Ming Zhan (mzhan@abo.fi)

Exploring the Feasibility of Applying Data Mining for Library Reference Service Improvement

A Case Study of Turku Main Library

Master thesis in
Information and Knowledge Management
Master's Programme
Supervisors: Gunilla Widén
Faculty of Social Sciences, Business and Economics
Åbo Akademi University
Åbo 2016

Abstract

Data mining, as a heatedly discussed term, has been studied in various fields. Its possibilities in refining the decision-making process, realizing potential patterns and creating valuable knowledge have won attention of scholars and practitioners. However, there are less studies intending to combine data mining and libraries where data generation occurs all the time. Therefore, this thesis plans to fill such a gap. Meanwhile, potential opportunities created by data mining are explored to enhance one of the most important elements of libraries: reference service. In order to thoroughly demonstrate the feasibility and applicability of data mining, literature is reviewed to establish a critical understanding of data mining in libraries and attain the current status of library reference service. The result of the literature review indicates that free online data resources other than data generated on social media are rarely considered to be applied in current library data mining mandates. Therefore, the result of the literature review motivates the presented study to utilize online free resources. Furthermore, the natural match between data mining and libraries is established. The natural match is explained by emphasizing the data richness reality and considering data mining as one kind of knowledge, an easy choice for libraries, and a wise method to overcome reference service challenges. The natural match, especially the aspect that data mining could be helpful for library reference service, lays the main theoretical foundation for the empirical work in this study.

Turku Main Library was selected as the case to answer the research question: whether data mining is feasible and applicable for reference service improvement. In this case, the daily visit from 2009 to 2015 in Turku Main Library is considered as the resource for data mining. In addition, corresponding weather conditions are collected from Weather Underground, which is totally free online. Before officially being analyzed, the collected dataset is cleansed and preprocessed in order to ensure the quality of data mining. Multiple regression analysis is employed to mine the final dataset. Hourly visits are the independent variable and weather conditions, Discomfort Index and seven days in a week are dependent variables. In the end, four models in different seasons are established to predict visiting situations in each season. Patterns are realized in different seasons and implications are created based on the discovered patterns. In addition, library-climate points are generated by a clustering method, which simplifies the process for librarians using weather data to forecast library visiting situation. Then the data mining result is interpreted from the perspective of improving reference service. After this data mining work, the result of the case study is presented to librarians so as to collect professional opinions regarding the possibility of employing data mining to improve reference services. In the end, positive opinions are collected, which implies that it is feasible to utilizing data mining as a tool to enhance library reference service.

Key words: Data mining, Library reference service, Service improvement, Case study

Content

1. Introduction

Libraries, as the nexus of information and knowledge, provide various services for satisfying citizens' requirements. Historically, the quality of a library was mainly evaluated by its collections (Hernon and Altman, 2010). Vast volume of collection means plentiful materials for patrons' self-learning process. Nevertheless, what a library has has gradually been replaced by what a library does owing to the development of technology and the increasing needs in personal information management. Therefore, library service quality has been distinguished from library quality. How to improve the service becomes a major question within various libraries.

However, what a library does ought to consider what a library has. Therefore, resources of a library need to be carefully viewed before carrying out detailed service providing schemes. After critically reviewing empirical studies of public service improvement, Boyne notices that resources play a role to improve the quality of services. It is because the more resources can be utilized, the better the service will be (Boyne, 2003). This viewpoint can also be merged to the theory of library service improvement as most libraries are crucial elements of public sectors. Nevertheless, the acute awareness of ecological destruction is not only accompanied by the concept "Sustainable Development" (Dempsey et al., 2011), but also highlights a tough situation in the world: resource limitation which potentially constrains public services from being developed to a higher level. However, there is one exception: data, the volume of which, as oppose to such limited condition, has been dramatically increasing. In 2008, Anderson put forward a term: Petabyte Age, which is used to describe that the amount of data is so vast that it should be stored in the cloud rather than on practical discs (Anderson, 2008). Gordon-Murnane (2012) notices the advent of data generation as well and puts forward opportunities and possibilities for libraries' responsibility. Therefore, data could be a reasonable choice of resource for libraries developing their services. Because data is not as limited as other resources. Since what a library dose might define the service quality, what a library dose with data could better assist the formation of library service in this data explosion generation.

Data mining, referring to extracting or mining knowledge from a large amount of data (Uppal and Chindwani, 2013), has been suggested then practically applied to organizations. Its great potential for decision-making, forecasting, pattern realization (Perner, 2002) are widely recognized by scholars and practitioners. Nonetheless, there are few studies focusing on the applicability of data mining on library service improvement. As a matter of fact, a library is a natural data generator. That is to say, data mining could be readily achieved within a library. Therefore, studying data mining application concerning library service improvement offers a great opportunity to gain valuable outcomes. Moreover, a data-intensive generation is coming upon us, which would accelerate this process. As such, could data mining be an effective solution to balance resource limitation and service improvement in libraries where the existence of data growth is already a reality? This is the leading question of the current study to carry out the detailed research steps.

This study aims at exploring the possibility to apply data mining to improve library services. In order to generate a more specific outcome, library reference service is selected as the major study domain because providing reference service is considered as one of the main functions of current libraries (Standerfer, 2006, p. 139). In order to demonstrate the feasibility of data mining in enhancing reference service, a case study is conducted. Turku Main Library is chosen as the case, the number of daily visits in this library is collected for data mining. In addition, the weather condition data is collected from the website Weather Underground in the same time period as daily visits. The combination of these two datasets aims to discover patterns which can be employed to enhance reference service in Turku Main Library. After interpreting the result of mining these datasets, the implication is presented to librarians. The opinions of these professionals will provide strong evidence to explain whether data mining is feasible for the improvement of reference service.

1. 1 The motivation of the thesis work

Firstly, with the development of information communication technology (ICT), the amount of data generated, stored or processed has been surging dramatically. As is pointed out by Chen et al. (2014), advances of IT make it easier to generate data and the fast development of cloud computing techniques further accelerate such processes. Furthermore, cloud techniques can simplify data access and storage as well. As such, the general scientific paradigms have also evolved from empirical science, theoretical science, computational science to data-intensive science (Chen and Zhang, 2014). Therefore, it is an up-to-date approach to conduct research through the lens of data-intensive method. In this study, the significance of data around libraries and the potential of such data are explored, which disclose a path of libraries to enter into this new scientific paradigm. Therefore, this thesis is motivated to follow the new trend: data-intensive science.

There is a research gap between data mining and reference service. As is shown in Table 1:

Table 1: The result of data mining and library related key words in Web of Science

Science Category	Retrieval Method	Results
Information science and library science	Topic “data mining” and Title “library”	57
	Topic “data mining” and Title “library service”	6
	Topic “data mining” and Title “reference service”	0

Taking Web of Science as an example, the key word combination is presented in Table 1. It can be concluded that there are not many studies concerning introducing data

mining into libraries. When it comes to the library service, the result (six) indicates that fewer studies have thoroughly seen data mining as a tool for improving library service. Even though only considering Web of Science is not representative enough because Web of Science does not include all disciplines, the zero combination of data mining and reference service still discloses a gap or at least a lack of attention to these two areas. Therefore, another motivation for this study is to demonstrate the possibilities of utilizing data mining to enhance reference service in order to fill in the gap to some extent.

Last but not least, most studies concerning data mining in libraries only shed light on data sets within or related to library content as is discussed in the second section, such as bibliometrics, user borrowing history, comments under library social media homepages. There are few studies mining data sets which are outside of the library domain. The combination between such data with certain library data might have a chance for pattern recognition or knowledge creation. Therefore, the final motivation of this thesis work is to enlarge the scope of useful data sets for library data mining mandates.

1.2 The aim of the thesis work

Considering the leading research question and research motivations, three aims are decided in this study:

Firstly, discovering useful free databases for reference service improvement.

Various online data, such as social medium data, online consumer comments, stock index etc. can be easily approached. Therefore, one of the main reasons for this study will be to achieve some useful free and open resources for library use.

Secondly, putting forward pragmatic ideas for applying data mining in libraries.

In order to come up with practical knowledge for libraries, machine learning methods or algorithms will be employed so as to put forward concrete ideas to present that data mining could be employed to enhance reference service. Achieving this goal will enrich the research about data mining application in the library domain.

Last but not least, demonstrating the feasibility of data mining in the context of reference service.

Although data mining can be proved as a worthwhile approach, how to launch it in a library to achieve better reference services would be a complex procedure. Various issues need considering. Therefore, after getting pragmatic application interpretation, interviews will be carried out to demonstrate the applicability and feasibility of data mining for reference service improvement. With interviewing librarians, how to explore the benefits of data mining will be discussed with professional perspectives. In

the end, the feasibility of applying data mining for reference service improvement will be demonstrated theoretically and practically.

1.3 The organization of the thesis work

This thesis work is organized as follows: first of all, the result of a literature review is presented, which includes two contents: literature about data mining in the library context and literature about library reference service. In the first content, the meaning, forms, benefits, challenges and the process of data mining in the library is explained. In the second content, the definition, forms, challenges of library reference service are reviewed. Meanwhile, ideas to enhance the reference service are summarized as well. In the third section, the natural match of data mining and the library is explicated based on previous studies. Then the methodology of this thesis is discussed in the fourth section. The result of mining the eventually chosen data sets is presented meanwhile the implications generated on the result are also displayed. Then the evaluation on the feasibility of the implication is presented in order to eventually achieve an understanding of whether data mining could be helpful in library reference service improvement. After that, the result of the empirical work is discussed and the conclusion of this study is made. In the final section, the expectations for future studies are put forward.

2. Literature review

As current study is partly motivated by filling the research gap, literature reviews concerning data mining in the context of libraries and library reference service were conducted. In the review, definition, forms, benefits and challenges on each topic are discussed. The results of the literature review provide hints to combine these two topics and thus lay the theoretical foundation for the study.

2.1 Data mining in the library environment

Libraries are confronting various data every day, e.g. the author demographic information, publication time or citation information of library collections, user profiles, website browsing histories, state policies, daily news, healthcare information, company whitepapers. It can be obviously summarized that data is generated in and outside the library. With carefully handling, data would be a valuable sources for libraries and data mining, referring to extracting or mining knowledge from large amount of data (Uppal and Chindwani, 2013), would be a proper and suitable approach to explore values from such resources. As is mentioned by Banerjee, "There is simply too much information to process manually, so increasing our reliance on…data mining tools seems to be just a matter of time". (Banerjee, 1998, p. 29) Therefore, it might be an up-to-date idea to understand data mining in the context of libraries.

2.1.1 The meaning of data mining in libraries

According to Wang et al., data mining is a process to extract connotative and unknown but useful information and knowledge from data (Wang et al., 2005), which enrich technologies concerning library databases. This definition may not be stated exactly the same with those cited or understood in other studies, but its main content is widely agreed by (Banerjee (1998), Dumouchel and Demaine (2006), Yan et al. (2010)). Additionally, knowledge discovery has been emphasized in library data mining correspondingly owing to the fact that libraries are the kernel of information and knowledge. In the view of Dumouchel and Demaine (2006), digital libraries, "institutions or organizations that provide information services" (Borgman, 1999, p. 239), provide a condition where knowledge discovery techniques could be utilized for advancing work

Moreover, the unique term, bibliomining combining data mining and bibliometrics, has been created to explicate a special data mining application in the context of libraries. Bibliomining was firstly created by Scott Nicholson, who created this word to differentiate key words between "data mining for libraries" and "libraries for data mining" and eventually help researchers easily approach their needed resources (Nicholson, 2003b) out of information retrieval. Bibliomining is defined as the process of pattern recognition from behavior-based datasets of library systems (Nicholson, 2003b). One summary could be made here: all data mined in the process of bibliomining is produced within the library. Or put in another way, merely part of data around the library is reached. In the light of Nicholson and Stanton's research, there are

three main data sources for bibliomining: the creation of the library system, the usage of the library system and the external data sources (Nicholson and Stanton, 2003).

Bibliographic information is the main type of data during the creation of the library system. The records of collections represent the storage capability of a library. During the operation of catalogue, numerous data can be generated for the access and location of materials.

There are three types of data generated during the usage of library system (Nicholson and Stanton, 2003, pp. 253-254):

- User information refers to demographic information of library patrons'. It can be analyzed so as to realize proper user classification.
- Circulation information concerns the situation of library items circulated among library users, for instance, the borrowing history of a book. Such data could be useful for library managers to purchase or remove materials.
- Searching and navigation information is generated when individuals try to find bibliometric information in library databases or websites, potential needs or service improvement could be achieved through mining such information.

Data stemmed from the following three sources are considered as the external data sources (Nicholson and Stanton, 2003, pp. 255-256)

- Reference desk interactions. In libraries, the reference desk is an important interface to communicate with library users. Questions, issues or concerns are put forward here, which provides the basis for interaction data for user understanding;
- Item use information. This is related to the in-house using of library items.
- Interlibrary loan. The cooperation among libraries leads to the creation of great amounts of data. Furthermore, such data is not only related to users, but it also discloses some requirements for library staff.

Even though, diverse data resources are covered under the concept of bibiliomining, data from social media, public websites, governments etc. are not included. If all these data resources are considered, more opportunities would be created to support library data mining conduction. Thereby, bibliomining cannot thoroughly reflects the whole picture of library data mining. Data mining has different meanings according to the data resource employed for pattern recognition in the library environment.

2.1.2 Different forms of data mining in libraries

Since data resources are diverse around a library, the form of utilizing data shows differences as well. Based on the result of literature review, there are generally four types of data resources employed for data mining in the library context: library websites, social media websites, bibliometric data and user-based data. Thus, the forms of data mining in libraries could be classified into: web mining, text mining, bibliometric data mining and user-based data mining.

2.1.2.1Web mining

Web mining refers to all the methods aiming at extracting valuable information from data generated on the web (Velásquez, 2013). When it comes to the domain of libraries, web mining means extracting information or knowledge from the library website. In light of Wang et al., there are three types of web mining in the library: web content mining, web structure mining and web usage mining (Wang et al., 2005). Web content mining creates information from web page content (Pol et al., 2008). Thereby, all the images, information, audios etc. recorded on a library web page could be valuable resources for web content mining. For example, effective information retrieval models are established resulting from mining the documents recorded on the library website. Advanced searching methods could be created during this process (Klampfl et al., 2014). Web structure mining comprises link mining, inter structure mining and HTML mining (Pol et al., 2008). Zuccala et al. consider links to a library web page as significant information because it reflects why users view a page. As such, they launch a case study of National Electronic Library for Health (NeLH) to demonstrate the feasibility of mining library web structures to satisfy user information needs. It is concluded that insights concerning who is using the website or when is the busy time for the website are attained through analyzing transaction log files. (Zuccala et al., 2007) Web usage mining aims to extracte knowledge from data disclosing user behaviors. User profiles, logs are the main resource. The application design put forward by Kamdar and Joshi is a good example of web usage mining. They take user prior traversal patterns as the database designing a fuzzy incremental clustering algorithm. In the end, personalized web pages are created. (Kamdar and Joshi, 2005)

2.1.2.2 Text mining

Text mining is the process of discovering useful patterns or knowledge from a text. Within a library, text mining is a twofold concept: on one hand, it is a data mining method used in web content mining; on the other hand, it is one of data mining types. In this section, text mining means the latter, mining the textual materials recorded in the library system and tweets from libraries' twitter homepage. Text mining can facilitate searching process by enlarging metadata and highlighting items in the document (Witten et al., 2004) in the library system. Meanwhile, mining tweets of library followers' is promoted as an effective way to communicate with library users. (Cuddy et al., 2010, Sewell, 2013)

2.1.2.3 Bibliometric data mining

Bibliometric data includes information about articles such as the authorship and citation situation. Meanwhile, metadata related to the article, for instance key words, studying areas, journals who published that article is also considered as bibilimetric data (Nicholson, 2006). Bibliometric data is normally known as quantitative and literature-based indicators, for instance publication and citation data (Moed et al., 1985, p. 131). To mine such data is called bibilometrics. According to Pendlebury (2010), one

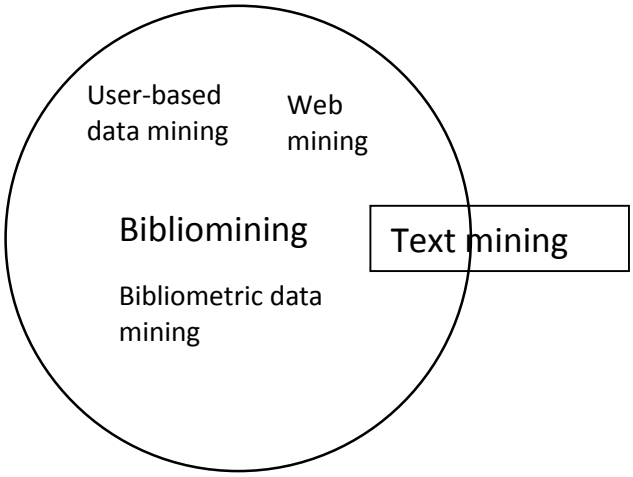
of the most biggest business intelligence corporation, bibilometrics is the process applying quantitative analysis and statistics on articles published in journals, newspapers, websites etc. and the citation data of these articles. During this process, useful patterns are discovered from citation data, publication data or other kinds of data related to circulation. In the whitepaper of Thomson Reuters (Pendlebury, 2010), it is pinpointed that bibliometrics is helpful for libraries to realize journal usage patterns moreover to identify the most required journals for future subscription. The importance of bibilometrics has been mentioned by studies. For example, Dumouchel and Demaine (2006) employ former cases about Main Path Analysis and Linked Literature Analysis, both of which are main methods to conduct social network analysis and they indicate that scientific information with proper operation will make great scholarly progress. Cleyle and Nicholson (2006) demonstrate the effectiveness of bibilometrics for evidence-based librarianship and put forward the future paths for this field. Meanwhile, empirical research is conducted as well to further prove the significance of bibilometrics. Alotaibi et al. (2015) mine bibliometric data to present the most cited works in aneurysmal subarachnoid hemorrhage and a paradigm shift from clinical practice to endovascular treatment is identified. Ajay and Sangamwar (2014) achieve 10 year chronological changes, international filing and grant trend, patent licensing pattern etc. with the help of mining bibliometric data of patents. The results map out a clear picture of Indian intellectual properties and implies the feasibility of mining bibiliometric data.

