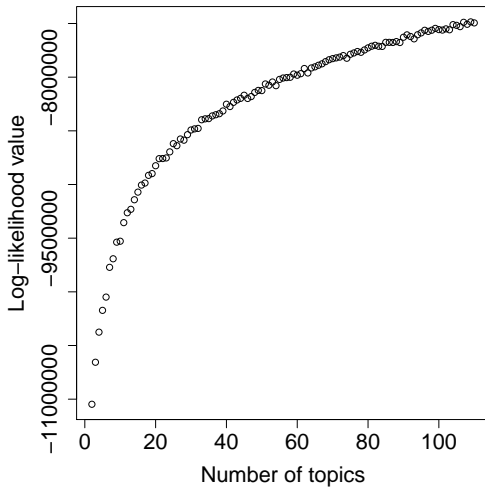
(a) Log likelihood against number of topics
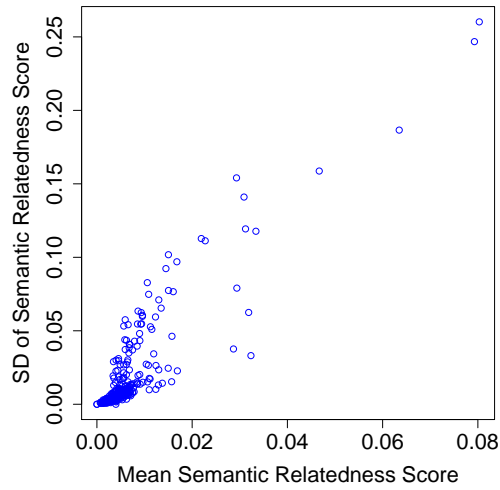
(b) Heteroscedasticity in semantic relatedness scores

Figure 5.2: Performance of LDA to derive semantically related tags

has been conducted. M_p denotes the set of all manually assigned tags for place p. θ_p is the topic distribution inferred from LDA for place p. ϕ_t is the word distribution for topic index t. $relVector_p$ is a vector storing the semantic relatedness scores computed between all the manually assigned tags and simulated words for place p. We have computed semantic relatedness scores using the explicit semantic relatedness technique

(a) Distribution of semantic relatedness scores

(b) Semantic relatedness scores at various percentile values
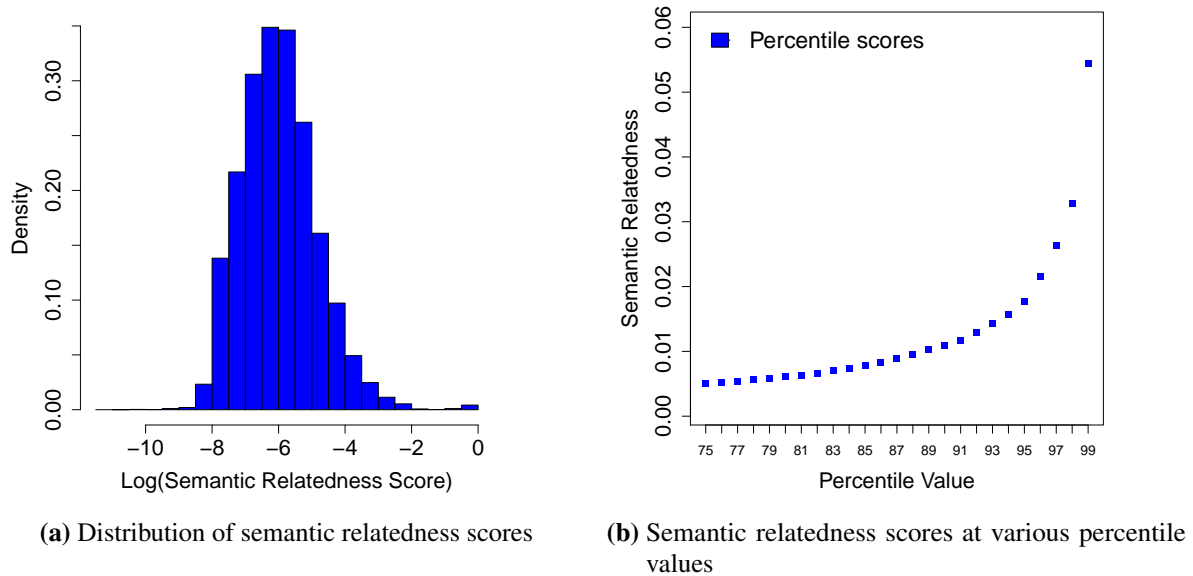
Figure 5.3: Statistics of semantic relatedness scores

implemented by Freitas et al. [123]. This Web service gives a semantic relatedness score in the range (0, 1). This ensures that we draw the words relevant to the place document more often than drawing 'noisy' words in the place document.

Deriving Threshold Semantic Relatedness Score

We have considered 100 random places present on both Meetup and Foursquare and simulated the semantic relatedness scores using Algorithm 3. We have retained only non-zero semantic relatedness scores as the majority of manual tag and derived tag pairs have semantic relatedness score of zero. We show the mean and standard deviation of semantic relatedness scores computed for each place in Figure 5.2(b). We can see that there is heteroscedasticity in the semantic relatedness scores. This means that even for places with high mean semantic relatedness scores, there is a need to filter out irrelevant tags which are contributing to the higher standard deviation. So, we have further analysed the simulated semantic scores between derived tags and manual tags to determine threshold semantic relatedness score of relevance. This threshold score is used to analyse the effect of topic and word rank on the relevance of derived tags. We will use this threshold score in Section 5.2 to analyse the effectiveness of our tag generation approach. In Figure 5.3(a), we show the distribution of semantic relatedness scores obtained for all place documents. It can be noted that only a minor portion of scores at the tail end of the distribution are very high and there is high

Component Index	Mixing Probability	Mean	Variance
1	0.1296	5.33e-4	2.41e-08
2	0.1601	1.11e-3	7.89e-08
3	0.2565	1.93e-3	3.91e-07
4	0.2449	3.80e-3	1.94e-06
5	0.1517	8.76e-3	1.18e-05
6	0.0485	2.39e-2	1.01e-04
7	0.0084	3.42e-1	1.25e-01

