Application of Big Data and Data Mining Techniques to Designing Building

Hyoung-Seon Choe, Jinhwa Kim

Abstract—Background/Objectives: ***This study suggests an approach in designing new building, specifically a library with peer to peer social big data and survey data.***

Methods/Statistical analysis: ***Big data techniques such as social data mining, text mining, and association rule analysis are used in this study. This study uses sentiment analysis and opining mining in analyzing social data. Association rule analysis is used to understand the behavioral patter in survey data on daily movement of users in libraries.***

Findings: ***This study confirms that big data techniques such as social data mining, text mining, and association rule analysis can be efficiently applied to designing a building such as a library. Nouns related to library extracted from social media such as Twitter &blogs describe major services and facilities many people want in libraries. Adjectives from social data show that users' feeling on the libraries. An analysis of data set from actual movement behaviors in the library shows efficient routing for library users. The study finds that data mash-up and big data techniques can help design new building, which is more efficient and convenient for users.***

Improvements/Applications: ***Designing a building using more advanced technique such as an artificial intelligence technique is possible with more diverse applications in design areas.***

Index Term—Designing Building, Data Mash-Up, Text Mining, Social Mining, Sentiment Analysis, Opinion Mining

I. INTRODUCTION

The information age, which is also called the computer age, new media age or digital age, is in rapid shift from traditional economy to a new economy based on information technologies in the 21st century. In recent days, data availability has increased explosively. Data has become a raw material of new business and a new source of tremendous social and economic value. The growth of digital networks, sensors, and devices has revolutionized communication, sharing, creation and access to data. In addition, government also collects and shares data for creating values in economy.

Especially, social data can provide diverse information on users with their idea, opinion, and needs. People communicate using social media such as Facebook, Twitter, and Blogs these days. The data from these Peer-to Peer communication systems provides lots of information on users. Peer-to-Peer(P2P) streaming data such as a social data represents a scalable, robust, and economical alternative to data from traditional client–server(CS). P2P cloud computing refers to a system that uses information technology resources as needed, and provides services by using information convergence technology. Cloud systems can be classified into groups based on the methods providing the services and their goals[1].The basic idea is that instead of streaming with media on a dedicated server, the user terminal that coordinates the streaming itself configures the application-level overlay. While receiving the stream, the terminal acts as a hub to distribute the stream at the same time. Cloud computing technology is basically a tool for sharing resources. It provides the possibility of sharing data between different organizations. In addition, different organizations can maintain their own cloud environment while exchanging data with peer-to-peer streaming based on a specific object[2].

As the size of data has become bigger and increasingly complicated, we need skills that can analyze and extract information from the data. Therefore, it is necessary to analyze large-scale stream data more efficient way. As the data size increases, the run time is increased. In many studies these days, researchers introduce diverse mining algorithms to extract useful information from large-sized data sets. This study uses social and survey data and it also uses big data techniques to analyze these data sets.

Data mash-up is an application approach enabling users to generate application by reusing and integrating data from different sources. Data mash-up integrates information and data from more than one source into a single application. A variety of data providers provides a highly customized professional service platform to users to provide flexible deliveries[3], [4]. A typical type of mash-up application combines data from various data providers according to needs by users. It then integrates data and information from various sources to satisfy business needs[5].

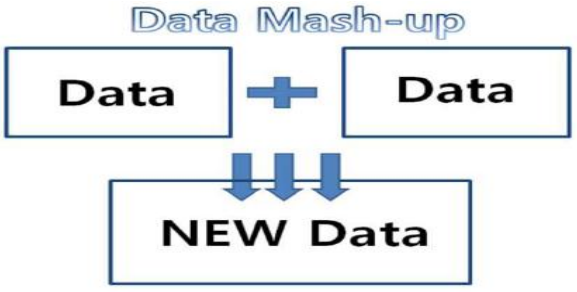

Figure 1. Data mash-up

Data mash-up can produce new business values, which is not possible with single set of data. Figure1and Figure2 show the way to generate

Revised Manuscript Received on May 09, 2019.

Hyoung-Seon Choe, School of Business, Sogang University, 1 Shinsoo-Dong, Mapo-Gu, Seoul, Korea

Jinhwa Kim, School of Business, Sogang University, 1 Shinsoo-Dong, Mapo-Gu, Seoul, Korea

new application data with data mash-up approach.