2.1.2.4 User-based data mining

As is recorded in the library system, the detailed information of users is a main composition of library data. The education background, jobs, ages, gender and so forth all make a difference in users' behavior of library item usage. Therefore, focus has been made on such data. Yan et al. (2010) establish a network based on user booking loan records. It is confirmed that the major is the key factor to influence students' book loan behaviors. Hajek and Stejskal (2012) conduct a project to measure the value of library service, and they employ K-mean algorithms to mine user-based data. In the end, typical readers are classified by socio-economic and demographic characteristics. The frequency of visits to the library is identified as well. The result contributes to the improvement of library service with the perspective of typical readers rather than financial allocation. Karno et al. (2012) collect user information records from the library data source including 957,224 pieces of borrowing history. Then borrowing trend, types of degrees and schools, popular works etc. are realized and user patterns are reflected correspondingly.

According to the review above, it can be concluded that data mining in the library world is mostly in the domain of bibliomining. As is shown in Figure 1:

Figure 1: The relationship between bibliomining and four data mining types in the library



Web mining, bibliometric data mining and user-based data mining are all included in bibliomining according to the bibliomining resources summarized in 2.1.1. Only mining texts from twitter or other social media is outside the domain. However, texts from social media are generated on the homepage of the library. Therefore, such data can also be considered as one part of library data. However, various data generated on the Internet are open and free, such online data have been empirically indicated that they could be a valuable recourse for organizations. For example, Xiang et al. (2015) utilized online consumer reviews as a data source to understand users behaviors in the hotel industry. In the end, a key word dictionary is created, which can be used to locate consumer needs. When it comes to the library, few similar studies are carried out. Simply put, there is a research gap between data mining and the library, which is mining online free data.

2.1.3 The benefits and issues of data mining for libraries

The increasing understanding of data mining in the domain of libraries indicate that benefits could be disclosed through mining library data. As is earlier stated by Banerjee (1998), there are two main potentials of data mining: firstly, faster and wider access can be provided compared with manually cataloging; secondly, it is easy to be learnt by librarians or users with low abilities in analytic skills so that their needs can be readily satisfied. With the application scope of data mining getting larger, more benefits have been explored. These benefits can be classified into six categories:

2.1.3.1 Creating new knowledge.

Libraries are containing various textual documents, which makes them an environment for knowledge discovery thus advancing users' work (Dumouchel and Demaine, 2006). Wisely applied, data mining would be an effective approach to generate new knowledge.

Cases listed by Dumouchel and Demaine (2006), undiscovered public knowledge and social network analysis, are good examples to prove this viewpoint.

2.1.3.2 Improving collection process

How to execute a logical collection process is of prime significance for libraries because collection is one of the most indispensable factors to evaluate library service quality. (Yankova, 2013). Therefore, employing data mining to refine collection process could be one of the benefits. Nicholson (2003a) designed a tool to automatically search web pages including scholarly research work with the help of bibliomining. Four exploration models were established with logistic regression, discriminant analysis, classification tree and neural work to discover web-based scholarly research outcomes. In the end, a tool is produced as a filter to aid the collection task. Based on the review work of Wang et al. (2005), online investigation, message note or other types of data recorded in digital library system can be analyzed to guide the library information source collection.

2.1.3.3 Refining recommendation

According to Nicholson and Stanton (2003), circulation histories could be a useful resource to advise users which work would be more related to their current task or help them locate the correct material. Such operation has been utilized by many libraries with bibiliomining. Hwang and Lim (2002) shed light on an approach to recommend library books. They set users' preference as the core standard to research the recommendation approach, precious transaction records, web logs and cookies data are mined. Then, a new approach is pinpointed by relating demographic information with product types.

2.1.3.4 Supporting decision making

How to make a rational decision has a strategic impact on organizations so that scholars devote themselves to finding possibilities in data mining to help libraries make decisions. Cleyle and Nicholson (2006) present a different path to evidence-based librarianship. Data mining is employed to measure and evaluate library service. As a result, a well-defined decision could be made to further support the development of libraries. The decisions made by library staff can be aided by bibilomining because patterns of behaviors can be noticed. This is advantageous to decide detailed operations, such as arranging the optimized staff schedule or reasonable number of on-duty librarians (Nicholson and Stanton, 2003).

2.1.3.5 Assisting library marketing

Twitter has been empirically proved that it is an important implementation for library promotion activities. Ads concerning library events or new arrivals can be easily reached by users through posting tweets on libraries' Twitter homepage. (Cuddy et al., 2010) As such, Sewell (2013) launched a study to explore insights of Twitter user accounts. In the end, the behaviors of different user groups, for instance students,

faculty and staff, Texas A&M Universities, alumni, corporations and outside followers and other library/librarian, were discovered. Even though Twitter is not the only communication method in the studied case, the feasibility of mining twittes implies potentials of such method to plan marketing events.

2.1.3.6 Understanding library users

Data of users are widely mined and explored in libraries. The aforementioned benefits, improving collection process, refining recommendation, supporting decision making and assisting library marketing, are all based on the understanding of library users. Furthermore, other forms of user behaviors are noted as well through data mining. Yan et al. (2010) conduct a study to analyze users' book-loan behaviors. Book-loan logs are mined and different borrowing trends and knowledge dependency are discovered. Karno et al. (2012) also consider book borrowing histories as the main datasets for exploring user patterns. Eventually, clusters with different user groups and book categories are created. The similar results are also achieved by studies of Hajek and Stejskal (2012).

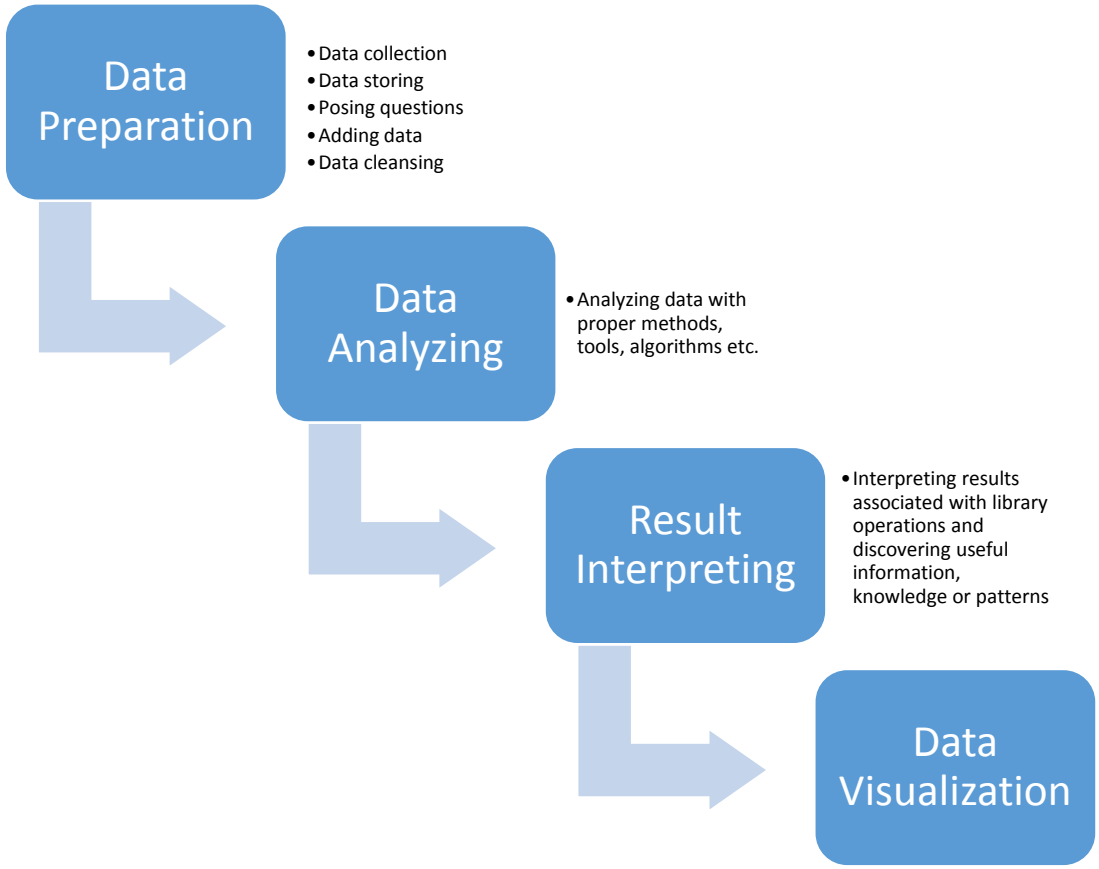
Although many benefits are generated, issues caused by data mining cannot be neglected either. From the theoretical point of view, currently most papers regarding data mining in library science are written by scholars, few librarians participate in this group. Thereby, the result could be separated by practical operations. As was proposed by Cleyle and Nicholson (2006), a great deal of librarians should contribute to the research, otherwise the content of the research cannot be sufficient. From the practical point of view, the development is blocked owing to no clear and universal standard to mine data in libraries. Issues like how to retrieve information from datasets, which operations should be employed for data pre-processing, how to extract information from shared databases etc. are obstacles to achieve a widely accepted data mining results. This point is reflected by studies (Banerjee, 1998, Hajek and Stejskal, 2012, Karno et al., 2012, Okerson, 2013). When it comes the more detailed level, issues are even more. One of the premises to mine data is to store it properly. This requires advanced capabilities for libraries to record various types of data (Hajek and Stejskal, 2012, Okerson, 2013). The access to different data resources is also complicated (Okerson, 2013, Nicholson and Stanton, 2003). Within one library, staff have diverse levels of access to databases, which mainly decides how should the data mining work be accomplished, not to mention, the access difference between library users and library staff, library networks. In addition, the skills needed to realize data mining is not well attained by librarians (Nicholson and Stanton, 2003). This issue directly determines how professional the data mining task could be handled in knowledge discovery or pattern analysis as data mining is a manual conduct. User information is clearly recorded in library systems, which leads to another concern: privacy issue (Nicholson and Stanton, 2003). User data mining is one of the main library data mining types. During this process, sensitive demographic information of users might be reached, such as home address, email address and phone number. How to protect user information and how to scrutinize such information properly need to be managed well.

Data mining as a rising topic in library science has been discussed, benefits are emphasized and implied by many pragmatic cases. Whereas, issues caused by data mining are worthy of attention as well. To gain a balance between benefits and issues of data mining could be a task for every library.

2.1.4 The process of library data mining

According to the studies (Cleyle and Nicholson, 2006, Hajek and Stejskal, 2012, Karno et al., 2012, Nicholson, 2003a, Sewell, 2013, Yan et al., 2010), the data mining process in libraries can be classified into four phases: data preparation, data analyzing, results interpreting and data visualization, as shown in Figure 2:

Figure 2: The process of library data mining



The process of data preparation starts with data collection. Data from different sources or forms are collected, such as user information, book loan histories, collection statistics etc. All these data will be stored in the library system, where data will be formatted in order to accelerate the analysis. Questions, concerning services, user satisfaction or library resource allocation, guide the forms and directions of data mining and they determine which data should be mined as well. With the posed questions, specific data

can be extracted. As such, the brief introduction about data could be reached. What the main data format is or how the data quality is could be realized. Meanwhile, additional data might be integrated if needed. Then data cleansing should be conducted. This is similar to the data pre-processing step. Since various forms and types of data could be explored, data noise or data duplication must exit in library data warehouses. Data cleansing is the process to limit the affect caused by such issues. For example, before text mining, stop words, words are indispensable for sentence structure but not useful for explaining the content, are deleted so that proper insights can be discovered. In a word, data cleansing is the process to merely keep the most related data for mining. When the data is ready, analytic tools should be applied to dig potential meanings under the cover of such data. This is the core process of data mining, which decides the reliability and credibility of the discovered information. With the result from the former process, practical interpretations upon the results are required. All the explored information, knowledge or patterns should be related to pragmatic library operations, and work for library improvement otherwise there would be no use for libraries to mine data. Last but not least, data visualization should be performed in order to make the result attained by larger audience.

2.2 Literature review on library reference service

Libraries are designed to serve citizens. As the kernel of information and knowledge, the library is one of the places to which individuals or even organizations would like to turn when confronting setbacks. In order to realize the mission, libraries provide diverse services as the approach to communicating with users. Among these, reference service makes a vital difference and it is even considered as the center of library service (Ranasinghe, 2012, Standerfer, 2006).

2.2.1 The definition of reference service

Reference service is the response from libraries regarding users' information requests (Standerfer, 2006), since different ways could be employed to reply to the request, hence reference service has been defined from various perspectives. Modern definitions of reference service are proposed based on the idea put forward by Green (1993), who originally assumed that positive results could be achieved if librarians maintain good interaction with users.

Han and Goulding (2003) consider reference service as services provided by reference libraries, which are professional and supportive for users;

Sharma (2006, p. 8) conceptualizes reference service as "a personal service which aims to provide information to the reader who requires it." According to this definition, reference service is transformed into three motions: to answer the questions asked by users; to provide personal instruction concerning how to utilize library resources; and to realize the maximum usage of libraries collections;

Kuruppu (2007) assimilates reference service as an interface where people interacts with the library the most.

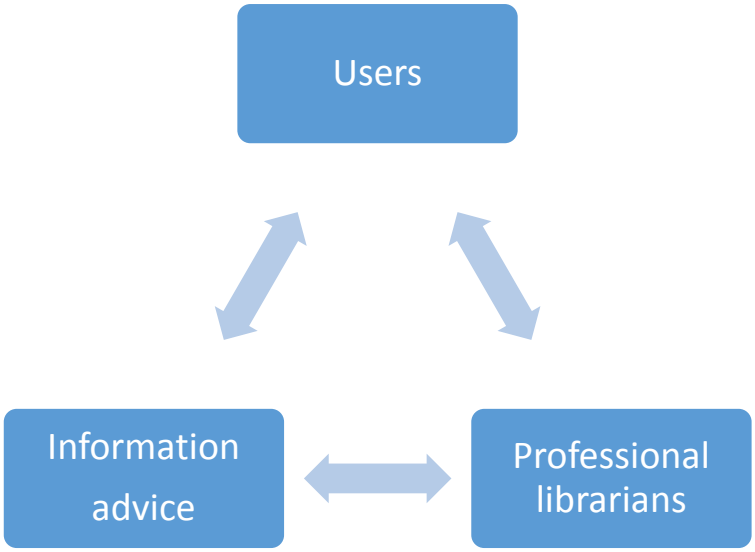
Adebayo (2009) thinks that reference service is composed of three sections: professional assistance, information products and delivery. Professional assistance means the advice presented by librarians based on their professional usage of library materials. Such advice will help users to solve personal issues. Information products include library databases, dictionaries or other infrastructure which can be used for satisfying information requests. Delivery refers to the patterns that librarians are employed to communicate with users so as to ensure that the library materials are effectively reached and properly gained by users. The delivery section includes delivering practical resources such as books, articles and invisible resources for instance ideas, instructions.

Inspired by Green (1993), the interaction or at least communication between the library and the user is emphasized by these aforementioned definitions. Han and Goulding (2003) and Standerfer (2006) share the similar point that the core of library reference service is to respond the information need. In these two definitions, libraries are more led by user requirements than actively provide services, which outlines the main communication pattern in the context of reference service. To further explain the kernel of reference service, both Sharma (2006) and Adebayo (2009) put forward three detailed actions to shed light on the whole process of reference service. In Sharma's opinion, responding to user information requirements is replaced by two activities: to answer the question and to provide helpful instruction. Meanwhile, how to optimize the value of libraries is taken into consideration. From the perspective of Adebayo, the responding part is simplified as professional assistance. Moreover, information products are generated to aid users. Compared with other definitions, Adebayo's uniquely highlights that libraries should take an active role when offering reference service, such as providing information products which can decrease the complexity for users to approach library materials. Even though specific questions may not be answered, patrons are helped by the product. In addition, the factor concerning how to deliver the reference service is covered by Adebayo's definition as well. Since the definition defined by Adebayo comparatively thoroughly contains aspects of reference service, this definition is used in the thesis to guide the study in the following sections.

2.2.2 Three main forms of reference service

According to the definition analysis above, three main elements of reference service can be summarized: users, information advise and professional librarians. These three elements are also recognized by Ranasinghe (2012). The connection between the elements is illustrated as:

Figure 3: The connection between three reference service elements



This connection is proposed in light of Adebayo's definition. Under the circumstance of reference service, users confronting difficulties turn to professional librarians. Librarians tend to explore library sources to solve the problem, and information advice is considered as the solution. The model illustrated in Figure 3 highlights the importance of communication. Because useful messages are exchanged between librarians and patrons in the process of reference service. Therefore, the form of reference service is categorized based on the communication method in this paper, which are face-to-face reference service, telephone based reference service and virtual reference service.