Table 5.1: Parameters of the Gaussian mixture distribution of semantic relatedness scores

variance in scores. These show the need to derive a threshold score to infer topic ranks and word ranks of relevant tags inferred from LDA. In Figure 5.3(b), we have plotted the semantic relatedness scores corresponding to various percentile values. We can see that there is a wide difference in the percentile scores at higher percentile values which indicates clear multimodality in the scores with outliers. These outlier scores can be used to determine the threshold score for semantic relevance of tags. Such multimodal data can be represented as a finite mixture of probability distributions and various parameters of mixture components can be inferred. Specifically, a multimodal data with K modes can be represented using the following probability density function.

$$f(x|\mu, \pi) = \sum_{k=1}^{K} \pi_k \phi(x|\mu_k, \sigma_k^2) \qquad (5.1)$$

$$\sum_{k=1}^{K} \pi_k = 1 \qquad (5.2)$$

Here we have used the Gaussian distribution as the generative distribution since we have continuous data and need flexibility in variances observed among sub populations. ϕ is the Gaussian probability density function with mean μ_k, variance σ_k^2 for k^{th} component. π_k is the mixing probability of kth component which is the proportion of the subpopulation corresponding to k^{th} component. Each component represents the sub population in the data and the corresponding mode. Detecting the outliers in multimodal data which is represented with a mixture model has been widely studied in the field of statistics. Yamanishi et al. [100] discuss the advantages of detecting outliers with finite mixture model. Specifically, the authors represent the data from a

finite mixture of Gaussian distributions as follows and detect the outliers in the data using simulation.

$$f(x) = (1 - \epsilon)f_n(x) + \epsilon f_0(x) \tag{5.3}$$

$$f_n(x) = \sum_{i=1}^{K-M} \pi_i \phi(x|\mu_i, \sigma_i^2) \tag{5.4}$$

$$f_0(x) = \sum_{j=K-M+1}^{K} \pi_j \phi(x|\mu_j, \sigma_j^2) \tag{5.5}$$

Here, ϵ is a small positive fraction which represents the proportion of data that are outliers. $f_n(x)$ is the probability density of the data that is non-outliers and $f_0(x)$ is the probability density of the outlier data. We need to infer ϵ for semantic relatedness scores which we can use to infer the threshold semantic relatedness score of relevance. We can consider one or more of the inferred components as outlier components and the data in those components as outlier data. In Table 5.1, we can see that component 7 and 6 are not separated due to the large variance of component 7. It is also evident that component 6 and 5 are clearly separated due to extremely small variance of component 6. Hence, we can compute that $\epsilon = 5.69\text{e-}2$ which is the sum of the mixing probabilities of component 6 and 7. This is also justified by the steep increase in the percentile score values at 95 percentile as seen in Figure 5.3b. So, we have used 1.78e-2 which corresponds to (1- ϵ) or 94.31 percentile as threshold score for semantic relatedness. It means that we consider all the derived tags having a semantic relatedness score of 1.78e-2 against manual tags as relevant derived tags for places and the rest of the derived tags as irrelevant.

Effect of Topic Rank and Word Rank

There are two problems in directly using the semantic relatedness score between probable words of place document and manually assigned tags to obtain relevant tags.

1. Since a place document collection can contain thousands of unique words, it is not scalable to compute the semantic relatedness score between all the words and manually assigned place tags.

2. Many of the places might not have any manually annotated tags to be used as ground truth to retrieve relevant tags from the set of derived tags

Due to these problems, it is desirable to know the characteristics of highly semantically related tags that are derived from place documents. So, we have studied the effect
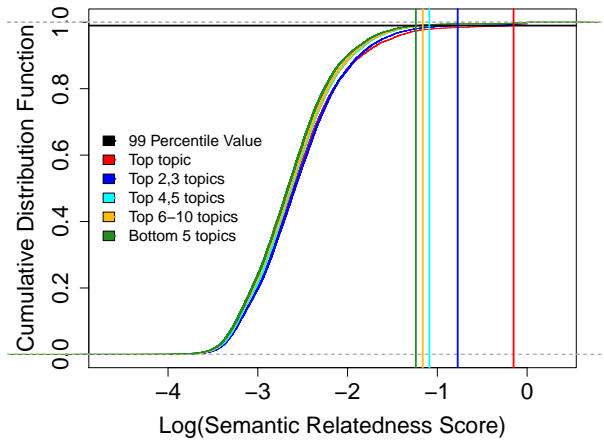
(a) Effect of topic rank on semantic relatedness
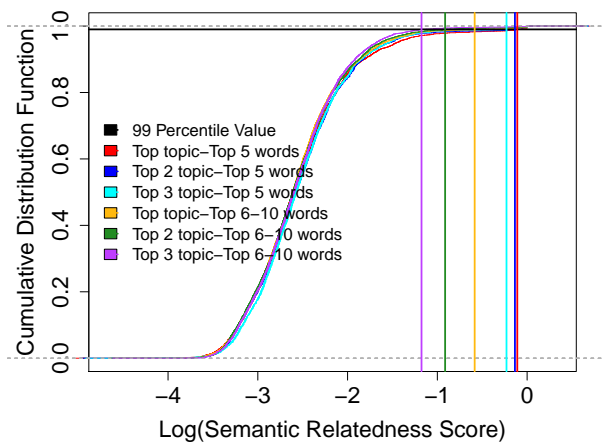
(b) Effect of word rank on semantic relatedness

Figure 5.4: Effect of top topics and top words on the semantic relatedness

of the ranks of topics and ranks of words for those topics on the relevance of derived place tags. We have plotted the Empirical Cumulative Distribution Function (ECDF) of semantic relatedness scores between top 10 words from top probable topics for place documents and the manually assigned tags for the corresponding places. Since all the ECDF curves in Figures 5.4(a), 5.4(b) are very similar and vary only at the higher percentile values, we have used the semantic relatedness score at the 99th percentile

to analyse ECDF curves corresponding to various topic and word ranks. So, we have plotted the semantic relatedness scores corresponding to the 99th percentile for the semantic relatedness scores obtained for various top probable topics. In Figure 5.4(a), we can see that ECDF curves have similar shapes except that there is huge difference in the 99th percentile score for various topics. It is evident that top 3 topics have much higher 99th percentile score than the rest of the topics. Also, 99th percentile scores for top 4-5 topics, top 6-10 topics and bottom 5 topics are almost same. So, word distributions of top 3 topics capture tags that are of high relevance to manual tags. In 5.4(b), we show the effect of word ranks among the top probable topics. It is evident that top 5 words of the top 3 topics clearly have high semantic relevance for the places. Even the top 6-10 words of the top topic are less relevant compared to the top 5 words of top 3 topics, on an average.

We now summarise the data analysis carried out and our findings. First, we applied LDA on a subset of place documents to derive place tags. These place documents were formed by aggregating event related data of places. We studied the quality of derived tags by simulating topics and words inferred from LDA against manual tags. We further analysed and noted that not all topics and words are relevant for a place to be used as tags. So, we studied the effect of topic and topic-word rank on the relevance of tags. We have found that top 5 words of top 3 topics for a place are highly semantically related to that place.

5.2 Experimental Evaluation

In the previous section, we used a subset of the data set and performed analysis using LDA and semantic relatedness scores to derive relevant tags representing various topics for places. In this section, we validate the statistical inferences about topic ranks, topic-word ranks on the relevance of derived tags for places in the entire data set. We analyse the relevance of derived tags by comparing average semantic relatedness score obtained by derived place tags against the semantic threshold score that we derived in Section 5.1.2. We have used the manual tags assigned by users to places on Foursquare corresponding to the places in the Meetup data set to validate our inferences.

There are 37124 unique words in the text corpus corresponding to Meetup events. In Figure 5.6a, we show the frequencies of these words. In Figure 5.6b, we show the frequency based ranks for top 3 words from the top topic for 300 random places. Here,

Meetup Name	Foursquare Name	Derived Tags	Foursquare Tags
Adobe SF Offices	Adobe	adobe, classes, development, digit, film, foods, photography, photography, programming, software, technology, web, workshops	adobe, adobe systems, flash platform, software development, great software, user groups
Pretty Lady Pole Fit	Pretty Lady Pole Fit	can, get, look, date, life, meet, night, relationships, single, singles, social, dance	aerobics, dance, pole dancing, zumba
Virginia Volleyball Center	Virginia Volleyball Center	self improvement, exercise, fun, hours, nutrition, social, volleyball, volleyball, women, indoor, networking, recreation, fitness, sport	courts, gym, pickup, sports, volleyball
Stuyvesant Yacht Club	Stuyvesant Yacht Club	sport, club, boating, sail	sailing, boat, yacht club, car, cars, fiction, organ, science, support, auto, discussion, book, gardening
Galaxy Games	Galaxy Games	board, card, friends, fun, game, games, make, meet, new, people, war	dnd, dungeons and dragons, gamers, games
Arastradero Preserve	Enid W. Pearson Arastradero Open Space Preserve	bicycling, dads, club, paddle, biking, run, running, cycling, train, mountain, road, stand, marathon	biking, running
Fit with Jenny - NEW LOCATION	Get Fit With Jenny	self improvement, basketball, cano, exercise, fun, group, kayaking, nutrition, training, weight, women, outdoor, fit, fitness	fitness, group fitness, kickboxing, personal training, women, zumba

Table 5.2: Manually annotated tags and derived tags for places
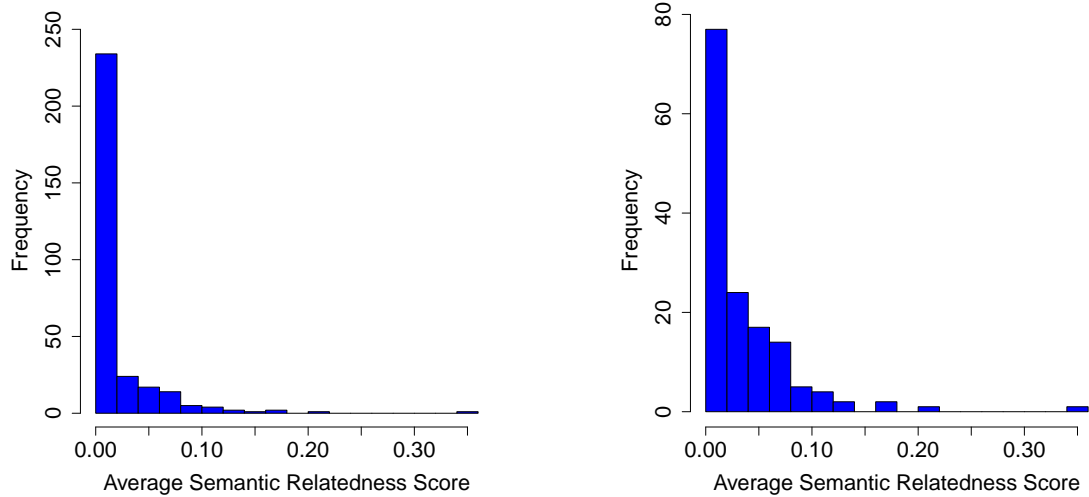
(a) Semantic relatedness for places of all categories

(b) Semantic relatedness for *other* category places

Figure 5.5: Semantic relatedness scores obtained for top 5 words of top 3 topics

Data	1st quartile	Mean	3rd quartile
Derived tags	9.1e-2	1.1e-1	1.2e-1
Foursquare tags	1.5e-1	3.8e-1	5.0e-1

Table 5.3: Semantic relatedness scores within tag sets

we can see that most of the top 3 words for places have bottom ranks and middle ranks when their frequency is considered. Only few of the top 3 words have higher frequency based ranks. This shows that

- Most frequently occurring words in the event text corpus cannot be directly used as most probable tags for places

- Words with least occurrence in the event text corpus can be highly relevant as tags for places

In Table 5.4, we can see that top 3 words for most of the places have very low rank when their frequency is considered in the text corpora. In Figure 6(b), we can see that few of the places have got same words (indicated by the flat portion of line segment) as
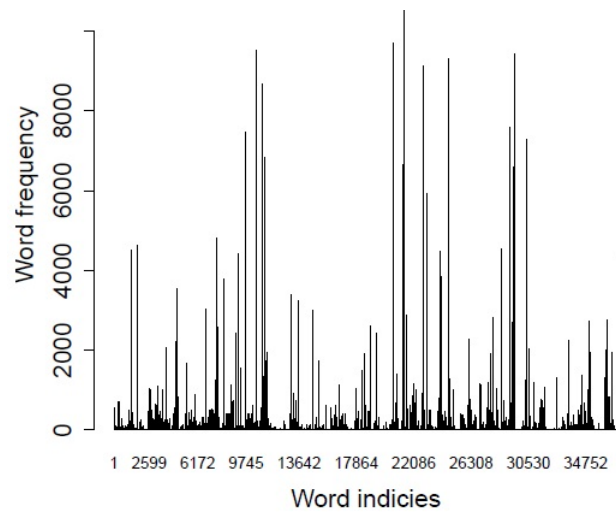
(a) Frequencies of words in text corpus
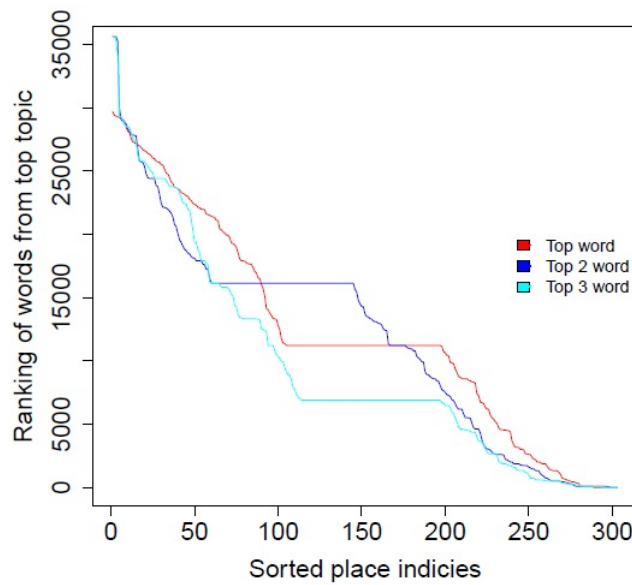
(b) Ranks of top 3 words of top topic for places