Luo and Weak (2013) launched a study to discuss the perception and usage condition of text reference service. During the data analysis process, an interesting point is raised: most people still would like to directly go to reference desk to ask for help. Based on the analysis results, 86% of participants mentioned the usage of reference desk which implies one of the reference service types: face-to-face. The services they accept are generally resource recommendations, materials identification and instructions on library facilitates. The reason why they still choose the reference desk, especially when many other tools are provided by libraries to boost reference service, is that they are psychologically comforted by librarians. Such feeling is generated from previous positive experiences. As such, face-to-face reference service is not totally obsolete with the advent of many advanced communication technologies. As is stated by Tyckoson (2011), no matter how convenient the communication modes are, users are still willing to come to the library in person to ask for help.

Telephone based reference service is provided along with the wide usage of telephones in individual homes and organizations (Tyckoson, 2011). This could be seen as the way for libraries to offer real time interaction. Nevertheless, limitations do exit in telephone

based reference service. One of the most problematic issues is that all the information is provided in audio format. This means that the possibility of misunderstanding would be greater than with the other two types because users might write down the wrong message during the librarians' dictation.

Virtual reference service is considered as an online based reference service to quickly achieve a satisfactory answer for patrons who will be connected with experts and provided with useful referrals. This definition is derived from the studies of Wasik (1999) and Katz (2013). There are three main forces to encourage the transformation of reference service from traditional ways to Internet-based approaches. They are the learning process towards to the asynchronous condition, the emergence of the Internet generation and the marketization of the library environment (Campbell, 2000). According to Wasik (1999), the origin of digital reference service can be traced back to the period when librarians were keen to enrich the content of reference service thus utilizing electronic resources to achieve the goal. Since digital techniques have been evolved gradually, the main methods to conduct such digital service can be classified into four groups:

2.2.2.1 Email-based reference service (Hull and Adams, 1995, Yang and Dalal, 2015).

Email is the main tool to realize the communication between patrons and libraries in email based reference service. For one thing, email communication can provide convenience for users living far from the library. Furthermore, the effectiveness of email communication is not high enough. Simply put, patrons still need to wait (for perhaps even days to get the feedback), as such, email dose not replace face-to-face models. Nevertheless, on a general level, email can still be used to handle some requires. This leads to the steady usage of email in the library (Hull and Adams, 1995).

2.2.2.2 Instant chat based reference service (Luo and Weak, 2013, Yang and Dalal, 2015)

This method is quite new but it has won worldwide attention. Many libraries employ instant message tools to offer reference service. Compared with email-based service, this method accomplishes the real time communication. The response waiting time can be saved. For those who prefer face-to-face services, instant message can satisfy their requirements of smooth communication and not going anywhere at the same time. Instant chat tools are usually integrated into academic library systems or displayed on the website of public library homepages. Instant chat tools shorten the communication distance in time and space. However, it puts pressure on library staffing and training and increases cost as well (Yang and Dalal, 2015) because libraries need to maintain the working condition of integrated tools whilst answering questions in a limited time and all of these tasks need proper human resources to finish.

2.2.2.3 Mobile application based reference service (Pun, 2015).

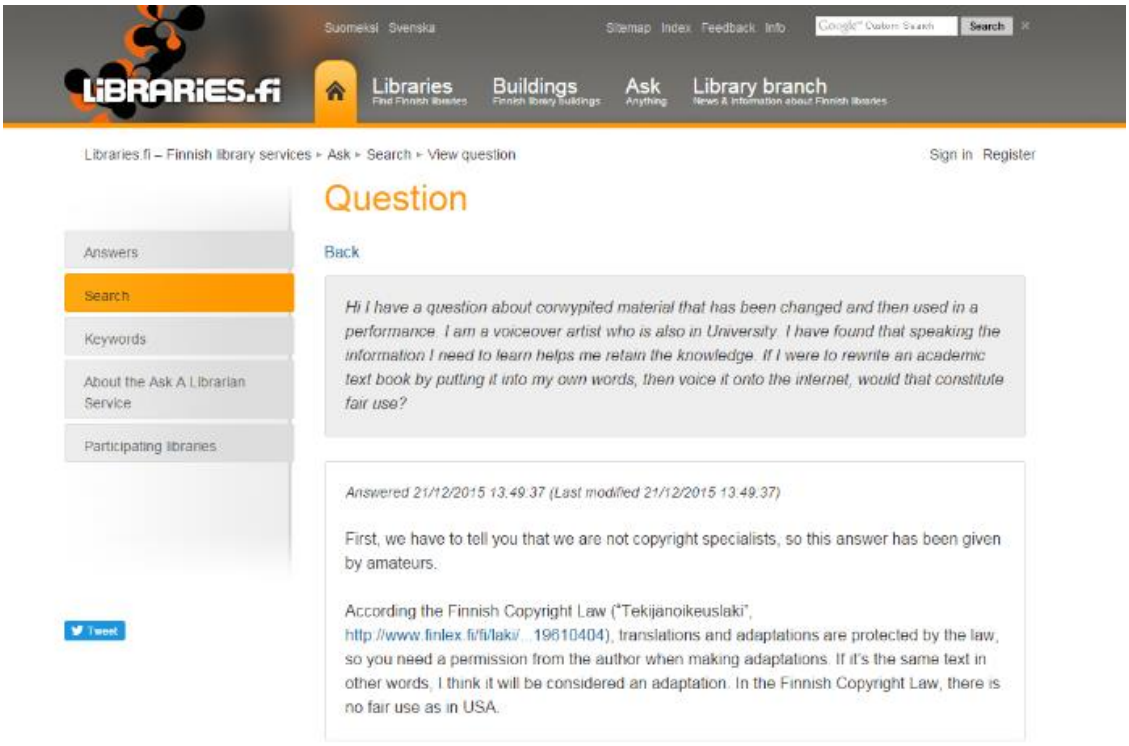
Technically, this service model is a combination of mobile communication and the Internet. Information retrieving is not a brand new activity. Libraries, as the knowledge

pioneer of the society, have feasible motivation to include mobile device for service improvement. Moreover, mobile applications show great potential in information seeking. Compared with other service providing models, the utilization pattern of mobile application in virtual reference service is not very much studied. Nonetheless, positive outcomes have still been achieved. In the view of Pun (2015), mobile applications can be used for multipurpose communication roles, such as encouraging user involvement and locating user communities. Apart from this, virtual services can also be enhanced by wisely applying those apps. The outcome of this study simply proves one point: mobile apps have advantages for discovering user related patterns, and thus eventually accomplish virtual reference service.

2.2.2.4 Homepage Q&A based reference service (Ranasinghe, 2012).

Such service can be found on the homepage of libraries, as is illustrated in Figure 4:

Figure 4: The example of homepage Q&A based reference service



Source: The picture is screenshot on the website: libraries.fi.
This is the official website of Finnish libraries. http://www.libraries.fi/en-GB/ask_librarian/question.aspx?id=db4f73be-7801-4792-8918-3efd8a86e923

One user asks one question on the website and it is answered later with a detailed explanation. Another useful resource is mentioned by the librarian. This virtual reference service model can be seen as a balance between email and instant chat. It not only declines the trouble for patrons using their own email to ask question thus

preventing them from the risk of personal information invading, but it also requires less investment than instant chat tools because the homepage Q&A model can easily be collectively used for libraries within one region.

These four models of virtual reference service almost cover the whole range of the Internet and make the communication between users and libraries much easier and faster. Even though challenges have arisen, the electronic environment can still be the main place for libraries to provide reference service.

2.2.3 Main challenges for reference service

Reference service has been offered in libraries over a long period of time, meanwhile technologies, user habits or information needs have been changed. Therefore, libraries need to update their abilities to refine reference service in order to better satisfy patrons. During this process, libraries will also confront various challenges which obstruct the development of the library service. According to a study conducted by Tyckoson (2011), there are five challenges for managing reference service. These challenges are:

2.2.3.1 Offering a suitable service model.

Tyckoson (2011, pp. 582-587) outlines a dilemma exiting at many libraries: library managers would like to arrange more staff to ensure the quality of service, whilst they would also like to lower costs. More on-duty staff would increase labor costs. Therefore, how to provide a reasonable and suitable service model is problematic. For small libraries, professional librarians provide reference service in a general domain. When it comes to big librarians, subject-oriented reference services are organized. Patrons will be provided with more specific advice. Nevertheless, users are not as familiar with subject classification as librarians are. For example, an individual has a question regarding biometrics and he needs guidance. When he is in the library, he would assume that libraries should have known some information to answer his question and he would just ask any library who is available. But the library might not be a professional in bioscience thereby a proper answer cannot be given. In the end, this person might have a negative opinion towards library service. That is to say, even though subject-oriented reference service will provide detailed services for users, the premise should be that users would behave as libraries wish, otherwise a professional librarian has a great chance of being asked a question which is out of his area of specialty. Therefore, which model could be the suitable one to deliver services is a complex situation for libraries to decide? This challenge is shared by Stevens (2013) as well when the author conducts a literature review to discuss the changes confronted by desk-centric reference services. It is pinpointed that requirements at the reference desk have been declining but libraries still spend the same money ensuring the normal operation of such a desk, because libraries are unable to shut this service model down. Thus, resources are not optimized. In Stevens' study, some cases combined the information desk with subject specialty. It means that basic questions are asked at the information desk and answered. But complex ones will be delivered to professionals

with the knowledge of certain subject. This model could effectively employ staff time and savvy, meanwhile it complicates the management of reference service. As such, how to deliver services with lower cost is still pending.

2.2.3.2 Balancing different communication modes.

According to Tyckoson (2011, pp. 587-590), there are generally four communication modes in libraries, which are fact-to-face mode, telephone mode, email mode, instant message mode. These modes simultaneously function in a library in order to make reference service attainable for patrons. For instance, at the reference desk, staff will not only answer basic questions through the fact-to-face mode, but also answer phone calls and provide guidance for people living in remote areas. Subject professionals will reply to queries through email, in the meantime, they need to write a response to some messages generated from instant chat software. Different communication modes need different resources to support, therefore, balancing these communication modes will be a challenge for library managers. With the development of technology, more options will be created. That is to say, such balancing tasks will continuously exist in libraries. In addition, the application of virtual communication media extends library working time from limited hours to all day long, which increases the need for staff to work in different periods. In the end, library costs will be increased. In Stevens' study (2013), the communication mode is extracted as two forms, physical communication mode and virtual communication mode. When delivering services through these two modes, libraries have to confront the difficulty of allocating resources. In the physical communication mode, some electronic resources would be required, which challenges the meaning and existence of the physical mode. For example, if electronic materials are requested a lot at the reference desk, the physical communication mode is not a good choice to deliver the material. When it comes to the virtual communication mode, abilities of librarians are highly required to ensure the whole process of the virtual communication. Simply put, physical resources are needed in the virtual communication mode and vice versa. Balancing these resources in these two modes requires the strategic views of library managers. And the difficulty of management within a library will be increased by such situations.

2.2.3.3 Lack of professional librarians.

One resource could be helpful to solve issues caused by aforementioned challenges: professional librarians. No matter which service model or communicating mode is used, human resources are greatly needed to accomplish these tasks. However, in many libraries, few staff are assigned for reference service. As is pinpointed by Standerfer (2006, p. 140), many libraries only have one or two employees to do reference service and in most cases they work part-time. As such, skills attained by librarians could be the key energy to ensure the whole reference service runs smoothly, which is realized by Standerfer (2006, pp. 143-144) and Tyckoson (2011, pp. 590-592). In their opinion, the main idea to handle this challenge is to launch education or training campaigns for librarians in order to improve their professional skills. Nevertheless, issues still appear. First of all, no matter how well librarians are trained, there is always a chance of them

being asked rather specific questions which they cannot answer. Secondly, with the development of knowledge and technology, training in the library should be a continuous procedure rather than one-time operation. Thereby, how to identify library skills or technique requirements and thus arrange corresponding courses might be a tough problem. Last but not least, even though a well-planned training campaign could be reached, lack of resources, time, places etc. will lead to the unsuccessful training results. Therefore, it is a challenge to effectively assign enough librarians and keep them professional and competent.

2.2.3.4 Lack of information resources.

Based on the definitions of reference service mentioned in the former section, information is the main source throughout the whole process of offering reference service. People contact libraries to ask for instruction or guidance and librarians reply with specific information. Then people use that information to finally achieve their goals. Thus, various sources of information could be valuable assets. According to Tyckoson (2011, pp. 592-594), one of the greatest reference service challenges is the access to information sources. In the 1990s, when Oberhauser (1991) discussed the interactive multimedia for library and information services, it was realized that owning copyrights and permissions to distribute materials is a challenging issue which hinders the development of library services. To own copyrights means costing money from libraries. Since money is limited, copyrights or information access owned by a library are limited as well. How to allocate money to purchase the most important access could be a challenge for libraries. Meanwhile, information needs from individuals are ever-changing. That is to say, updating information resources is another difficult task for libraries. On one hand, they would like to offer satisfactory reference service, on the other hand, they do not have sufficient assets to upgrade information content. All in all, libraries are in a situation where information sources cannot be approached substantially and library managers need solutions to maximize their current resources while maintaining the quality of reference service.

2.2.3.5 Hard to assess the outcome.

Money as a factor has been mentioned in all these discussed challenges. Simply put, money is the major reason for these challenges. Therefore, reference service need to be assessed in order to measure how much has been paid back considering the amount of spent money. The assessment methods have been evolved a lot. First of all, the amount of questions asked at the reference desk is calculated. But the result dose not disclose the quality of service. Then how many right questions are answered is employed as an indicator to evaluate reference service. However, not much information concerning service improvement could be realized. Therefore, a new measurement is created, which is called unobtrusive evaluation testing the correctness of reference service in diverse fields. In the end, low valid response rate makes such method problematic. This default can also be found in the Wisconsin-Ohio Reference Evaluation Program. Furthermore, user interaction is involved to assess the service with the advent of LibQUAL, which is a standardized questionnaire and widely used in different libraries.

Nonetheless, this method is not flexible. That is to say, new requirements, especially needs with latest popular topics could not be reflected through the result (Tyckoson, 2011, pp. 595-597). It is clear to summarize that there is no ubiquitous method for reference service evaluation, which aggregates the difficulty for assessment. According to the review of Kuruppu (2007, pp. 368-369), no universally agreed indicators (user satisfaction, efficiency of the service or quality or quantity of materials) can be used to test the service. Moreover, each indicator can be evaluated with various methods. This two conditions together make the assessment a tough task for libraries.

As is obviously demonstrated by the aforementioned challenges, even though reference service is not a new topic for library managers, to manage it is still a burdensome task. Effective solutions are greatly needed to conquer these challenges. Or at least some ideas of decreasing the negative impact generated from the challenges are also expected if there would be no way to avoid them.

2.2.4 Studies on improving reference service

Since challenges are confronted by libraries when providing reference service, studies are encouraged to help libraries survive conquering challenges. Weimer (2010) implements an innovative service in Alderman Library, University of Virginia. This service brings short message service (SMS) into the reference service system in Alderman Library. Scholars test how well SMS could function to provide reference service. In the end, continuous increment of usage in SMS is noticed and reference service are extended. This study highlights the significance to involve popular communication medium (in this case: mobile phone) to work for library reference service, which sets a good example for future researchers or librarians to pay attention on containing daily life resources to improve the service. Nunn and Ruane (2011) employ marketing theories to work for the improvement of reference service. They emphasize the importance to closely link users in order to manage issues caused by evolving user requirements, changing technology and increasing amount of students and long-distance citizens. Face-to-face communication is outlined in this study and librarians' social expertise is recognized as a key factor to enhance the user awareness of library service. Human resource plays a role in improving reference service in this study. Contrary to Nunn and Ruane's study, Aguilar et al. (2011) consider that face-to-face communication is out of date and they think highly of the virtual environment. They launch initiatives to provide reference service in a virtual environment, which, according to the result of the initiatives, are more similar with current ways through which users approach information. And in the end, positive relationships with users are established. Todorinova et al. (2011) notice that data-driven changes are happening in academic libraries and involving data management for general pattern recognition could be a path to improving reference service. Saunders (2013) suggests learning a lesson from past bad examples. Two pieces of advice are put forward to improve the service at reference desk: arranging staff training, highlighting the role of evaluation. Saunders also suggests that training and evaluation should be integrated with each other and aimed more at customers. As such, the reference desk can effectively function. Aggarwal and Powers (2013) encourage a shared service model to increase reference

service quality. This model was launched in the Career Education Cooperation and more access was reached with the help of this model.

These studies clearly present one key approach to improve reference service: to maximize resources owned by libraries, such as enhancing skills attained by librarians, utilizing data generated at reference desk or accessing materials through shared service models. Under the situation that limited money or other sources can be invested in libraries, making full use of resource around library could be a good idea for service improvement. In addition, focusing on trends in daily life and users' needs could also be an effective method.

3. The natural match between data mining and the library

Numerous data is being recorded in a library, for instance, the information about newly bought books, the borrowing record of individuals, the images of library daily visits etc. In addition, our daily lives are also confronting the situation of data explosion. "The match between librarians' deep knowledge base and expertise in the area of metadata specification and development makes libraries well-matched to undertake a long-term role in supporting data curation" (Gold, 2010). It might imply that to combine data mining with libraries will naturally be a new trend to conduct library service. This is because both librarians' working content and the responsibilities of libraries require data management as a major approach. As such, data mining could be a main component.