Figure 5.6: Frequencies and ranks of words in text corpus

one of their top 3 probable words. We observed that these words are very generic words namely 'can', 'get', 'look' and have frequency based ranks of 11242, 16413 and 6910 respectively. Further manual observation revealed that many of these places belong to one of categories - food, hotel and transport. This is in line with the conclusion that our technique is not effective in deriving tags for places belonging to these three categories. We now describe the effectiveness of our technique in deriving tags for places belonging to different categories.

Top Word Rank	Q1	Median	Q3
1	5628	11240	18320
2	2888	13740	16140
3	2659	6910	13580

Table 5.4: Statistical summary of frequency based ranks for top 3 words

We have derived the tags for the 356 place pairs in our data set using the tags and topics inferred using LDA on the place documents for places on Meetup. We used the top 5 words from the top 3 topics to obtain distinct tags for these places. We then manually observed the tags derived for many of the places. We noted the places which did not get many relevant tags were those for which the corresponding Foursquare place categories were *food*, *hotel* or *transport*. This is expected as events are conducted by different social groups with varying interests at such public places, event data generated is more about the interests of the groups than about *food*, *hotel* or *transport*. So, we categorized the places on Meetup to one of the four categories namely *food*, *hotel*, *transport*, *other* based on the category information of corresponding Foursquare places. We then analysed the semantic relatedness scores obtained by different categories of places. We can see from Figure 5.5(a) that average semantic relatedness score obtained by places using top 5 words of top 3 topics is low when we consider places of all categories. In Figure 5.5(b), we show the average semantic relatedness score obtained by places belonging to *other* categories which includes all the place categories except *food, hotel, transport*. We show distinct tags from the list of the top 5 words for top 3 topics in Table 5.2 for few places belonging to *other* category. We can see that most of the derived place tags are highly relevant to the places and are semantically comparable to manually annotated tags. Also, the derived tags correspond to various topics associated with a place. For example, consider the manual tags and derived tags for *Arastradero Preserve*. The tags *bicycling*, *biking*, *road*, *paddle*, *cycling* represent the topic *biking* whereas *run*, *marathon*, *mountain*, *run* correspond to the topic *running*. Similar observations can be made about the derived tags for other places shown. Derived tags for 67% of the places in the data set obtained a mean semantic relatedness score of 1.78e-2 or higher which is the threshold score of relevance as mentioned in Section 5.1.2. It shows these places obtained highly relevant semantic tags.

We then analysed the nature of the semantic relatedness exhibited by the manual tags and derived tags among themselves for each of the places. These pairwise semantic

Data	1st quartile	Mean	3rd quartile
Derived tags vs Foursquare tags (*Other*)	3.7e-3	2.5e-2	3.7e-2
Derived tags vs Foursquare tags	1.5e-3	4.7e-3	3.9e-3

Table 5.5: Semantic relatedness scores between tag sets

relatedness scores indicate the *cohesiveness* among the manual tags and derived tags. In Table 5.3, rows show that manual tags on Foursquare and derived tags from Meetup data have semantic relatedness scores of same order of magnitude and thus exhibit similar cohesiveness in terms of semantic relatedness. We also obtained the pairwise semantic relatedness scores obtained between the manual tags and derived tags. These scores indicate the *semantic relationship* between the manual tags and derived tags for any place. In Table 5.5, first row shows that the derived tags for places belonging to *other* categories have semantic relatedness scores higher than the threshold semantic score of 1.78e-2 on average. It is easy to see from second row that derived tags for places belonging to *food, hotel, transport* categories obtain very low semantic scores against the threshold semantic relatedness score and differ from it by an order of magnitude. We then analysed the number of tags manually annotated by users on Foursquare in order to analyse the number of extra tags we generated. Table 5.6 shows the statistics about the amount of that are manually annotated and the amount of automatically derived tags. We can see from Table 5.5, 5.6 that the derived tags are highly semantically related to manual tags and are also larger in number compared to manual tags.

5.3 Discussion

In this chapter, we have defined and implemented an approach to automatically generate semantically relevant tags for places using data generated on event based social networks. We have found that Latent Dirichlet Allocation (LDA), an unsupervised document clustering technique, can effectively be utilized to derive relevant place tags

Data	1st quartile	Mean	3rd quartile
Derived tags from Meetup	14	14.02	15
Manual tags on Foursquare	2	6.26	8
Manual tags on Foursquare (*other*)	2	7.13	9

Table 5.6: Number of manual tags and derived tags

while representing different topics that are relevant to a place. We have shown that ranks of top topics and top words in the topics inferred using LDA have an effect on the relevance of place tags. We have also found that relevant tags can be derived for places effectively by mining event related textual data. These tags are essential for automatic classification of places, and can therefore play a crucial role in providing accurate place recommendations and support more precise geospatial search queries. Our approach utilizes successful data mining techniques on unstructured textual data and does not require any manual supervision. It is cost-effective and provides very encouraging results as demonstrated by our evaluation. In its current form, the proposed approach does not use any rich temporal information about events. In future work, we plan to investigate the role of temporal properties of events such as periodicity, start time and duration. We believe this can help deriving place tags more effectively for places belonging to *food*, *hotel* and *transport* categories, which cannot be derived by our current approach. We will also investigate the effect of the amount of event data available for place events, and the accuracy of our algorithm in finding relevant place tags.