3.1 Data mining as knowledge

Regardless of all the requirements for libraries launching data mining activities, libraries, as the nexus of knowledge, have the responsibility to involve data mining and bring it to citizens while only considering data mining as one kind of knowledge. Based on this point, it is not awkward to relate data mining to libraries. Moreover, according to Keloğlu-İşler and Bayram (2014, p. 551), knowledge is the essence that connects individuals and society. Furthermore, society is confronting the tendency to highly rely on technologies and the distribution of information, knowledge which can be served as a technology and aims at dealing with information should be right to the point. Given such circumstance, data mining could be viewed as the knowledge well needed for people understanding the current society because data mining can disclose patterns hiding behind information. Libraries providing knowledge concerning data mining could easily play a bridge role to connect people and society. Furthermore, libraries are heading towards a digital environment, which indicates more virtual library interfaces for users. Gaining data mining skills to some extent would be helpful for users to understand the new environment. Therefore, it could be a wise idea for libraries to possess data mining knowledge.

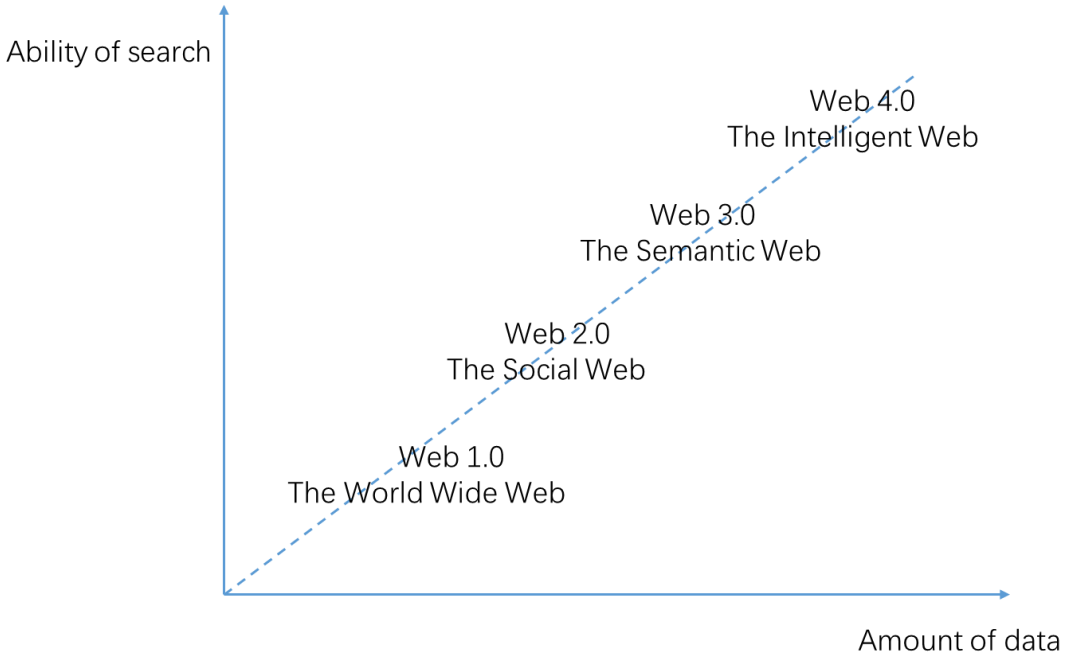
When introducing the new concept "Big Data", the idea that computers can gather numerous amounts of information and find values in such information to librarians, Hoy (2014, pp. 322-324) indicates three key ways for libraries to involve Big Data, which are providing guidance and materials for users owing to the their enhancing interest in Big Data, cooperating with other institutions to improve research in the corresponding field, and helping patrons understand what Big Data can do and cannot do. These three points are served to support the argument of Hoy that libraries are suitable for working with Big Data. In this case, Big Data is merely considered as knowledge. Since libraries generally share the mission to enhance the quality of life as a center of knowledge and learning (Heidorn, 2011, p. 662), there is no excuse for libraries not being involved with data mining. Therefore, libraries and data mining are naturally connected from the knowledge point of view.

3.2 Going for data mining: an easy choice for libraries

The library and information sciences have been discussing the issues of technology changing for a long time, among which the evolution from web 1.0 to web 4.0 has been mainly emphasized. Such evolution leads to the change of library features and the change is presented by the term library 1.0, library 2.0, library 3.0 and library 4.0. Noh (2015) launches a study to shed light on the future model in the context of library 4.0 and in this study, research papers about web 4.0 or library 4.0 are analyzed. Key advancements as well as key features of web 4.0 and key words of library 4.0 compose the main result of this study.

According to Noh, with the increment of data volume and the development of web search ability, the web has evolved as is shown in Figure 5:

Figure 5: The development of Web



Source: Figure 5 is developed from the Figure 2 in the study of Noh (2015, p. 789)

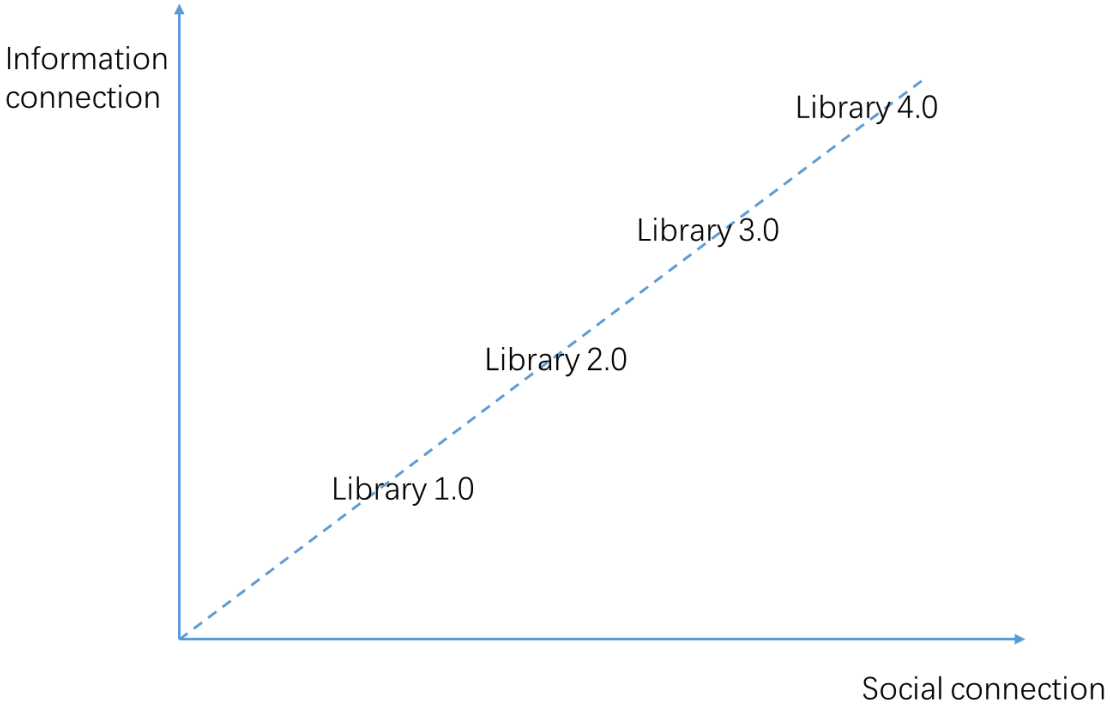
Figure 5 clearly shows that the key feature of web will be altered with different data amount and search ability. According to Noh (pp. 788-791)(pp. 788-791)(786-797), in the context of web 1.0, the main activities were information search and information consumption. That is to say, the usage of information is rather basic in web 1.0. When it comes to web 2.0, the information behavior of users becomes more diverse. Apart from information seeking, information creation, sharing, storage and evaluation are allowed in web 2.0. Simply put, user participation is realized in web 2.0. That is the main reason why web 2.0 is called the social web. Nevertheless, web 2.0 still lacks the capability to understand the meaning of information. Therefore, the advent of web 3.0 emerges to fill in this gap, to enhance the connection between knowledge and data, and

thus improve the meaning of information. In web 3.0, the communication is still ensured. However, the communication is not only human centered, but also concentrated on the conversation between human beings and machines. As for web 4.0, all the good characteristics generated from previous web versions are still kept. In addition, the analyzing function is highlighted in web 4.0. Thereby, new ideas or theories can be created through information analysis. Noh (p. 791)(p. 791)(786-797) names web 4.0 as Intelligent Web, which makes inference search possible. Furthermore, more decisions can be made in web 4.0 based on the over time record about what we want and how we live. In this context, the historical record about our life is data stored in the system. Hence decisions cannot be generated without the help of data mining.

It, as well, can be implied from Figure 5 that the increasing amount of data supports the enhanced ability of information searching and the improved searching ability requires greater volume of data. In this bidirectional interaction, the volume of data functions as the main factor. Or put it in another way, it is obvious that data and the meaning behind data have started to play a vital role in the development of the Internet. In web 1.0, data is merely considered as the source to locate information. The meaning of data or information in itself is not valued. Then with the advent of web 2.0, new data is generated during the social communication. Additionally, the link of data is noticed. In web 3.0, the understanding of data is emphasized and a knowledge connection is established. When it comes to web 4.0, the understanding of data is going further. As such, data mining shows a tendency to be more and more significant. This is because the core of data mining is analyzing data, to produce potential value hiding behind data. Since web 4.0 is more associated with intelligence, data mining as one of the tools for intelligence generation will be suitable in the meantime. All in all, the development of the web discloses the evolving utilization of data.

According to the study of Noh(pp. 791-796)(pp. 791-796)(786-797), the evolution of the web leads to the evolution of the library model, as shown in Figure 6:

Figure 6: The development of library model



Source: Figure 6 is developed from the Figure 5 in the study of Noh (2015, p. 795)

The social connection and information connection are considered as two main factors to decide the library model. The social connection reflects the development of technology and the information connection represents the extent of information understanding. Library 1.0 is associated with the application of web 1.0, the rest is connected in the same manner. Library 1.0 delegates the library which is operated in the context of low technology and basic information understanding. In Noh's study, the involvement of web 4.0 gives birth to the existence of library 4.0, which is an organic system and has the characteristic of web 4.0. After reviewing previous studies, Noh summarizes the key word of library 4.0 as: intelligent library, massive data library, augmented reality library, context aware library, cutting-edge recognition library and infinite creative space. As is discussed above, web 4.0 shows clearly the color of data mining. Therefore, when introducing library 4.0, data mining can also be easily noticed.

Intelligent library outlines the analysis of information and the usage of analysis result (for decision making or service developing). During this process, data mining techniques are indispensable. Massive data library emphasizes realizing the value of numerous library data. Data mining definitely plays a role here. Content aware services are applied in library fields such as book status information, book content information, personal library management service, internal library information etc. (Lee, 2013). Such service is developed from Big Data, hence it could be implied that the essence of content aware services stems from the analysis of data. Therefore, the key word, content aware library, is data mining related as well.

It can be summarized that data mining can be naturally approached in the library through indicating the relationship between data mining and library 4.0, especially in libraries of the future library. Moreover, data mining will be a necessary technique in the context of web 4.0. Therefore, it could be implied that to go for data mining would be an easy way for libraries in order to survive in the long run.

3.3 Libraries shifting from data poor to data rich

As is cited by Gordon-Murnane (2012), data is generated at an extraordinary speed and in a great volume by businesses, industries, universities, hospitals and individuals. The scale of data has been revolutionized to be measured by exabyte, which equals to 4000 times the information volume stored in the US Library of Congress. Therefore, it is quite obvious that current libraries are confronting a data explosion situation. According to the study of Heidorn (2011, pp. 663-664) and Gordon-Murnane (2012, p. 30), there are three main reasons leading to the richness of library data:

First of all, the worldwide availability, affordability and applicability of digital devices make access to the Internet rather easy. In Gordon-Murnane's opinion, every time of reaching to the Internet indicates the reality of data creation. Considering numerous smartphones, tablets, computers and laptops connecting the Internet, millions of pieces of data can be generated within just one second. Not to mention that the utilization of such digital devices has become a lifestyle. Currently, the amount of data created within one day or even one hour can be massive. Libraries, nowadays, are highly engaged in the Internet and Apps related to library services, such as booking information checking or renewing borrowing status are no longer a hype word. They do exit in many libraries. Thus, all these applications can be counted as the main power for enriching library data.

Secondly, the types of digital sources are increasing. People are generating data through emails, social media softwares, website browsing and instant chatting platforms. These data on one hand lead to the dramatic increment of data volume. On the other hand, they are in different forms, such as texts, numbers, images or even videos. Moreover, different forms of data are generated by different activities. Hence, the richness of library data is not only reflected in data volume, but also in data types.
Last but not least, the advanced technology in data collecting, recording, analyzing and aggregating plays an important role in enlarging the volume of library data. In light of Heidorn (2011), the digital storage evolves from notebooks to electronic devices and introduces greater amount of data for the utilization of librarians and scientists. Meanwhile, the change in computation and telecommunication has also caused a huge creation of data to the extent which is beyond the imagination of previous scholars. For example, the borrowing history of library patrons was not recorded effectively. However, with the current techniques, such history can be easily stored in the library system. When it increases to a large amount, the history can be used as a valuable asset for libraries to identify user requirements. The book recommendation mechanism of Amazon would be a perfect case to explicate how useful the borrowing history could be when the amount of the history reaches a high level.

According to these reasons, it would be easier to conclude that more data can be generated in libraries in the near future with the development of technology. As is pinpointed by Wittmann and Reinhalter (2014, p. 368), our libraries are entering into a data lifecycle. Three implications can be made based on this new library lifecycle with the angel of data mining:

From the source point of view, the emphasis of library material should be revolutionized from books to data generated in the library system. It does not mean that books are no longer important. On the contrary, they are still the basis for library service. Nevertheless, data around the library should be valued in the meantime. Wisely used, service gaps can be filled with the help of data. Wittmann and Reinhalter (2014) support this idea and they advocate libraries to be fitted in the data-fueled tide and provide data literacy. Data mining would be the suitable approach for wisely utilizing data as it has the feature to extract patterns behind data. In a word, data, as a main source in libraries, should be highly considered.

From the service point of view, services related to data should be established in libraries (Hoy, 2014, Wittmann and Reinhalter, 2014). At this point, services concerning data do not concentrate on data per se. They are more intended to guide individuals or organizations to benefit from data. For instance, Hoy (2014) put forward that libraries should educate citizens to understand the pros and cons of Big Data in order to help them achieve their own goals with Big Data analytics. It can be indicated that the instruction of data utilization could be the center of data related services. As such, the practical application of data could be a good education or instruction for library patrons. Since data mining has advantages in knowledge creating and decision making assisting, personal skills in data mining could be a good direction for libraries to pursue in developing data related services.

From the librarian skill point of view, this data lifecycle requires librarians to attain new skills. The two aforementioned implications also express the new requirement for librarians in such data intensive situations. Data curation is one of the most discussed skills of librarians (Gordon-Murnane, 2012, Heidorn, 2011) Since libraries have stored lots of data in the system, it is necessary to effectively curate it over a long period of time. In addition, excellent data sharing skills are also indispensable to librarians, because data is becoming richer and richer in libraries. There is no need for librarians to be aware of all kinds of data, nor is it possible. Thus, good data sharing networks could be a feasible solution to supply librarians' insufficient dataset knowledge. Data sharing skills are not only useful for the cooperation among librarians, they are also helpful for serving library users. Therefore, in the context of data explosion, data sharing skills are rather important for librarians to attain. It is also suitable for librarians to undertake the role of boosting data sharing (Gordon-Murnane, 2012, p. 34) Apart from these two skills, how to put data into practical use is foremost as well. In Gordon-murnane's opinion, it is a natural fit for librarians to learn how to utilize data and help individuals, business leaders, university lecturers and governmental staff to make better decisions. One of the common ways to realize this task is to analyze

available data. Therefore, it is not difficult to introduce data mining skills at this point. In a word, in order to make progress with data, librarians should know some data mining skills so as to discover more possibilities, which can be used for serving individual decision making process.

In summary, broad usage of digital devices, increasing types of digital sources and highly developed technology make libraries confront a data rich environment. As such, libraries' opinion on materials, services and staff skills should be changed correspondingly as well. Data mining can be naturally noticed and adjusted into this changing process because of its unique feature in employing data for creating more values. The shift from data poor to data rich naturally makes libraries involve data mining to a higher level.

3.4 Data mining: a wise method to confront reference service challenges

As is displayed above, there are five main challenges confronting libraries to deliver good reference services, which are offering a suitable service model, balancing different communication models, lack of professional librarians, lack of information resources and hard to assess service outcomes. Studies have been carried out to improve reference services, but it is rarely discussed how to decrease the negative impact of these challenges. Furthermore, as libraries are entering into a data rich environment, new insights should be created to resolve the reference service issues. As such, handling the problematic challenges with the angle of employing data could be a wise direction for libraries to head towards. Based on the nature and benefits, data mining could be a feasible tool to assist libraries.

First of all, when deciding which service models could be the most suitable, libraries put applicable resources into the core of the consideration. Regardless of the scale of the library, the model with investing the least of resources would be the foremost in the library managers' mind. Behr and Hill (2012) notice that wisely mining data generated from e-Reserves, the book reservation system in libraries, long distance users will be well served. In this research, the library in Central Michigan University is selected as the experimental unit, citation information about items reserved in the system is mined. In the end, a clear collection size, and a clear understanding about the utilization status for current materials are generated. The most needed books for long distance users are highlighted, which provides good hints for librarians to prepare hardcopies or scanned files for long distance users. In this case, the experimental library only makes use of on-hand resources, which indicates not much extra money is invested. Eventually, suggestions for establishing a service model for long distance users are purposed. This is a convincing example to illustrate that the result of data mining has a positive influence on deciding the reference service model.
When it comes the challenge of balancing communication models, data mining could also be a reasonable solution. As data mining has the target to realize patterns of information, librarians working in different communication models could mine all the questions asked in different models. In the end, a content of questions asked in models can be created. According to the content, which questions are mostly asked during face-

to-face communication, which is popular through instant message can be highlighted. Then, based on the specialty of librarians and the content of questions, librarians could be properly assigned to work in the corresponding communication model. Even though not every library has clear staff assignments in various models and few studies have been conducted to empirically prove the applicability of data mining in handling this challenge, it is still a key part to understand the nature and context of data to delivery library services (Ogier et al., 2014). Therefore, mining the content of the asked questions might be one of the pragmatic cases.