Chapter 6

Conclusion and Future Work

6.1 Summary

In this thesis, we have addressed the problem of automatic metadata generation for places or Points of Interests (PoIs). We have proposed and implemented various data mining and information retrieval techniques to address this problem. We have utilized user generated geolocation data captured by various types of online social forums, offline physical sensors to show the effectiveness of our solutions. We have developed three unique solutions to use various types of geolocation related data sets in order to generate metadata for places. Specifically, we have proposed effective solution to automatically generate textual snippets or tags descriptive of places based on user visits to locations and their social interests. We have further analysed data generated during participation of users in events organized on social networks to generate place tags. Furthermore, we have proposed a novel probabilistic algorithm to detect events at places that is more accurate than state of the art techniques and requires no manual tuning as it is an unsupervised learning algorithm. The above mentioned solutions can be used to automatically generate various types of metadata for places when they lack manually annotated metadata. Geospatial applications like location based services, geospatial web services can exploit these place metadata to provide accurate and relevant results against geospatial data queries. Place metadata can also be used for various applications like personalization of tour recommendations, place recommendations, geospatial aware advertisements etc. They can also be useful in financial planning activities such as setting up of stores in certain regions, decisions on logistics and supply chain among others.

6.2 Contributions

We have analysed three different types of geospatial data sets to propose various solutions to place metadata generation. Our solution to generate place tags by analysing social and check-in profiles of users shows that common interests of a group of people *checking-in* or visiting a place contains valuable information about that place. We have shown that common interests derived from the social interest profiles on online social networks can be used to derive rich set of highly descriptive place tags. We have defined a probabilistic model and a text processing algorithm for deriving relevant place tags. Furthermore, we have studied the effect of number of social profiles considered on the quality and stability of the tags derived. Since many users do not intend to share their check-ins or place visit information online, some of the data sources might not have enough check-ins information.

So we have studied the utilization of data generated during events organized at places. In this approach, we do not consider check-in information of users and utilize only textual data generated during events. We have used an unsupervised text clustering technique and probabilistic mixture modelling in order to infer about the nature of semantically relevant place tags. We have shown that the tags derived using our approach, are of the same quality as manual tags. Our method gives approximately three times more number of tags compared to the number of manual tags that people assign on a famous location based social network. Other prominent advantage of our approach is that topic specific tags are derived while we consider the relevance of multiple topics for a place.

It has been noted that users do not explicitly describe all the events that happen at a place. Knowledge about events happening at a place can be useful in recommending the place to appropriate Web users. Information about all the events that have happened at a place can also be used to analyse dynamics of population movements in a geographic region. So, we have developed an advanced event detection technique that uses the time series of count data of check-ins at places in order to detect events at places. This approach requires only the aggregated counts of user check-in information and does not require check-in information of individual users. Our approach is more accurate than the state of the art event detection technique and is completely unsupervised.

6.3 Future Work

We have exclusively used unstructured textual data from various types of online social networks to address the problem of generation of place tags. There are several approaches to text analysis that can be applied in order to achieve accurate tag generation using unstructured textual data. We have not considered the temporal aspect of users' social interest data such as time of creation, time of last update etc. We will investigate relevance of temporal profiles of social data of users on achieving better tag enrichment. Our current solution to analysing event related textual data does not consider the start time and duration of events. These data can potentially be used derive accurate tags for places specifically in the cases of places belonging to hotel, food, transport categories. We will also investigate utilizing other types of time series data generated for places in order to accurately detect events at places. We will further investigate scalable predictive models to infer amount of crowd at a location based on time series data of user presence obtained from sensors and user check-ins data obtained from location based social networks.

Bibliography

[1] S. Consolvo, I. E. Smith, T. Matthews, A. LaMarca, J. Tabert, and P. Powledge, Location disclosure to social relations: why, when, & what people want to share, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 81–90, ACM, 2005.

[2] J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman, I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application, in *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, pp. 2409–2418, ACM, 2011.

[3] A. J. Flanagin and M. J. Metzger, The credibility of volunteered geographic information, GeoJournal **72** (2008).

[4] J. Wang and J. Canny, End-user place annotation on mobile devices: a comparative study, in *CHI'06 Extended Abstracts on Human Factors in Computing Systems*, pp. 1493–1498, ACM, 2006.

[5] C. Keßler, K. Janowicz, and M. Bishr, An agenda for the next generation gazetteer: Geographic information contribution and retrieval, in *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 91–100, ACM, 2009.

[6] J. Lin, G. Xiang, J. I. Hong, and N. Sadeh, Modeling people's place naming preferences in location sharing, in *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, pp. 75–84, ACM, 2010.

[7] D. Lian and X. Xie, Learning location naming from user check-in histories, in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 112–121, ACM, 2011.

[8] M. Ye, D. Shou, W.-C. Lee, P. Yin, and K. Janowicz, On the semantic annotation of places in location-based social networks, in *Proceedings of the 17th ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 520–528, ACM, 2011.

[9] R. Mansour, N. Refaei, and V. Murdock, Augmenting business entities with salient terms from twitter., in *Proceedings of the 25th International Conference on Computational Linguistics*, pp. 121–129, 2014.

[10] M. Haklay and P. Weber, Openstreetmap: User-generated street maps, Pervasive Computing **7** (2008).

[11] I. Constandache, S. Gaonkar, M. Sayler, R. R. Choudhury, and L. Cox, Enloc: Energy-efficient localization for mobile phones, in *INFOCOM*, pp. 2716–2720, IEEE, 2009.

[12] D. Ashbrook and T. Starner, Using gps to learn significant locations and predict movement across multiple users, Personal and Ubiquitous Computing **7**, 275 (2003).

[13] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava, Using mobile phones to determine transportation modes, ACM Transactions on Sensor Networks **6**, 13 (2010).

[14] Z. Zhuang, K.-H. Kim, and J. P. Singh, Improving energy efficiency of location sensing on smartphones, in *Proceedings of the 8th International Conference on Mobile systems, Applications, and Services*, pp. 315–330, ACM, 2010.

[15] A. Pozdnoukhov and C. Kaiser, Space-time dynamics of topics in streaming text, in *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, pp. 1–8, ACM, 2011.

[16] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil, Exploiting semantic annotations for clustering geographic areas and users in location-based social networks, The Social Mobile Web **11**, 02 (2011).

[17] S. Y. Ho and S. H. Kwok, The attraction of personalized service for users in mobile commerce: an empirical study, ACM SIGecom Exchanges **3**, 10 (2002).