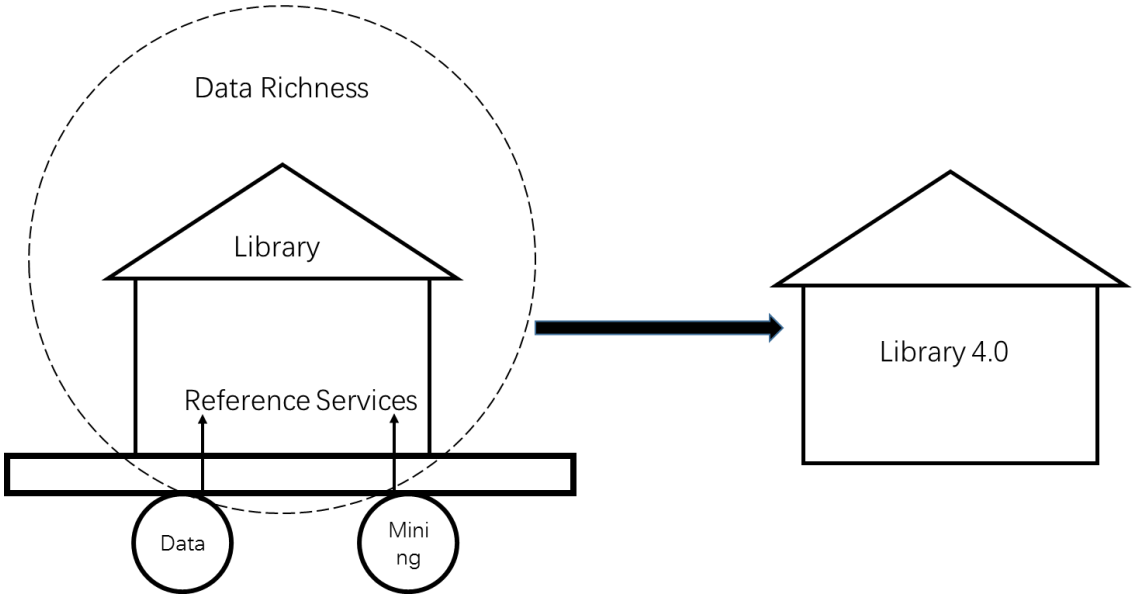
Since society is developing at a rapid speed, it is a challenging task for librarians to keep pace with the speed. Put in another way, the skill of librarians is hardly enough to support the library operation. As for the provision of reference services, the lack of professional librarians is one of the biggest challenges. Nevertheless, scholars find an effective approach to confront this challenge. Xia and Wang (2014) target all the job vacancies in libraries posted on the website IASSIST since 2005, and they collect job titles, qualifications, professional preparation and responsibilities into their datasets for further analysis. In the end, the categories of employers, library units, job titles, degree requirements etc. are generated. With the help of such categories, library directors would gain a better picture of librarian competencies in the industry. In addition, these advertisements are mined chronologically. That is to say, the key word for librarian jobs in different years can be obviously noticed, which reflects the evolving skill requirement in libraries. Even though data mining cannot directly resolve the lack of professional skills, the result of it can provide a good preparation for library human resource development strategy. For example, the result of mining job ads explicates the newly needed skills in the near future, such as data curation or data analysis. Thus, library directors will take these skills into consideration when recruiting employees. It will help libraries hire the right employees with the most useful and updated knowledge. Then the status of lacking professional librarians could be relieved. In a word, data mining indirectly assists libraries in the circumstance of having less professionals.

Data mining will have a positive influence in the similar way whilst handling the issues caused by lack of information resources. Since libraries have limited materials to provide, it will be a problem to overcome the situation of limited information resources. As such, how to effectively utilize current resources to serve as many as possible patrons could be a smart idea. Thus, data mining can make a difference here. With the help of mining user queries and material content, the relationship between user requirements and library materials can be established. Therefore, when a user is asking for something new or without any access, librarians can firstly provide the most relevant items based on his query. Then the lack of information resources could be conquered in light of deeply and effectively employing every library resources.

All in all, four of reference service challenges could be wisely conquered with the application of data mining. Even though the challenge cannot be solved totally, the issues caused by them could be decreased through employing data mining. Considering the source of data mining in libraries is the already existing data, it might also be a cost

reduction method to improve reference services from the perspective of data mining. Therefore, it can be concluded that choosing data mining as a method to enhance library reference services could be a natural choice for library directors.

Figure 7: Illustrating the natural match between data mining and the library



As is discussed in section three, data mining and the library are naturally connected with each other. First of all, libraries are facing a situation in which they are surrounded by more and more data. Therefore, it could be feasible and sensible for libraries to shed light on the usage of data. Then data mining could be one of the choices to use data. In addition, libraries are heading towards the generation of Library 4.0 with the development of web 4.0. During this process, data mining could function as a car for libraries to move to the Library 4.0 direction because intelligence creation and information analysis are emphasized in this new library version. If the library has gained the techniques to conduct data mining already, it would be easier and faster for the library to reach the new version. Thus, data mining could just be like a car to carry the library moving forward smoothly. Meanwhile, data mining has a positive impact on reference services, especially handling the challenges rooted in reference services for long periods of time. In this process, data mining is mainly viewed as one kind of knowledge and libraries function as the entity to disseminate knowledge. Additionally, data mining makes difference through creating knowledge as well.

In conclusion, data mining and libraries closely match with each other in light of the development of technology, the evolution of libraries, the mission of libraries and the advent of library data richness.

4. Methodology

Since the goal of this study is to find useful online free databases and provide pragmatic data mining applications for library reference service improvement, thus to outline the feasibility of libraries utilizing data mining, thereby case study could be an effective method. Firstly, the options of data sets within the library is various and the resources of online free data are even more diverse. Therefore, it could be a wise and effective idea to narrow down the list. A few data sources are chosen for mining, which can be set as the example to demonstrate the feasibility of combining data mining with reference service enhancement. Secondly, this study is an exploratory research. There are few prior theoretical notions that can be started up with. Few theories or ideas could be generated until the case is approached and studied (Gillham, 2000, p. 2), which is considered as a feature of case study. This feature corresponds with the current study well. Therefore, case study is chosen as the main method.

In order to explore the potential of data and data mining in improving library reference service, Turku Main Library, the biggest public library in Turku region, is chosen as the case. New services are offered through modern technology at the Library, which makes it easier to use the library independently. As one of the busiest libraries in Finland, Turku Main Library provides its customers with a wide and varied collection of library materials in both printed and electronic form. The collections are complemented by the various exhibitions and events organized every day at the library (Librarybuildings.Info). Thus, the study of such busy library could be representative in establishing theories of reference service improvement. As a municipal service provider, the public library undertakes the role to serve various customers, which, to some extent, toughens the difficulty to accomplish the job. Hence to chase citizen requirements down and to satisfy them are usually the priority of library tasks. After the communication with a project leader in Turku City Library, he mentioned that to know library patrons is “always the No. 1 task”. Therefore, the dataset of daily visits was selected. The motivation to choose this dataset is: according to Nicholson, to better understand library user communities is the key task to tailor services to meet user needs (Nicholson, 2003b). Since this dataset records the daily visiting situation, it is highly related to users thus directly reflected certain user requirements and user behavior. A minor consideration to choose such data set is that it records daily situation, which means the amount of data can be assured at a large volume.

In order to cooperate with daily visits, various data resources were viewed. In the end, the climatic condition data on the website, Weather Underground, were collected. The motivations to choose such data are: first of all, the climatic condition is daily data as well, which could be easily related to library daily visits; secondly, the climatic condition will influence the transportation, thus the willingness of citizens to go to the library in person might be affected too. Therefore, a pattern could be discovered after considering the weather condition with library daily visits. Moreover, there is no study concentrating on the relationship between climatic condition and library daily visits, hence such combination would add creativity into the presented work.

After data analysis, the result was interpreted from the reference service point of view. And implications of enhancing service were created. Then the result and the implications were presented to librarians. The presentation was performed in the interview with librarians, the interview was constructed and each interview lasted around 5 to 10 minutes. Owing to the lack of theories, there is no feasible way to test whether data mining could be helpful to improve the reference service. Therefore, the opinions from librarians are valuable information for creating a decent judgement. Simply put, the interview with librarians could be considered as the evaluation of the result in data mining process. Based on the opinions, the feasibility of data mining in improving reference service could be demonstrated and the research question of the thesis work can be answered.

4.1 Data collection

The daily visits were downloaded from the former website of Turku Main Library. The time period is from 1st January 2009 to 23rd of June 2015. (Because the websites of Finnish public libraries are under reconstruction, data concerning 24th of June 2015 onwards was not available.) As such, 2365 pieces of data were collected. In this dataset, date, seven days in week and daily visits were recorded.
The information of weather condition in the same time period as daily visits was collected. Actual daily average temperature, average humidity and weather conditions (clear, fog, mostly cloudy, overcast, partly cloudy, rain, scattered cloudy, snow, thunderstorm with hail and Tstorm) which are categorized on the web site were collected.

Additionally, in order to ensure the accuracy, seasonal information was also collected. The breaking points of different seasons in each year were decided based on the information on the website: timeanddate.com, as is shown in Table 2:

Table 2: The breaking point of seasons in Turku from 2009 to 2014

Year	Season	Date
2009	Winter	1.1 – 19.3, 21 – 31.12
	Spring	20.3 – 20.6
	Summer	21.6 – 22.9
	Autumn	23.9 – 20.12
2010	Winter	1.1 – 19.3, 22 – 31.12
	Spring	20.3 – 20.6
	Summer	21.6 – 22.9
	Autumn	23.9 – 21.12
2011	Winter	1.1 – 20.3, 22 – 31.12
	Spring	21.3 – 20.6
	Summer	21.6 – 22.9
	Autumn	23.9 – 21.12
2012	Winter	1.1 – 19.3, 21 – 31.12
	Spring	20.3 – 20.6
	Summer	21.6 – 21.9
	Autumn	22.9 – 20.12
2013	Winter	1.1 – 19.3, 21 – 31.12
	Spring	20.3 – 20.6
	Summer	21.6 – 21.9
	Autumn	22.9 – 20.12
2014	Winter	1.1 – 19.3, 22 – 31.12
	Spring	20.3 – 20.6
	Summer	21.6 – 22.9
	Autumn	23.9 – 21.12

The data collection process on the aforementioned sources was conducted manually. All of data were recorded in Excel 2010 and analyzed with SPSS 17.0.

4.2 Data cleansing

The daily visits of 187 days were zero owing to the close of the library in holidays or data deficiency. Therefore, they were deleted from the data set. Then, 2178 pieces of data were stored at this stage.

Since daily average temperature and daily average humidity were collected, the Discomfort Index (DI) could be calculated based on the formula posted on the website: Keisan Online Calculator. The reason to choose this website was because the information on the website is totally free and cited by some scholars.

$$\text{Formula 1:DI} = \mathrm{T} - 0.55(1 - 0.01\mathrm{H})\mathrm{x}(\mathrm{T} - 14.5)$$

T: air temperature ℃, H: relative humidity % (KeisanOnlineCalculator)

The motivation to compute DI was that studies have considered DI as a main factor to discuss individual behaviors (Liu et al., 2015, Mazon, 2014). Since the purpose of the presented study was to mine user patterns thus to make progress in library reference service, DI could also be an appropriate factor for this case study. As is cited by KeisanOnlineCalculator , there are six conditions of discomfort:

Table 3: Various discomfort conditions

DI(℃)	**Discomfort Condition**
～21	No discomfort
21～24	Under 50% population feels discomfort
24～27	Most 50% population feels discomfort
27～29	Most population feels discomfort
29～32	Everyone feels severe stress

According to Formula 2 and Table 3, DI of Turku within selected period was calculated. Out of 2178 cases, only 25 cases were classified into conditions other than No discomfort. Or put in another way, more than 98% cases belong to “No discomfort”. Therefore, these 25 cases were deleted from the dataset as outliers since the number of them is much less than that of No discomfort. Finally, 2153 cases were stored.
As is shown in Table 4, there are eleven types of weather conditions recorded on Weather Underground regarding Turku in the last 6 years.

Table 4: Weather condition types and frequencies in Turku from 2009.01.01-2015.06.23

Weather condition	Frequency
thunderstorms with hail	1
overcast	26
mostly cloudy	32
Tstorm	66
clear	91
scattered clouds	213
fog	233
partly cloudy	251
snow	643
rain	809

Since the weather condition marked as "thunderstorms with hail" was only in one case, therefore this case was deleted from the dataset as an outlier as well.
Last but not least, further investigation on the outliers in the dataset was operated. Because some outliers could not be easily realized through physical observation as shown in previous steps. In this case, the function: standardized values as variables was used and the values which were greater than the absolute value of 2 were deleted from the dataset (Field, 2009). Thus, 111 cases were deleted. In the end, the data cleansing process was completed and 2041 cases were eventually mined in the following process.

4.3 Data preprocessing

The analysis was adjusted from a library point of view and it was decided to use basic statistical methods to mine collected data. A multiple regression analysis was used. In this case, hourly visits were considered as the dependent variable. Because Turku Main Library has different opening hours on weekdays (11 hours) and weekends (6 hours). The hourly visits were calculated by Formula 2:

$$\text{F2: Hourly visites} = \frac{\text{Daily visits}}{\text{Opening hours}}$$

Seven days in a week, weather conditions and discomfort index were considered as independent variables. The influence on these independent variables on hourly visits was explored and a regression model was established in different seasons respectively. Since independent variables include two categorical variables, dummy variables should be generated to replace the categorical variables (Suits, 1984). In this case, it could be effective to pre-test the correlation between the dependent variable and two categorical independent variables respectively. If there is no significant relationship between the dependent variable and each of the categorical variable respectively, the categorical variable cannot be included in the regression model. Thus, it is no need to create dummy variables for the categorical one. Meanwhile, the efficiency of the data

mining process could be increased. Therefore, the relationship between the dependent variable and each of the categorical variable were examined. Before the examination, the number of groups within one categorical variable was deducted from the simplifying the data mining process point of view.

In this case, seven days in a week have clear division between groups. Therefore, they were not conducted for group number deduction. The groups under weather conditions were merged in a qualitative way. Tstorm and rain were merged into a new group "rainy" according to the meaning of these two words. Mostly cloudy, scattered clouds and partly cloudy were merged to the new group "cloudy" after reviewing the images found in Google Image with these three terms as key words.

Table 5: Weather condition after merging groups

Weather condition	**Frequency**
clear	91
cloudy	496
overcast	26
rainy	875
fog	233
snow	643

In order to utilize the suitable method to pretest the relationship between hourly visits and weather condition and seven days in a week respectively, the normality of hourly visits was checked. As is presented in Table 6:

Table 6: The result of dependent normality test

	Kolmogorov-Smirnov[a]			**Shapiro-Wilk**		
	Statistic	**df**	**Sig.**	**Statistic**	**df**	**Sig.**
Hourly visits	,067	2152	,000	,941	2152	,000

a. Lilliefors Significance Correction

P-value is 0, which rejects the null hypothesis and it can be concluded that the distribution of the dependent variable is not normal. Therefore, the command: nonparametric test was employed to explore the relationship. In this case, both categorical variables have more than two categories, therefore the function K Independent Samples was chosen (Field, 2009). The result of the test is shown in Table 7:

Table 7: The summary of nonparametric tests

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of hourly visits is the same across seven days in a week	Kruskal-Wallis Test	0	Reject the null hypothesis
2	The distribution of hourly visits is the same across categories of weather conditions	Kruskal-Wallis Test	0	Reject the null hypothesis

Asymptotic significances are presented and the significant level is 0.05.

According to Table 7, it could be easily concluded that hourly visits change greatly in different groups of these categorical variables. Such result could be implied that certain groups in the categorical variable might have a strong influence on hourly visits. Therefore, both categorical variables should be introduced into the regression model and dummy variables of them must be made.

Dummy variables were created with the function "recode into different variables" in SPSS under the command "Transform". This process was followed in the instruction of Field (2013, pp. 253-257), and the dummy coding result was listed in Table 8 and Table 9. It can be concluded that Sunday and snow were chosen as the baseline category.

Table 8: Dummy coding for seven days in week

	Dummy Variable: Mon	Dummy Variable: Tus	Dummy Variable: Wed	Dummy Variable: Thu	Dummy Variable: Fri	Dummy Variable: Sat
Monday	1	0	0	0	0	0
Tuesday	0	1	0	0	0	0
Wednesday	0	0	1	0	0	0
Thursday	0	0	0	1	0	0
Friday	0	0	0	0	1	0
Saturday	0	0	0	0	0	1
Sunday	0	0	0	0	0	0

Table 9: Dummy coding for weather conditions

	Dummy Variable: clear	Dummy Variable: cloudy	Dummy Variable: fog	Dummy Variable: overcast	Dummy Variable: rain
clear	1	0	0	0	0
cloudy	0	1	0	0	0
fog	0	0	1	0	0
overcast	0	0	0	1	0
rain	0	0	0	0	1
snow	0	0	0	0	0

5. Result of Data mining

The data mining process is composed of three parts. First of all, the visiting situations in different seasons are discussed and four regression models are put forward. Secondly, visiting situations are compared in different seasons based on the result in the first stage. Then, the classifications of discomfort index in the library environment are generated.