[18] H. H. Bauer, S. J. Barnes, T. Reichardt, and M. M. Neumann, Driving consumer acceptance of mobile marketing: A theoretical framework and empirical study, Journal of Electronic Commerce Research **6**, 181 (2005).

[19] K. Partridge and B. Price, Enhancing mobile recommender systems with activity

inference, in *User Modeling, Adaptation, and Personalization*, pp. 307–318, Springer, 2009.

[20] A. Miele, E. Quintarelli, and L. Tanca, A methodology for preference-based personalization of contextual data, in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pp. 287–298, ACM, 2009.

[21] A.-K. Pietiläinen, E. Oliver, J. LeBrun, G. Varghese, and C. Diot, Mobiclique: middleware for mobile social networking, in *Proceedings of the 2nd ACM Workshop on Online Social Networks*, pp. 49–54, ACM, 2009.

[22] C. Ratti, S. Williams, D. Frenchman, and R. Pulselli, Mobile landscapes: using location data from cell phones for urban analysis, Environment and Planning B: Planning and Design **33**, 727 (2006).

[23] J. J.-C. Ying, W.-C. Lee, T.-C. Weng, and V. S. Tseng, Semantic trajectory mining for location prediction, in *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 34–43, ACM, 2011.

[24] F. Calabrese, F. C. Pereira, G. Di Lorenzo, L. Liu, and C. Ratti, The geography of taste: analyzing cell-phone mobility and social events, in *Pervasive Computing*, pp. 22–37, Springer, 2010.

[25] P. Zhou, Y. Zheng, and M. Li, How long to wait?: predicting bus arrival time with mobile phone based participatory sensing, in *Proceedings of the 10th International Conference on Mobile systems, Applications, and Services*, pp. 379–392, ACM, 2012.

[26] V. Hegde, J. X. Parreira, and M. Hauswirth, Semantic tagging of places based on user interest profiles from online social networks, in *Proceedings of the 35th European Conference on Information Retrieval*, pp. 218–229, Springer, 2013.

[27] V. Hegde, M. Krnjajic, and A. Pozdnoukhov, Unsupervised event detection with infinite poisson mixture model, in *Proceedings of the IEEE International Congress on Big Data*, pp. 567–575, IEEE, 2015.

[28] V. Hegde, A. Mileo, and A. Pozdnoukhov, Events describe places: Tagging places with event based social network data, in *Proceedings of the 3rd IKDD Conference on Data Science*, p. 11, ACM, 2016.

[29] C. V. D. Weth, V. Hegde, and M. Hauswirth, Virtual location-based services: Merging the physical and virtual world, in *Proceedings of the IEEE International Conference on Web Services*, pp. 113–120, IEEE, 2014.

[30] M. Ye, P. Yin, W.-C. Lee, and D.-L. Lee, Exploiting geographical influence for collaborative point-of-interest recommendation, in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 325–334, ACM, 2011.

[31] A. X. Zhang, A. Noulas, S. Scellato, and C. Mascolo, Hoodsquare: Modeling and recommending neighborhoods in location-based social networks, in *Proceedings of the International Conference on Social Computing*, pp. 69–74, IEEE, 2013.

[32] E. Cho, S. A. Myers, and J. Leskovec, Friendship and mobility: user movement in location-based social networks, in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1082–1090, ACM, 2011.

[33] H. Gao, J. Tang, and H. Liu, gscorr: modeling geo-social correlations for new check-ins on location-based social networks, in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pp. 1582–1586, ACM, 2012.

[34] Y. Zheng, L. Zhang, X. Xie, and W.-Y. Ma, Mining interesting locations and travel sequences from gps trajectories, in *Proceedings of the 18th International Conference on World Wide Web*, pp. 791–800, ACM, 2009.

[35] A. Noulas, S. Scellato, N. Lathia, and C. Mascolo, Mining user mobility features for next place prediction in location-based services, in *Proceedings of the IEEE 12th International Conference on Data Mining*, pp. 1038–1043, IEEE, 2012.

[36] B. De Longueville, R. S. Smith, and G. Luraschi, Omg, from here, i can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires, in *Proceedings of the 2009 International Workshop on Location Based Social Networks*, pp. 73–80, ACM, 2009.

[37] L. Ferrari, A. Rosi, M. Mamei, and F. Zambonelli, Extracting urban patterns from location-based social networks, in *Proceedings of the 3rd ACM SIGSPA-TIAL International Workshop on Location-Based Social Networks*, pp. 9–16,

ACM, 2011.

[38] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo, Geospotting: Mining online location-based services for optimal retail store placement, in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 793–801, ACM, 2013.

[39] S. Bauer, A. Noulas, D. O. Séaghdha, S. Clark, and C. Mascolo, Talking places: Modelling and analysing linguistic content in foursquare, in *Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2012 International Confernece on Social Computing*, pp. 348–357, IEEE, 2012.

[40] J. J.-C. Ying, E. H.-C. Lu, W.-C. Lee, T.-C. Weng, and V. S. Tseng, Mining user similarity from semantic trajectories, in *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks*, pp. 19–26, ACM, 2010.

[41] S. Scellato, A. Noulas, and C. Mascolo, Exploiting place features in link prediction on location-based social networks, in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1046–1054, ACM, 2011.

[42] Y. Zheng, Y. Chen, X. Xie, and W.-Y. Ma, Geolife2. 0: a location-based social networking service, in *Proceedings of the 10th International Conference on Mobile Data Management: Systems, Services and Middleware*, pp. 357–358, IEEE, 2009.

[43] X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, and J. Han, Event-based social networks: linking the online and offline social worlds, in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1032–1040, ACM, 2012.

[44] K. Feng, G. Cong, S. S. Bhowmick, and S. Ma, In search of influential event organizers in online social networks, in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pp. 63–74, ACM, 2014.

[45] P. Yin, Q. He, X. Liu, and W.-C. Lee, *It Takes Two to Tango: Exploring Social Tie Development with Both Online and Offline Interactions* (SIAM, 2014),

chap. 38, pp. 334–342.

[46] K. Li, W. Lu, S. Bhagat, L. V. Lakshmanan, and C. Yu, On social event organization, in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1206–1215, ACM, 2014.

[47] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, An introduction to mcmc for machine learning, Machine learning **50**, 5 (2003).

[48] T. Bailey and C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, in *Proceedings of International Conference on Intelligent Systems for Molecular Biology* Vol. 2, p. 28, 1994.