5.1 The visiting situation in winter

According to Table 10, 548 cases were explored. Pearson's correlation efficient expresses the relationship between two variables in the model. Therefore, it can be summarized that among all the relationships, the relationship between discomfort index and hourly visits are the strongest (r=0.346). Four other variables' (Mon, Fri, Sat and Clear) strongest relationships are with hourly visits. There are no values greater than 0.9, which indicates that there is no multicollinearity between independent variables. The value of one-tail test shows that there are nine variables show significant relationship with hourly visits and all these results are contributed by the whole 548 cases, which means no cases are left out in the analysis.

Table 10: The descriptive statistics and correlation between variables in winter

		Mean	Std. De	N	HV	DI	Mon	Teu	Wed	Thu	Fri	Sat	CLE	CLO	Fog	Rain	OV
Pearson Correlation	**HV**	377.36	42.16	548		**0.35**	0.25	0.10	0.03	-0.10	-0.21	0.28	0.13	0.12	0.05	0.09	0.06
	DI	-3.16	5.57	548	**,346****		-0.02	0.03	0.01	-0.02	0.01	-0.02	0.08	0.15	0.05	0.32	0.10
	Mon	0.15	0.36	548	**,248****	-0.02		-0.17	-0.17	-0.17	-0.18	-0.17	-0.01	0.00	-0.02	0.02	-0.06
	Teu	0.14	0.35	548	0.095	0.03	**-0.17**		-0.16	-0.17	-0.17	-0.16	0.05	-0.06	0.11	0.01	-0.06
	Wed	0.14	0.35	548	0.029	0.01	**-0.17**	-0.16		-0.17	-0.17	-0.16	0.02	0.08	0.01	-0.06	0.06
	Thu	0.14	0.35	548	-.104**	-0.02	**-0.17**	-0.17	-0.17		-0.17	-0.17	-0.03	0.02	0.01	-0.04	0.02
	Fri	0.15	0.36	548	**-,211****	0.01	-0.18	-0.17	-0.17	-0.17		-0.17	0.02	-0.02	-0.04	0.02	0.02
	Sat	0.14	0.35	548	**0.278****	-0.02	-0.17	-0.16	-0.16	-0.17	-0.17		-0.03	-0.04	-0.04	0.05	0.02
	CLE	0.04	0.20	548	**0.132****	0.08	-0.01	0.02	-0.03	-0.03	0.02	-0.03		-0.07	-0.05	-0.06	-0.03
	CLO	0.10	0.30	548	0.124**	**0.15**	0.00	0.08	0.02	0.02	-0.02	-0.04	-0.07		-0.07	-0.09	-0.05
	Fog	0.05	0.21	548	0.045	0.05	-0.02	0.01	0.01	0.01	-0.04	-0.04	-0.05	-0.07		**-0.06**	-0.03
	Rai	0.07	0.26	548	0.086	**0.32**	0.02	-0.06	-0.04	-0.04	0.02	0.05	-0.06	-0.09	-0.06		-0.04
	OC	0.02	0.13	548	0.062	**0.10**	-0.06	0.06	0.02	0.02	0.02	0.02	-0.03	-0.05	-0.03	-0.04	

**p <0,01

Table 11: The regression model summary in winter

Model	Variables Entered	R Square	Adjusted R Square	R Square Change	F	Sig
1	Discomfort Index	.012	.0118	.120	74.206	0.00
2	Wed, Fri, Sat, Thu, Mon, Tues	.400	.392	.281	51.487	0.00
3	Clear, Fog, Cloudy, Rain，OC	.425	.412	.025	32.950	0.00

According to Table 11, there are three models established during the regression analysis. First of all, only Discomfort Index is considered as the independent variable. The value of adjusted R square is 11.8%, which means only 11.8% variance of the dependent could be explained by the independent based on the model. However, when seven days in a week are introduced to the model, the value of adjusted R square is increased to 39.2%, which accounts for almost 40% of the changes in the dependent variables. After considering the last variable: weather conditions, the adjusted R square goes up to 41.2%. That is to say, the variance in hourly visits could be explained 41.2% considering Discomfort Index, seven days in a week and weather conditions. Model 3 works well in this situation. In addition, among these three independent variables, seven days in a week have more influence to predict hourly visits. Meanwhile, the p values in these three models are all zero, which imply that the models significantly fit the overall data. Since Model 3 is the best model based on the value of adjusted R square, therefore only the coefficients of this model is listed below:

Table 12: The coefficients of regression model in winter

	B	Std. Error	Beta	Sig.
(Constant)	345.996	3.997		0.000
DI	2.485	0.273	0.329	0.000
Mon	62.200	5.169	0.527	0.000
Tues	45.318	5.280	0.374	0.000
Wed	37.428	5.265	0.309	0.000
Thu	26.259	5.220	0.219	0.000
Fri	14.781	5.183	0.125	0.005
Sat	66.887	5.254	0.552	0.000
Clear	25.333	7.012	0.121	0.000
CLO	14.759	4.872	0.104	0.003
Fog	5.933	6.752	0.029	0.380
Rain	-1.793	5.708	-0.011	0.754
Overcast	15.414	10.510	0.049	0.143

There are nine variables that are significant when p<0.01 and these variables are in bold in Table 12. A model can be defined as:

Model Winter:

$$Hourly\ visits = 346 + 2{,}49DI + 62{,}2Mon + 45{,}32Tues + 37{,}43Wed + 26{,}26Thu + 14{,}78Fri + 66{,}89Sat + 25{,}33clear + 14{,}76cloudy + 5{,}93fog - 1{,}79rain + 15{,}41overcast$$

Since dummy variables are included in the model, therefore the interpretation of the model should be noticed that each dummy variable means the difference between the group and the baseline group. For example, the B value of Mon is 62,2, which means that hourly visits would increase if the day changes from Sunday to Monday. As the p value is significant, it can be concluded that such a change is dramatically obvious. The beta value also tells that within one week, hourly visits in Saturday are the most because the beta value of Sat is the highest compared with the others'. In the same manner, the B value of rain can be interpreted that hourly visits would decrease if the weather changes from snowy to rainy, however such change is not significant. In other words, the change in hourly visits cannot be predicted through whether the weather is rainy compared with if it is snowy. But compared with sunny days, snowy days have less hourly visits and such a comparison is significant.

5.2 The visiting situation in spring

As is shown in Table 13, overcase is deleted from the analysis owing to the fact that there is no overcast weather in sping. There were 551 cases being explored at this stage. There is only one varaible (Mon) whose strongest relationship is with hourly visits. The strongest relationship in the table is between cloudy and rain. In addition, the

relationships between different days in a week is also strong even though they are less than 20%. All weather conditions show no significant relationship with hourly visits in spring.

Table 13: The descriptive statistics and correlation between variables in spring

		Mean	Std. De	N	HV	DI	Mon	Teu	Wed	Fri	Sat	CLE	CLO	Fog	Rain
Pearson Correlation	**HV**	387.83	47.63	551		-0.13	**0.26**	0.11	0.04	-0.14	0.14	-0.02	-0.01	0.07	-0.03
	DI	8.42	5.22	551	**-,129***		0.01	0.02	0.01	0.03	0.01	0.01	0.11	0.04	0.16
	Mon	0.15	0.36	551	**0,26****	0.01		-0.18	-0.19	-0.17	-0.17	-0.04	0.02	0.06	-0.06
	Teu	0.15	0.36	551	0,107*	0.02	-0.18		**-0.19**	-0.17	-0.17	-0.02	0.01	-0.07	0.04
	Wed	0.16	0.37	551	0.037	0.01	-0.19	**-0.19**		-0.18	-0.18	-0.03	0.05	0.08	-0.08
	Thu	0.15	0.35	551	-0.098	-0.02	-0.17	-0.18	**-0.18**	-0.16	-0.17	0.03	0.04	0.01	-0.03
	Fri	0.14	0.34	551	-.135*	0.03	-0.17	-0.17	**-0.18**		-0.16	0.04	-0.08	-0.04	0.11
	Sat	0.14	0.35	551	0,135*	0.01	-0.17	-0.17	**-0.18**	-0.16		0.03	-0.05	-0.05	0.08
	CLE	0.06	0.23	551	**-.017***	0.01	-0.04	-0.02	-0.03	0.04	0.03		-0.17	-0.07	-0.22
	CLO	0.33	0.47	551	-0.009	0.11	0.02	0.01	0.05	-0.09	-0.05	-0.17		-0.21	**-0.61**
	Fog	0.08	0.27	551	0.07	0.04	0.06	-0.07	0.08	-0.04	-0.05	-0.07	-0.21		**-0.26**
	Rain	0.44	0.50	551	-0.027	0.16	-0.06	0.04	-0.08	0.11	0.08	-0.22	**-0.61**	-0.26	

**p <0,01, *p<0,05

Table 14: The regression model summary in spring

Model	Variables Entered	R Square	Adjusted R Square	R Square Change	F	Sig
1	Discomfort Index	.017	.015	.017	9.27	0.00
2	Wed, Fri, Sat, Thu, Mon, Tues	.243	.233	.226	24.69	0.00
3	Clear, Fog, Cloudy, Rain	.250	.235	.008	16.37	0.24

Like the process in the former analysis in winter (Table 14), these three variables were not introduced to the model at the same time. When Discomfort Index was firstly and only used to forcast hour visites, only 1.5% (the value of adjusted R square) changes of hourly visits could be explained. Whereas, when introducing seven days in a week in the model, the explaining ablity of the model is improved a lot as the value of adjusted R square increases by 21.8%. when it comes to consider weather conditions, the explaining ablity is enhanced slightly (only 0.2% of increment). Therefore, even though all these models show significant possibilities to forecast hourly visits, weather conditions are not considered in spring owing to the slight enhancement in adjusted R square and strong correlations between different weather conditions.

Table 15: The coefficients of regression model in spring

	B	Std. Error	Beta	Sig.
(Constant)	350.446	5.929		.000
DI	-1.389	.342	-.152	.000
Mon	78.859	7.063	.590	.000
Tues	61.352	7.027	.463	.000
Wed	53.153	6.926	.413	.000
Thu	37.472	7.095	.277	.000
Fri	33.412	7.205	.241	.000
Sat	64.959	7.119	.478	.000

First of all, it should be noticed that Wed does not show clear relationship with hourly visits in Pearson correlation analysis, but it is significant to forecast hourly visits working with other variables. The model to predict hourly visits in spring can be defined as:

Model Spring:

$$Hourly\ visits = 350{,}45 - 1{,}39DI + 78{,}86Mon + 61{,}32Tues + 53{,}15Wed + 37{,}47Thu + 33{,}41Fri + 64{,}96Sat$$

According to the interpretation in Model Winter, Monday is the busiest day in spring. In addition, the Discomfort Index shows negative relationship in spring, which is opposite to that in winter.

5.3 The visiting situation in summer

According to Table 16, in summer time, there are only five variables (DI, Mon, Tues, Fri and Clear) having significant relationships with hourly visits and 500 cases contribute to the analysis. Similar with previous analysis, the strongest relationship of Discomfort Index is with hourly visits. Cloudy days and rainy days tend to relate to rainy days more, which conforms to common sense that in summer cloudy or foggy days are more easily to be followed by rainy days. All in all, no multicollinearity between independent variables should be considered still because the value is less than 90%.

Table 16: The descriptive statistics and correlation between variables in summer

		Mean	Std. De	N	HV	DI	Mon	Teu	Wed	Thu	Fri	Sat	CLE	CLO	Fog	Rain	OV
Pearson Correlation	**HV**	376.19	48.30	500		-0.36	**0.43**	0.16	0.03	-0.09	-0.25	0.02	-0.13	-0.05	0.05	0.03	0.08
	DI	15.35	2.76	500	**-,357****		-0.01	0.02	0.05	-0.03	-0.03	0.00	0.07	0.07	-0.11	0.00	-0.01
	Mon	0.16	0.36	500	**,426****	-0.01		-0.19	-0.18	-0.17	-0.18	-0.17	-0.01	0.01	-0.05	0.02	0.15
	Teu	0.16	0.37	500	,157**	0.02	**-0.19**		-0.18	-0.18	-0.18	-0.17	0.02	0.04	-0.11	0.05	-0.03
	Wed	0.15	0.36	500	0.03	0.05	-0.18	**-0.18**		-0.17	-0.17	-0.17	-0.01	-0.01	0.01	0.00	-0.03
	Thu	0.14	0.35	500	-0.09	-0.03	-0.17	**-0.18**	-0.17		-0.17	-0.16	-0.04	0.03	0.08	-0.08	-0.03
	Fri	0.15	0.35	500	**0,246****	-0.03	-0.18	-0.18	-0.17	-0.17		-0.16	0.06	0.02	-0.02	-0.02	-0.03
	Sat	0.14	0.34	500	0.02	0.00	-0.17	**-0.17**	-0.17	-0.16	-0.16		0.00	-0.01	0.03	-0.01	-0.03
	CLE	0.03	0.17	500	**-,130****	0.07	-0.01	0.02	-0.01	-0.04	0.06	0.00		-0.10	-0.08	-0.19	-0.01
	CLO	0.25	0.43	500	-0.05	0.07	0.01	0.04	-0.01	0.03	0.02	-0.01	-0.10		-0.28	**-0.61**	-0.04
	Fog	0.19	0.39	500	0.05	-0.11	-0.05	-0.11	0.01	0.08	-0.02	0.03	-0.08	-0.28		**-0.51**	-0.03
	Rain	0.53	0.50	500	0.03	0.00	0.02	0.05	0.00	-0.08	-0.02	-0.01	-0.19	-0.61	**-0.51**		-0.07
	OC	0.00	0.06	500	0.08	-0.01	**0.15**	-0.03	-0.03	-0.03	-0.03	-0.03	-0.01	-0.04	-0.03	-0.07	

**p <0,01

Table 17: The regression model summary in summer

Model	Variables Entered	R Square	Adjusted R Square	R Square Change	F	Sig
1	Discomfort Index	.128	.126	.128	72.967	0.00
2	Wed, Fri, Sat, Thu, Mon, Tues	.481	.474	.353	65.164	0.00
3	Clear, Fog, OC, Rain	.498	.487	.017	44.062	0.00

The variables are entered as the same manner with the former analysis in winter and spring. Compared with spring, Discomfort Index explains more about hourly visits in summer. It is the same that after introducing seven days in a week, the value of adjusted R square is increased dramatically as that in winter. After considering weather conditions, the model can explain almost half of the variance in hourly visits. Cloudy is deleted from the model automatically owing to the great possible linear relationship with other weather conditions.

Table 18: The coefficients of regression model in summer

	B	Std. Error	Beta	Sig.
(Constant)	415.176	10.600		.000
DI	-6.101	.567	-.349	.000
Mon	97.515	6.126	.733	.000
Tues	69.643	6.103	.526	.000
Wed	55.323	6.127	.409	.000
Thu	38.162	6.212	.276	.000
Fri	21.291	6.189	.156	.001
Sat	52.284	6.258	.371	.000
Clear	-24.369	9.483	-.086	.010
Fog	11.924	4.821	.097	.014
Rain	5.492	3.789	.057	.148
Overcast	15.783	24.925	.021	.527

One special case should be noticed in summer which is no snowy days in summer. That is to say no comparison with snow is needed when interpreting the result. As such, rainy days and overcast days have no significant influence on forecasting hourly visits in summer. The model in summer can be defined as:

Model Summer:

$$Hourly\ visits = 415.18 - 6.1DI + 97.51Mon + 69.64Tues + 55.32Wed + 38.16Thu + 21.29Fri + 52.28Sat - 24.37clear + 11{,}92fog + 5{,}49rain + 15{,}78overcast$$

In summer, Discomfort Index shows negative influence in predicting hourly visits and such a negative influence is stronger than that in spring. Monday is much busier than other days in a week. But unlike cases in winter and spring, Saturday is not that busy in summer. When there is one clear day in summer, it will cause 24 units of decline in hourly visits. And such a declining trend is very obvious when other variables are held constant.

5.4 The visiting situation in autumn

Based on Table 19, 442 cases belong to the season group: autumn. Hourly visits have the strongest relationship with the Discomfort Index. Both Discomfort and Mon accounts for almost 65% of the dependent variable's correlation.

Table 19: The descriptive statistics and correlation between variables in autumn

		Mean	Std. De	N	HV	DI	Mon	Teu	Wed	Thu	Fri	Sat	CLE	CLO	Fog	Rain	OV
Pearson Correlation	**HV**	386.68	40.06	442		**0.34**	0.30	0.16	0.01	-0.14	-0.26	0.27	-0.06	0.17	0.12	0.06	-0.01
	DI	3.37	5.44	442	0.341**		0.02	0.01	-0.04	0.01	0.02	-0.02	-0.05	0.05	0.10	**0.44**	0.05
	Mon	0.15	0.35	442	**0.303****	0.02		-0.17	-0.17	-0.17	-0.17	-0.16	-0.06	0.04	0.02	0.00	0.01
	Teu	0.15	0.36	442	0.157	0.01	-0.17		**-0.18**	-0.17	-0.17	-0.16	0.04	0.04	0.00	,01	0.05
	Wed	0.15	0.36	442	0.031	-0.04	-0.17	**-0.18**		-0.18	-0.17	-0.16	0.03	0.02	0.04	-0.02	0.05
	Thu	0.15	0.36	442	-.136**	0.01	-0.17	-0.17	**-0.18**		-0.17	-0.16	0.09	-0.05	0.00	-0.02	0.01
	Fri	0.14	0.35	442	-,258**	0.02	-0.17	-0.17	**-0.17**	-0.17		-0.15	-0.05	0.04	-0.11	0.06	-0.03
	Sat	0.13	0.34	442	0.270**	-0.02	-0.16	-0.16	**-0.16**	-0.16	-0.15		0.00	-0.04	0.04	-0.05	-0.06
	CLE	0.02	0.13	442	0.062	-0.05	-0.06	0.04	-0.03	**0.09**	-0.05	0.00		-0.07	-0.04	-0.12	-0.02
	CLO	0.20	0.40	442	-.168**	0.05	-0.04	0.04	0.02	-0.05	0.04	-0.04	-0.07		**-0.16**	-0.43	-0.08
	Fog	0.09	0.29	442	0.122**	0.10	0.02	0.00	0.04	0.00	-0.11	0.04	-0.04	-0.16		**-0.28**	-0.05
	Rai	0.44	0.50	442	0.055	0.44	0.00	0.01	-0.02	-0.02	0.06	-0.05	-0.12	-0.43	-0.28		-0.15
	OC	0.03	0.16	442	-0.009	0.05	0.01	0.05	0.05	0.01	-0.03	-0.06	-0.02	-0.08	-0.05	**-0.15**	

**p <0,01

Table 20: The regression model summary in autumn

Model	Variables Entered	R Square	Adjusted R Square	R Square Change	F	Sig
1	Discomfort Index	.116	.114	.116	57.829	0.00
2	Wed, Fri, Sat, Thu, Mon, Tues	.461	.452	.345	53.041	0.00
3	Clear, Fog, Cloudy, Rain，OC	.486	.472	.025	33.809	0.00

Once again, seven days in a week contributes more variance in hourly visits compared with other the two variables from the perspective of entering the model. After considering all the variables and dummy variables, 47,2% changes in hourly visits could be explained, which implies a powerful ability to demonstrate the variance.

Table 21: The coefficients of regression model in autumn

	B	Std. Error	Beta	Sig.
(Constant)	338.958	4.492		0.000
DI	2.364	0.341	0.321	0.000
Mon	62.111	5.208	0.546	0.000
Tues	48.714	5.211	0.431	0.000
Wed	36.433	5.177	0.327	0.000
Thu	22.034	5.175	0.196	0.000
Fri	7.771	5.274	0.067	0.141
Sat	63.303	5.357	0.530	0.000
Clear	-12.193	10.893	-0.041	0.264
CLO	15.587	4.924	0.154	0.002
Fog	11.139	6.087	0.081	0.068
Rain	2.152	4.750	0.027	0.651
Overcast	-3.188	9.397	-0.013	0.735

According to the numbers listed in Table 22, the model in autumn can be defined as:
Model Autumn:

$$Hourly\ visits = 338.96 + 2.36DI + 62.11Mon + 48.71Tues + 36.43Wed + 22.03Thu + 7.77Fri + 63.3Sat - 12.19clear + 15.59cloudy + 11{,}14fog + 2.15rain - 3.19overcast$$

Model Autumn shows that Friday and Sunday show no significant influence on forecasting hourly visits and Saturday is the busiest day in autumn. When it comes to weather conditions, only cloudy days have a great difference compared with snowy days. That is to say, if one day changes from snowy into cloudy, there will be 15 more people visiting Turku Main Library and such change is significant.

5.5 Model checking and summary

There are 21 cases listed in the case-wise diagnostics table (three in winter, eight in spring, two in summer and eight in autumn), which means the residual statistics in these cases are extreme. As for an ordinary sample, it is reasonable to expect 95% of cases to have standardized residuals. Thereby, such an amount of cases is acceptable.
In order to check the assumptions of these four models, the plot of *ZRESID and *ZPRED and the histogram of residuals to show the distribution were required.

Figure 8: Model checking in winter

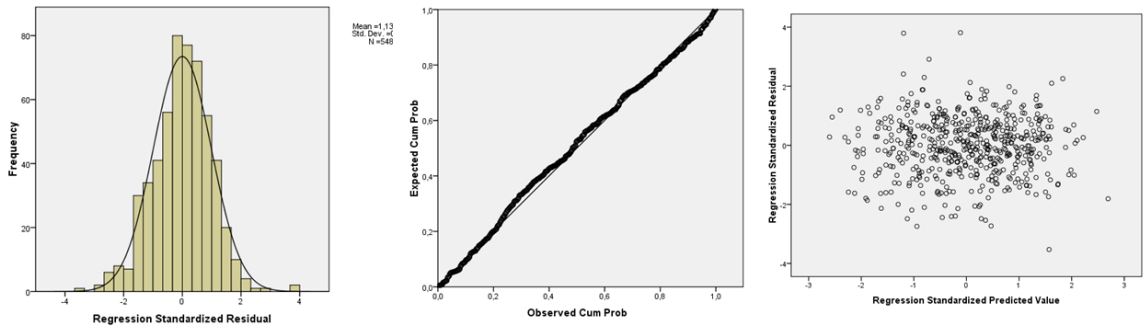


Figure 9: Model checking in spring

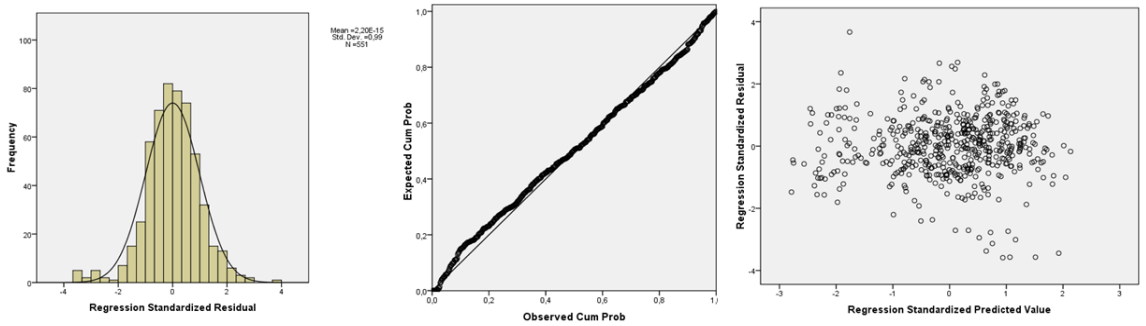


Figure 10: Model checking in summer

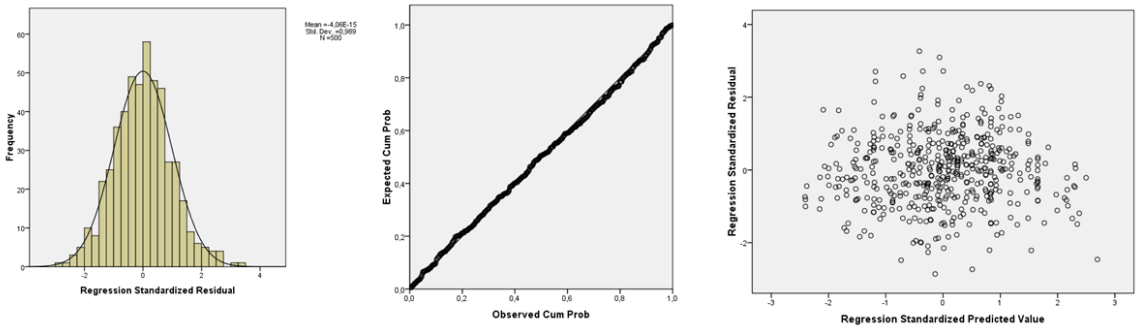
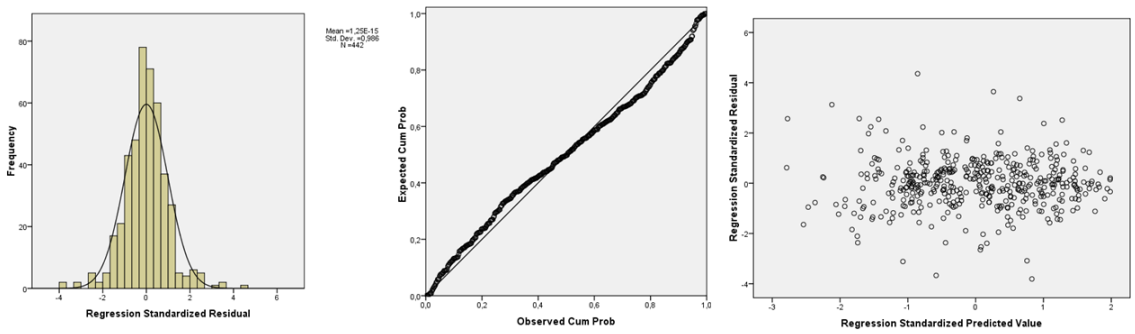


Figure 11: Model checking in autumn



All the pictures in the left-hand side from Figure 8 to Figure 11 present the distribution of residuals in each model and all of them appear as normal distribution even though such distribution is rough normal. The middle pictures also represent the normal distribution and it can be assumed that Model Winter and Model Summer have better normal distribution compared with Model Spring and Model Autumn, which further implied that the result generated from Model Winter and Model Summer is more reasonable and trustable. The right-hand side pictures in these figures illustrate the assumptions of linearity and homoscedasticity. Since there are no clear patterns that can be drawn from all these pictures, it indicates that all the assumptions are met. Therefore, Model Winter to Autumn show the feasibility of practical application even though they are not perfectly matched with the requirements in regression analysis.

All in all, information interprets from the models is listed in Table 22

Table 22: Model Summary

Seasons	N	Mean	Model Summary
Winter	548	377,36	1. Saturday is the busiest day
			2. When weather changes from snowy to sunny, hourly visits will increase dramatically
			3. DI has the biggest relation with hourly visits
Spring	551	387,83	1. Only seven days in a week are useful to predict hourly visits
			2. Monday is the busiest day
Summer	500	376,19	1. Rainy and overcast weather cannot predict hourly visits
			2. Monday is the busiest day
			3. Saturday is not as busy as those in other seasons
Autumn	442	386,68	1. DI has the biggest relation with hourly visits
			2. The difference between Friday and Sunday cannot predict hourly visits
			3. Saturday is the busiest day
			4. When weather changes from snowy to cloudy, hourly visits will increase dramatically
Total	2041	381,92	1. Seven days in a week play the most important role in predicting hourly visits
			2. DI has the opposite influence in different seasons
			3. In different seasons, only certain weather conditions are helpful to predict hourly visits

5.6 Classifying Discomfort Index in the library context

As is demonstrated in the data collection stage, most weather conditions in the Turku region were grouped into "No discomfort". This, on one hand, implies that Turku is a living friendly city. On the other hand, owing to the correlation with hourly visits, wisely classified groups within "No discomfort" could be effective for librarians to fast evaluate the visiting situation. Considering this, classification within "No discomfort" was conducted. The "Tree" function under "Classify" command was employed. Since the explaining ability of Discomfort Index in Model Spring is not very strong even though p value is 0, such classification was not operated in spring. This is because most of the variance in hourly visits is explained by seven days in a week based on the adjusted R value in Table 15. The breaking point in each group within "No discomfort" is named as "Library-Climate point".

Table 23: Library-Climate points in winter, summer and autumn

Season	Library-Climate point	N	Mean value of Hourly Visits
Winter	≤-8.099	109	352.42
	(-8.099, -0.659]	220	374.21
	>-0.659	219	392.94
Summer	≤14.088	199	393.87
	(14.088, 17.134]	151	376.61
	>17.134	150	352.32
Autumn	≤-0.194	86	360.99
	(0.194, 5.502]	179	387.93
	>5.502	177	397.9

Table 23 illustrates the classification result in each season. Three groups of DI are generated within "No discomfort" in seasons. As in Model Summer, Discomfort Index shows negative relationship with hourly visits, this is also reflected in Table 24. When DI increases from 14,088 to 17,134, hourly visits decrease. All these Library-Climate points could be used as a hint to predict how many users will come to the library. For example, if DI in one day is 2,4 in winter, it means that roughly 400 people will come to the library.

5.7 The interpretation of data mining result

Since the main goal of this thesis is to explore the feasibility of data mining in enhancing library reference service, the interpretation is accomplished with the emphasis of this goal. Whether data mining could be helpful in service improvement will be evaluated through the facts of whether it is helpful to conquer challenges in current reference services. As is mentioned in the third section that four reference service challenges could be overcome with the help of data mining, which are offering a suitable service model, balancing different communication models, lack of professional librarians, lack of information resources. Therefore, the result will be interpreted from these four aspects.

First of all, being aware of the hourly visiting situation would be useful for staffing. Saturday is the busiest day in winter and autumn, but it is not that busy in summer. As such, a reasonable schedule for librarians on Saturday could be created. In summer time, sunny days lead to less users in the library, considering that day as a Sunday with 20 ℃ Discomfort Index, then the hourly visits would become much lower. Under this situation, there will be no use for Turku Main Library to open so many reference desks. Only one or two desks could be enough and librarians at the open desk could combine general information service and specific information service together. However, when it comes to a very busy day, for instance, one cloudy Saturday following some snowy days in autumn, the library could open as many reference desks as possible in order to give fast answers to users' questions. In addition, desks for general information and

desks for specific information could be separately worked on such days. According to this interpretation, it could be concluded that how to offer a suitable service model can be handled through the reasonable allocation of librarians, which can be finished in accord with the change in hourly visits.

In a similar manner, different communication models could be balanced. During the snowy days in winter, less people would go to the library. However, it does not mean that they have no requirements in information searching or material locating. Thus, such requirements might be met through remote communication methods, such as making a phone call, sending out an email or posting a question on the library website. Thereby, the working content of librarians should be more concerning these communication platforms. In a word, the working content on different communication models could be adjusted to the hourly visits. And how to adjust can be accomplished by mining data.

The aforementioned two challenges could also be conquered in a way with a flexible opening hours. Since Friday and Sunday have less visitors, especially Fridays in autumn, therefore, fewer opening hours on Fridays and more opening hours on Saturdays and Mondays could effectively meet users' needs. Offering suitable service models and balancing communication models could be achieved under different opening hours.
When it comes to a lack of professional librarians and information resources, data mining per se could be an effective tool. Since knowledge and information could be produced in the process of data mining, the more often data mining is conducted, the more knowledge and information will be attained by librarians. Furthermore, librarians would become more knowledgeable and professional in a general way. Since online free databases are encouraged to be discovered during data mining, they can be seen as valuable information resources. Furthermore, the amount of information could be increased as new pieces of information are generated during data mining. Simply put, data mining is a self-learning process for librarians becoming more professional and an information generation process to enrich library information sources.
For assessing the reference service, data mining might not be a very useful solution. Whereas, since online free databases and basic statistics methods are advocated, the cost of data mining is not very high, especially compared with other huge investment for libraries improvement. The improvement stemming from data mining is not pricy, or put in another way, very cheap. Hence, assessing the reference services generated from data mining is not as necessary as that for services of big investment. That is to say, services introduced by data mining have less urgency to evaluate.

All in all, even though data mining may not be a solution to radically resolve all the issues caused by the challenge, it is still an effective tool to overcome the challenge with adjustable and flexible ideas. In this case, data mining is not obviously useful in assessing service, but it is very useful to confront the other four challenges for Turku Main Library. Because all the results generated from data mining and all the models established for predicting hourly visits are interpreted with practical solutions to handle pragmatic issues. In this case, reference services are improved with the help of data mining.

5.8 Evaluation by Librarians

Even though it could be concluded that data mining is feasible to improve reference service by overcoming challenges of current library reference service, models were still presented to librarians in order to achieve a more professional evaluation. During this process, librarians who work at the reference desk in the public library were considered as the main interviewees because they might have a better understanding of reference service. In addition, since the case study is concerning a public library, librarians working in the public library were preferred in this study. Structured interviews were conducted after presenting the result of data mining. The interview language was English and it was recorded manually in the textbook. Ten librarians agreed to listen to the presentation of the result and six librarians wanted to have the interview after the presentation. For those who did not take the interview, there were two reasons: two librarians were not confident to express their opinions in English and the other two were not confident enough to give professional feedback on data mining.

The result of the interview is listed in Table 24:

Table 24: Interview summary

Librarian ID	How long have you worked in the public libraries?	Do you know what Data Mining is?	After the presentation, do you understand Data Mining?	Do you think this case could be useful for improving reference service?	From 1 to 10, how much would you grade data mining as a tool to improve reference service?
A	More than 20 years	No	Yes	Yes	8
B	15 years	No	Yes	Yes	9
C	5.5 years	No	Absolutely	Yes	8
D	1 month	No	Roughly yes	Yes	6
E	1.5 year	No	Yes	Yes	7
F	15 years	No	Yes	Yes	8

It is worth of highlighting that among all these interviewed librarians, none of them were aware of data mining. Even though six librarians are not enough to be considered as representatives, it could still be implied that data mining is not a popular concept known by librarians. After the presentation, all these librarians attained an idea about what data mining is and they showed an interest in using data mining during work if possible. All of them considered this case study as a valuable example and they believed that the result from the case could be used for service improvement. When it comes to

evaluating how much data mining could be used, they were asked to grade their opinions from one to ten, one means not at all and ten means extremely useful. It turns out that all of their opinions are on the positive side. The lowest grade is six, but the interviewee's opinion is that she can see a future with data mining in libraries. Since she has only one month of experience as a librarian, she is not sure how useful it could be. Therefore, she just graded six to show her positive opinion. As for the librarian who graded nine, she thinks that data mining is a "must-do" rather than a "have-to-do".