[49] P. Deb and P. K. Trivedi, Demand for medical care by the elderly: a finite mixture approach, Journal of Applied Econometrics **12**, 313 (1997).

[50] M. Alfò, L. Nieddu, and D. Vicari, A finite mixture model for image segmentation, Statistics and Computing **18**, 137 (2008).

[51] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent dirichlet allocation, The Journal of Machine Learning Research **3**, 993 (2003).

[52] S. K. Lukins, N. A. Kraft, and L. H. Etzkorn, Bug localization using latent dirichlet allocation, Information and Software Technology **52**, 972 (2010).

[53] G. Maskeri, S. Sarkar, and K. Heafield, Mining business topics in source code using latent dirichlet allocation, in *Proceedings of the 1st India Software Engineering Conference*, pp. 113–120, ACM, 2008.

[54] D. Xing and M. Girolami, Employing latent dirichlet allocation for fraud detection in telecommunications, Pattern Recognition Letters **28**, 1727 (2007).

[55] I. Bhattacharya and L. Getoor, A latent dirichlet model for unsupervised entity resolution, in *Proceedings of the International Conference on Data Mining* Vol. 5, p. 59, SIAM, 2006.

[56] C. Wang, D. Blei, and F.-F. Li, Simultaneous image classification and annotation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1903–1910, IEEE, 2009.

[57] I. Titov and R. McDonald, Modeling online reviews with multi-grain topic models, in *Proceedings of the 17th International Conference on World Wide*

Web, pp. 111–120, ACM, 2008.

[58] T. L. Griffiths and M. Steyvers, Finding scientific topics, Proceedings of the National academy of Sciences of the United States of America **101**, 5228 (2004).

[59] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, Hierarchical dirichlet processes, Journal of the American Statistical Association **101** (2006).

[60] T. Chen, J. Morris, and E. Martin, Probability density estimation via an infinite gaussian mixture model: application to statistical process monitoring, Journal of the Royal Statistical Society: Series C (Applied Statistics) **55**, 699 (2006).

[61] M.-S. Shin, M. Sekora, and Y.-I. Byun, Detecting variability in massive astronomical time series data–i. application of an infinite gaussian mixture model, Monthly Notices of the Royal Astronomical Society **400**, 1897 (2009).

[62] J. P. Huelsenbeck and P. Andolfatto, Inference of population structure under a dirichlet process model, Genetics **175**, 1787 (2007).

[63] A. Kottas, J. A. Duan, and A. E. Gelfand, Modeling disease incidence data with spatial and spatio temporal dirichlet process mixtures, Biometrical Journal **50**, 29 (2008).

[64] K. A. Heller and Z. Ghahramani, Bayesian hierarchical clustering, in *Proceedings of the 22nd International Conference on Machine learning*, pp. 297–304, ACM, 2005.

[65] C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda, Learning systems of concepts with an infinite relational model, in *Proceedings of the National Conference on Artificial Intelligence* Vol. 21, p. 381, Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.

[66] L. Zhang, J. Meng, H. Liu, and Y. Huang, A nonparametric bayesian approach for clustering bisulfate-based dna methylation profiles, BMC Genomics **13**, S20 (2012).

[67] J. Sethuraman, A constructive definition of dirichlet priors, Statistica Sinica **4**, 639 (1994).

[68] M. K. Titsias, The infinite gamma-poisson feature model, in *Advances in Neural Information Processing Systems*, pp. 1513–1520, 2008.

[69] M. Krnjajić, A. Kottas, and D. Draper, Parametric and nonparametric bayesian model specification: A case study involving models for count data, Computational Statistics & Data Analysis **52**, 2110 (2008).

[70] M. Zhou, L. Hannah, D. B. Dunson, and L. Carin, Beta-negative binomial process and poisson factor analysis, in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 1462–1471, 2012.

[71] V. Zeimpekis, G. M. Giaglis, and G. Lekakos, A taxonomy of indoor and outdoor positioning techniques for mobile location services, ACM SIGecom Exchanges **3** (2002).

[72] G. Sun, J. Chen, W. Guo, and K. R. Liu, Signal processing techniques in network-aided positioning: a survey of state-of-the-art positioning designs, Signal Processing Magazine **22**, 12 (2005).

[73] J. Paek, K.-H. Kim, J. P. Singh, and R. Govindan, Energy-efficient positioning for smartphones using cell-id sequence matching, in *Proceedings of the 9th International Conference on Mobile systems, Applications, and Services*, pp. 293–306, ACM, 2011.

[74] C.-C. Yu and H.-P. Chang, Personalized location-based recommendation services for tour planning in mobile tourism applications, in *Proceedings of the International Conference on Electronic Commerce and Web Technologies*, pp. 38–49, Springer, 2009.

[75] Z. Cheng, J. Caverlee, K. Lee, and D. Z. Sui, Exploring millions of footprints in location sharing services., in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media* Vol. 2011, pp. 81–88, 2011.

[76] B. Berjani and T. Strufe, A recommendation system for spots in location-based online social networks, in *Proceedings of the 4th Workshop on Social Network Systems*, ACM, 2011.

[77] A. Kittur, E. Chi, B. Pendleton, B. Suh, and T. Mytkowicz, Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie, World Wide Web **1**, 19 (2007).

[78] A. Fuxman, P. Tsaparas, K. Achan, and R. Agrawal, Using the wisdom of the crowds for keyword generation, in *Proceedings of the International Conference on World Wide Web*, ACM, 2008.

[79] P. Mendes, A. Passant, and P. Kapanipathi, Twarql: tapping into the wisdom of the crowd, in *Proceedings of the 6th International Conference on Semantic Systems*, p. 45, ACM, 2010.

[80] M. Goodchild, Citizens as sensors: web 2.0 and the volunteering of geographic information, GeoFocus (Editorial) **7** (2007).

[81] A. Sheth, Citizen sensing, social signals, and enriching human experience, IEEE Internet Computing **13** (2009).

[82] B. Sigurbjörnsson and R. Van Zwol, Flickr tag recommendation based on collective knowledge, in *Proceedings of the International Conference on World Wide Web*, pp. 327–336, ACM, 2008.

[83] E. Moxley, J. Kleban, and B. Manjunath, Spirittagger: a geo-aware tag suggestion tool mined from flickr, in *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, ACM, 2008.

[84] S. Johnson, Hierarchical clustering schemes, Psychometrika **32** (1967).