After answering the questions listed in Table 25, these librarians were also asked to make additional comments on the topic. Two aspects were put forward. First of all, the combination of different databases could not only be very useful and valuable, but challengeable for practical conductions. Librarians need training before they actually start to use data mining in their work. Whereas, libraries are confronting a situation of investment reduction, and it could be difficult to get enough money to support the training campaign. Second of all, how to make sure that which database could be useful is hard to decide, especially when online free databases are considered. Owing to these two concerns, there is no "ten" for the fifth question in the table.

To sum up, the librarians' opinion regarding this case study is positive and they think data mining could be used for the enhancement of reference services.

5. Discussion and Conclusion

In order to explore the possibility of applying data mining in libraries for service improvement, a case study was conducted in this thesis work. The case study is, on one hand, a simulation of a practical process using data mining, from data collection to data interpretation; on the other hand, this case study is an example of presenting how public libraries could use data mining for their reference service. According to the case study, a combination of library system database and online free database is achieved. It is worth being highlighted that mining such combined database could generate valuable information. Different visiting situations in different seasons, weather conditions and days in one week could be clearly pictured based on the result in section five. Wisely interpreting, it could provide pragmatic hints for library mandates. As in this thesis work, challenges in library reference service could be overcome with the help of the information attained from data mining. This indicates that online free databases could be an important resource for library data mining as well with reasonable combining of databases recorded within the library.

In addition, data mining could be a tool for improving library reference service the as four challenges of reference service could be conquered by data mining according to the review of previous studies. Then the interpretation of the data mining result empirically demonstrates that data mining is a feasible solution to enhance library reference service. To sum up, two research gaps, one which is between data mining and reference service and the other which is between online free data sources and library data mining, could be filled in with the result of the case study.

In order to answer the research question: is data mining feasible to improve library reference service, a constructed interview was arranged to collect opinions from a professional point of view although the data volume is not large enough to demonstrate a phenomenon in the case study. In the end, six interviewees all gave positive feedback regarding the case study result. Four of them gave more than eight points when being asked to grade how helpful data mining could be in service improvement from one to ten. This implies a promising future of data mining developed in the library. It is also worth noting that no interviewees graded ten, which indicates that concerns still remain. During the conversation with the interviewees, technological difficulties were mentioned a few times. How to store, analyze and interpret data were considered as challenges. Furthermore, no interviewees being aware of data mining might be a sign that there is a gap between academic and practical understanding of data mining. One interviewee mentioned that it could be a tough task to evaluate the quality of the database, which decides how reliable the result generated from data mining is. All of these reflections are covered by the content of section 2.1.3 where issues of library data mining are discussed by reviewing former studies.

All in all, this study not only enriches theories generated in former studies, but also answers the research question well. Meanwhile, the aim of this thesis work is achieved. A potentially useful online free database was discovered and utilized. Valuable solutions are put forward from a service improvement point of view. Even though

issues exist, it is still an acceptable idea to apply data mining to reference service improvement in libraries. Considering the situation of resource limitation and data explosion, such an application could establish a path for future libraries to develop services.

6. Expectations for future studies

For future study, more combinations of different datasets are expected. A method to help establish such combination is expected to be explored. The domain of the research field is also expected to be limited to a more specific field of reference service, such as virtual reference service. Since this thesis concentrates more on how to apply data mining, the feasibility of data mining is the center of the whole research. In the future, challenges caused by data mining should be discussed as well in order to achieve a more thorough perspective. Moreover, solutions to overcome those challenges should be explored. In addition, the volume of data could be increased, or the concept Big Data could also be a target of future studies. When the amount of data increases to certain level, Big Data characters would appear. Under these circumstances, it could be an interesting topic to explore the benefits of Big Data for library services. In this study, the type and scale of the library is not considered. It is expected that in the future, more studies could be conducted on different kinds of libraries, such as data mining in public libraries, data mining in digital libraries.

References

ADEBAYO, O. 2009. Quality Reference Service: The Fulcrum For Users' Satisfaction. *In:* OYESIKU, F. A. (ed.) *Current Trends in Library and Information Science : Essays in Honour of Late O.K. Odusanya* Ibadan: BIB Press Nig.

AGGARWAL, I. & POWERS, M. 2013. Increasing Library Access and Enhancing Reference Support through a Shared Services Model. *The Reference Librarian,* 54**,** 236-244.

AGUILAR, P., KEATING, K., SCHADL, S. & VAN REENEN, J. 2011. Reference as Outreach: Meeting Users Where They Are. *Journal of Library Administration,* 51**,** 343-358.

AJAY, D. & SANGAMWAR, A. T. 2014. Identifying the patent trend, licensing pattern and geographical landscape analysis of the Council for Scientific & Industrial Research (CSIR) of India between 2000 and 2011. *World Patent Information,* 38**,** 42-49.

ALOTAIBI, N. M., NASSIRI, F., BADHIWALA, J. H. & AL, E. 2015. The Most Cited Works in Aneurysmal Subarachnoid Hemorrhage: A Bibliometric Analysis of the 100 Most Cited Articles. *World neurosurgery,* 89**,** 587-592.

ANDERSON, C. 2008. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *WIRED MAGAZINE.* San Francisco: Wired.

BANERJEE, K. 1998. Is Data Mining Right for Your Library? *Computers in Libraries: Complete Coverage of Library Information Technology***,** 28-31.

BEHR, M. & HILL, R. 2012. Mining e-Reserves Data for Collection Assessment: An Analysis of How Instructors Use Library Collections to Support Distance Learners. *Journal of Library & Information Services in Distance Learning,* 6**,** 159-179.

BORGMAN, C. L. 1999. What are Digital Libraries? Competing Visions. *Information Processing and Management: an International Journal - Special Issue on Progress toward Digital Libraries,* 35**,** 227-243.

BOYNE, G. A. 2003. Sources of Public Service Improvement: A Critical Review and Research Agenda. *Journal of public administration research and theory,* 13**,** 367-394.

CAMPBELL, J. D. 2000. Clinging to Traditional Reference Services: An Open Invitation to Libref. com. *Reference & User Services Quarterly***,** 223-227.

CHEN, C. L. P. & ZHANG, C.-Y. 2014. Data-intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data. *Information Sciences,* 275**,** 314-347.

CHEN, M., MAO, S. & LIU, Y. 2014. Big Data: A Survey. *Mobile Networks and Applications,* 19**,** 171-209.

CLEYLE, S. & NICHOLSON, S. 2006. Approaching Librarianship from the Data: Using Bibliomining for Evidence-based Librarianship. *Library hi tech,* 24**,** 369-375.

CUDDY, C., GRAHAM, J. & MORTON-OWENS, E. G. 2010. Implementing Twitter in a Health Sciences Library. *Medical Reference Services Quarterly,* 29**,** 320-330.

DEMPSEY, N., BRAMLEY, G., POWER, S. & BROWN, C. 2011. The Social Dimension of Sustainable Development: Defining Urban Social Sustainability. *Sustainable Development,* 19**,** 289-300.

DUMOUCHEL, B. & DEMAINE, J. 2006. Knowledge Discovery in the Digital Library: Access Tools for Mining Science. *Information Services and Use,* 26**,** 39-44.

FIELD, A. 2009. *Discovering Statistics Using SPSS*, Sage Publications.
FIELD, A. 2013. *Discovering Statistics Using IBM SPSS Statistics*, Sage.
GILLHAM, B. 2000. *Case Study Research Methods*, Bloomsbury Publishing.
GOLD, A. 2010. Data Curation and Libraries: Short-term Developments, Long-term Prospects. *Office of the Dean (Library)*, 27.
GORDON-MURNANE, L. 2012. Big Data: A Big Opportunity for Librarians. *Online,* 36, 30-34.
GREEN, S. S. 1993. Personal Relations between Librarians and Readers. *Library Journal,* 118, S4.
HAJEK, P. & STEJSKAL, J. 2012. Analysis of User Behavior in a Public Library Using Bibliomining. *Advances in Environment, Computational Chemistry and Bioscience*, 339-344.
HAN, L. & GOULDING, A. 2003. Information and Reference Services in the Digital Library. *Information Services & Use,* 23, 251-262.
HEIDORN, P. B. 2011. The Emerging Role of Libraries in Data Curation and E-science. *Journal of Library Administration,* 51, 662-672.
HERNON, P. & ALTMAN, E. 2010. *Assessing Service Quality: Satisfying the Expectations of Library Customers*, American Library Association.
HOY, M. B. 2014. Big Data: An Introduction for Librarians. *Medical reference services quarterly,* 33, 320-326.
HULL, T. J. & ADAMS, M. O. 1995. Electronic Communications for Reference Services: A Case Study. *Government Information Quarterly,* 12, 297-308.
HWANG, S.-Y. & LIM, E.-P. 2002. A Data Mining Approach to New Library Book Recommendations. *In:* LIM, E.-P. E. A. (ed.) *The 5th International Conference on Asian Digital Libraries.* Singapore: Springer.
KAMDAR, T. & JOSHI, A. 2005. Using Incremental Web Log Mining to Create Adaptive Web Servers. *International Journal on Digital Libraries,* 5, 133-150.
KARNO, M. R., NOORDIN, S. A., TALIB, M. A. & RAHMAN, M. S. A. Facilitating Resource Allocation Decision through Data Mining: The Case of UTM Library. *In:* TEKTAS， ARZU, ed. The 18th IBIMA International conference 2012.
KATZ, L. S. 2013. *Digital Reference Services*, Routledge.
KEISANONLINECALCULATOR. *Discomfort Index Calculator* [Online]. Available: http://keisan.casio.com/exec/system/1351058230.
KELOĞLU-İŞLER, E. İ. & BAYRAM, Ö. G. 2014. Commodification of Knowledge Communication Mediums: From Library to Social Media. *Procedia-Social and Behavioral Sciences,* 147, 550-553.
KLAMPFL, S., GRANITZER, M., JACK, K. & KERN, R. 2014. Unsupervised Document Structure Analysis of Digital Scientific Articles. *International Journal on Digital Libraries,* 14, 83-99.
KURUPPU, P. U. 2007. Evaluation of Reference Services—A Review. *The Journal of Academic Librarianship,* 33, 368-381.
LEE, J.-M. 2013. Understanding Big Data and Utilizing its Analysis into Library and Information Services. *Journal of the Korea Biblia Society for Library and Information Science,* 24, 53-73.

LIBRARYBUILDINGS.INFO. *Turku Main Library* [Online]. Available: http://www.librarybuildings.info/finland/turku-main-library-part-turku-city-library.
LIU, Y., KOSTAKOS, V. & LI, H. 2015. Climatic Effects on Planning Behavior. *PLoS One,* 10**,** 9.
LUO, L. & WEAK, E. 2013. Text Reference Service: Teens' Perception and Use. *Library & Information Science Research,* 35**,** 14-23.
MAZON, J. 2014. The Influence of Thermal Discomfort on the Attention Index of Teenagers: an Experimental Evaluation. *International Journal of Biometeorology,* 58**,** 717-724.
MOED, H. F., BURGER, W., FRANKFORT, J. & VAN RAAN, A. F. 1985. The Use of Bibliometric Data for the Measurement of University Research Performance. *Research Policy,* 14**,** 131-149.
NICHOLSON, S. 2003a. Bibliomining for Automated Collection Development in a Digital Library Setting: Using Data Mining to Discover Web‐based Scholarly Research Works. *Journal of the American Society for information science and technology,* 54**,** 1081-1090.
NICHOLSON, S. 2003b. The Bibliomining Process: Data Warehousing and Data Mining for Library Decision Making. *Information technology and libraries,* 22**,** 146-151.
NICHOLSON, S. 2006. The Basis for Bibliomining: Frameworks for Bringing Together Usage-based Data Mining and Bibliometrics through Data Warehousing in Digital Library Services. *Information Processing & Management,* 42**,** 785-804.
NICHOLSON, S. & STANTON, J. M. 2003. Gaining Strategic Advantage through Bibliomining: Data Mining for Management Decisions in Corporate, Special, Digital, and Traditional Libraries. *Organizational data mining: Leveraging enterprise data resources for optimal performance***,** 247-262.
NOH, Y. 2015. Imagining Library 4.0: Creating a Model for Future Libraries. *The Journal of Academic Librarianship,* 41**,** 786-797.
NUNN, B. & RUANE, E. 2011. Marketing Gets Personal: Promoting Reference Staff to Reach Users. *Journal of Library Administration,* 51**,** 291-300.
OBERHAUSER, O. C. 1991. Interactive Multimedia in Library and Information Services. *Audiovisual Librarian,* 17**,** 17-25.
OGIER, A., HALL, M., BAILEY, A. & STOVALL, C. 2014. Data Management Inside the Library: Assessing Electronic Resources Data Using the Data Asset Framework Methodology. *Journal of Electronic Resources Librarianship,* 26**,** 101-113.
OKERSON, A. 2013. Text & Data Mining-a Librarian Overview. *79th IFLA General Conference and Assembly* Singapore: IFLA World Library and Information Congress
PENDLEBURY, D. 2010. White paper using bibliometrics in evaluating research.
PERNER, P. 2002. *Advances in Data Mining*, Springer.
POL, K., PATIL, N., PATANKAR, S. & DAS, C. A Survey on Web Content Mining and extraction of Structured and Semistructured data. Emerging Trends in Engineering and Technology, 2008. ICETET'08. First International Conference on, 2008. IEEE, 543-546.
PUN, R. 2015. WeChat in the Library: Promoting a New Virtual Reference Service Using a Mobile App. *Library Hi Tech News,* 32**,** 9-11.

RANASINGHE, W. 2012. New Trends of Library Reference Services. *Prof. Jayasiri Lankage Felicitation Volume.* Colombo: Godage.
SAUNDERS, L. 2013. Learning from Our Mistakes: Reflections on Customer Service and How to Improve It at the Reference Desk. *College & Undergraduate Libraries,* 20**,** 144-155.
SEWELL, R. R. 2013. Who is Following Us? Data Mining a Library's Twitter Followers. *Library Hi Tech,* 31**,** 160-170.
SHARMA, C. 2006. *Reference Service and Sources,* New Delhi, Atlantic Publishers & Distributors Pvt Ltd
STANDERFER, A. E. 2006. Reference Services in Rural Libraries. *The Reference Librarian,* 45**,** 137-149.
STEVENS, C. R. 2013. Reference Reviewed and Re-Envisioned: Revamping Librarian and Desk-Centric Services with LibStARs and LibAnswers. *The Journal of Academic Librarianship,* 39**,** 202-214.
SUITS, D. B. 1984. Dummy Variables: Mechanics v. Interpretation. *The Review of Economics and Statistics***,** 177-180.
TODORINOVA, L., HUSE, A., LEWIS, B. & TORRENCE, M. 2011. Making Decisions: Using Electronic Data Collection to Re-envision Reference Services at the USF Tampa Libraries. *Public Services Quarterly,* 7**,** 34-48.
TYCKOSON, D. A. 2011. Issues and Trends in the Management of Reference Services: A Historical Perspective. *Journal of Library Administration,* 51**,** 259-278.
UPPAL, V. & CHINDWANI, G. 2013. An Empirical Study of Application of Data Mining Techniques in Library System. *International Journal of Computer Applications,* 74**,** 42-46.
VEL SQUEZ, J. D. 2013. Web Mining and Privacy Concerns: Some Important Legal Issues to Be Consider Before Applying Any Data and Information Extraction Technique in Web-based Environments. *Expert Systems with Applications,* 40**,** 5228-5239.
WANG, B., XU, R., ZHU, J. & AL, E. 2005. Study on Applications of Web Mining to Digital Library. *Artificial Intelligence Applications and Innovations.* Springer.
WASIK, J. M. 1999. Building and Maintaining Digital Reference Services. . *ERIC Digest.*
WEIMER, K. 2010. Text Messaging the Reference Desk: Using Upside Wireless' SMS-to-email to Extend Reference Service. *The Reference Librarian,* 51**,** 108-123.
WITTEN, I. H., DON, K. J., DEWSNIP, M. & TABLAN, V. 2004. Text Mining in a Digital Library. *International Journal on Digital Libraries,* 4**,** 56-59.
WITTMANN, R. J. & REINHALTER, L. 2014. The Library: Big Data's Boomtown. *The Serials Librarian,* 67**,** 363-372.
XIA, J. & WANG, M. 2014. Competencies and Responsibilities of Social Science Data Librarians: An Analysis of Job Descriptions. *College & Research Libraries,* 75**,** 362-388.
XIANG, Z., SCHWARTZ, Z., GERDES JR, J. H. & UYSAL, M. 2015. What Can Big Data and Text Analytics Tell Us about Hotel Guest Experience and Satisfaction? *International Journal of Hospitality Management,* 44**,** 120-130.
YAN, F., ZHANG, M., TANG, J., SUN, T., DENG, Z. & XIAO, L. 2010. Users' Book-loan Behaviors Analysis and Knowledge Dependency Mining. *Web-Age Information Management.* Springer.

YANG, S. Q. & DALAL, H. A. 2015. Delivering Virtual Reference Services on the Web: An Investigation into the Current Practice by Academic Libraries. *The Journal of Academic Librarianship,* 41**,** 68-86.

YANKOVA, I. V. 2013. Marketing of the Library-information Services. *Journal of Balkan Libraries Union,* 1.

ZUCCALA, A., THELWALL, M., OPPENHEIM, C. & DHIENSA, R. 2007. Web Intelligence Analyses of Digital Libraries: A Case Study of the National Electronic Library for Health (NeLH). *Journal of Documentation,* 63**,** 558-589.