[85] J. Paolillo and E. Wright, Social network analysis on the semantic web: Techniques and challenges for visualizing foaf, Visualizing the Semantic Web **2** (2005).

[86] G. W. Milligan, An examination of the effect of six types of error perturbation on fifteen clustering algorithms, Psychometrika **45**, 325 (1980).

[87] A. El-Hamdouchi and P. Willett, Comparison of hierarchic agglomerative clustering methods for document retrieval, The Computer Journal **32** (1989).

[88] P. Langfelder, B. Zhang, and S. Horvath, Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r, Bioinformatics **24** (2008).

[89] I. Witten and D. Milne, An effective, low-cost measure of semantic relatedness obtained from wikipedia links, in *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, 2008.

[90] C. A. Lampe, N. Ellison, and C. Steinfield, A familiar face (book): profile elements as signals in an online social network, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 435–444, ACM, 2007.

[91] S. Zhao, S. Grasmuck, and J. Martin, Identity construction on facebook: Digital empowerment in anchored relationships, Computers in Human Behavior **24** (2008).

[92] H. Liu, P. Maes, and G. Davenport, Unraveling the taste fabric of social networks, International Journal on Semantic Web and Information Systems **2** (2006).

[93] E. Gabrilovich and S. Markovitch, Wikipedia-based semantic interpretation for natural language processing, Journal of Artificial Intelligence Research **34** (2009).

[94] J. Gracia and E. Mena, Web-based measure of semantic relatedness, in *Proceedings of the 9th International Conference on Web Information Systems Engineering* Vol. 5175, p. 136, Springer, 2008.

[95] C. Shannon, W. Weaver, R. Blahut, and B. Hajek *The mathematical theory of communication.* Vol. 117 (University of Illinois press, Urbana, 1949).

[96] J. Lin, Divergence measures based on the shannon entropy, IEEE Transactions on Information Theory **37** (1991).

[97] A. Ihler, J. Hutchins, and P. Smyth, Adaptive event detection with time-varying poisson processes, in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 207–216, ACM, 2006.

[98] H. Heffes and D. Lucantoni, A markov modulated characterization of packetized voice and data traffic and related statistical multiplexer performance, IEEE Journal on Selected Areas in Communications **4**, 856 (1986).

[99] S. L. Scott, A bayesian paradigm for designing intrusion detection systems, Computational Statistics & Data Analysis **45**, 69 (2004).

[100] K. Yamanishi, J.-I. Takeuchi, G. Williams, and P. Milne, On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms, Data Mining and Knowledge Discovery **8**, 275 (2004).

[101] A. Haghighi and D. Klein, Unsupervised coreference resolution in a nonparametric bayesian model, in *Proceedings of the Annual meeting-Association for Computational Linguistics* Vol. 45, p. 848, 2007.

[102] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky, Describing visual scenes using transformed dirichlet processes, Advances in Neural Information Processing Systems **18**, 1297 (2006).

[103] E. Marshall and D. Spiegelhalter, Identifying outliers in bayesian hierarchical models: a simulation-based approach, Bayesian Analysis **2**, 409 (2007).

[104] J. Wakefield, A. Smith, A. Racine-Poon, and A. Gelfand, Bayesian analysis of linear and non-linear population models by using the gibbs sampler, Applied Statistics , 201 (1994).

[105] F. A. Quintana and P. L. Iglesias, Bayesian clustering and product partition models, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **65**, 557 (2003).

[106] K. Chaloner and R. Brant, A bayesian approach to outlier detection and residual analysis, Biometrika **75**, 651 (1988).

[107] T. S. Ferguson, A bayesian analysis of some nonparametric problems, The Annals of Statistics , 209 (1973).

[108] R. M. Neal, Markov chain sampling methods for dirichlet process mixture models, Journal of computational and Graphical Statistics **9**, 249 (2000).

[109] D. J. Aldous, *Exchangeability and related topics.* (Springer, 1985).

[110] M. D. Escobar and M. West, Bayesian density estimation and inference using mixtures, Journal of the American Statistical Association **90**, 577 (1995).

[111] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil, and C. Mascolo, A tale of many cities: universal patterns in human urban mobility, PloS One **7**, e37027 (2012).

[112] L. Barkhuus, B. Brown, M. Bell, S. Sherwood, M. Hall, and M. Chalmers, From awareness to repartee: sharing location within social groups, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 497–506, ACM, 2008.

[113] A. Le, M. Gertz, and C. Sengstock, An event-based framework for the semantic annotation of locations, in *Advances in Databases and Information Systems*, pp. 248–262, Springer, 2014.

[114] T. Rattenbury, N. Good, and M. Naaman, Towards automatic extraction of event and place semantics from flickr tags, in *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 103–110, ACM, 2007.

[115] C. Biancalana, F. Gasparetti, A. Micarelli, and G. Sansonetti, An approach to social recommendation for context-aware mobile services, ACM Transactions on Intelligent Systems and Technology **4**, 10 (2013).

[116] H. Yin, Y. Sun, B. Cui, Z. Hu, and L. Chen, Lcars: A location-content-aware recommender system, in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 221–229, ACM, 2013.

[117] X. Liu, Y. Tian, M. Ye, and W.-C. Lee, Exploring personal impact for group recommendation, in *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pp. 674–683, ACM, 2012.

[118] Z. Qiao, P. Zhang, Y. Cao, C. Zhou, L. Guo, and B. Fang, Combining heterogenous social and geographical information for event recommendation, in *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, 2014.

[119] M. Sahami and T. D. Heilman, A web-based kernel function for measuring the similarity of short text snippets, in *Proceedings of the 15th International Conference on World Wide Web*, pp. 377–386, ACM, 2006.

[120] Y. Li, Z. A. Bandar, and D. McLean, An approach for measuring semantic similarity between words using multiple information sources, IEEE Transactions on Knowledge and Data Engineering **15**, 871 (2003).

[121] E. Gabrilovich and S. Markovitch, Computing semantic relatedness using wikipedia-based explicit semantic analysis, in *Proceedings of the International Joint Conference on Artificial Intelligence* Vol. 7, pp. 1606–1611, 2007.

[122] D. Sankoff and J. B. Kruskal, *Time warps, string edits, and macromolecules: the theory and practice of sequence comparison.* (Addison-Wesley, 1983).

[123] A. Freitas, D. Carvalho, and E. Curry, EasyESA: A Low-effort Infrastructure for Explicit Semantic Analysis, in *Proceedings of the 13th International Semantic Web Conference*, 2014.

List of Figures

List of Tables