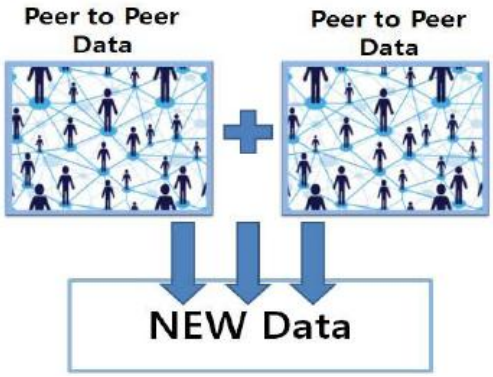

Figure 2. Peer to Peer data mash-up

Figure 3 shows the design processes in this study:1). Research planning, 2). Design concept establishment, 3). Design development, 4). Design control, and 5). Design fallow-up. The design concept, with its meaning combined with required functions, plays an important role in the performance of modern design industry. In addition, as the flow of people and their optimal movement to give convenience with reduced times is getting more and more important, there is a need for clear and accurate design concept. Concept development during the design process is a very important process in the whole process of design.

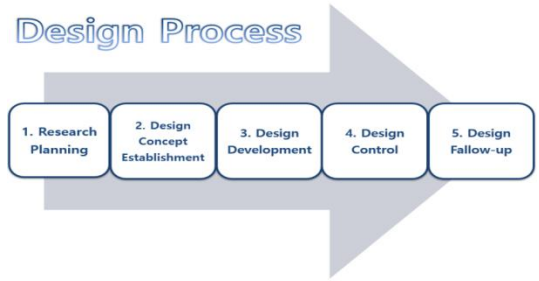

Figure 3. Design processes in this study

In this study, we recognize the importance of concept in drawing design and look for a new approach to design. By using big data techniques and data mash-up, the study tries to find needs of libraries users and presents the design concept of library. Major tools used in this study are text mining, sentiment analysis, and opinion mining.

Text mining is a technology that analyzes text data. Text mining tries to extract important information from unstructured text data. Text mining help users find meaningful models, patterns, rules, or trends in unstructured text data such as text files, HTML files, chat messages, blogs, and Twitter. Unlike traditional content analysis techniques, text mining is primarily data based, and its main purpose is to automatically identify hidden patterns in the text data. It also creates models that describe interesting patterns and trends in text data[6], [7]. As the importance of unstructured data increases in these days, text mining became an important technique providing unique business values to companies.

Sentiment analysis refers to an analytical method that extracts sentiment information from text data. It determines the polarity (positive or negative) and strength of polarity (weakly negative, mildly negative, strongly negative) of major key words in the text.

Opinion mining can be defined as extracting people's opinion from the web site or social media and it generates useful information from this opinion. With the recent explosion on internet uses, users can contribute and express themselves through social networks, videos, blogs, and web news. Analyzing the data in the form of text data and finding the opinions from users' social data with statistical and data mining techniques provide a lot of valuable information. It is also considered to be a method to discover useful information from many useful data such as user's thoughts and expressions, replacing asking and answering in traditional survey approaches[8].

In emotional analysis, they use emotional and sentiment categories for words that express emotions in advance. Emotional and sentiment extraction, or emotional and sentiment evaluation are used here. For the sake of simplicity in this paper, this study follows polarity of words in existing emotion dictionary. The polarity here means positive or negative emotion regarding a topic.

Data mining has gained great attention due to the arrival of the data explosion age. Data mining is called knowledge discovery or data discovery from database or data warehouse. It is used to discover useful information or knowledge from data sets such as databases, news, web, and social media. It consists of various fields like statistics, machine learning, artificial intelligence, databases, information retrieval, and data visualization.

Association rules mining is a technique of creating meaningful connections or more specifically connection rules between variables. Association rule mining is an important sub-field of data mining and is widely applied in many application areas[9]. The association rule analysis is rule-based method for discovering a small set of rules in the database that forms an accurate classifier and attractive relations between variables in huge databases. It consists of two main steps:the first step is finding the set of all frequent itemsets[10] and the second step is generating and testing all high confidence rules among item sets. It is for identifying strong association rules discovered in databases using some measures of attractiveness. An association rule is divided into two parts: a condition part, 'if' and a result part, 'then', 'if' part is an item found in the data. A result, 'then' part is an item found in combination with the predecessor. Association rules are created using baseline support and assurance to analyze data for frequent if /then patterns and to identify the most important relationship. Support indicates how often an item is displayed in the database. Reliability or confidence indicates the number of times a if / then statement has been found to be true.

One of the most popular applications of association rule analysis is market basket analysis, which is often seen at marts or department stores. This is a tool for recording POS(Point Of Sale) data and analyzing their behavioral patterns through association rules. The most famous example is a marketing case where men bought diapers & beer at the same time.

Rule induction method is

an analytical method that can classify and predict with the decision. In addition, a rule induction method is a method which is used in many research fields. Decision tree is an analytical method composed of nodes and branches. A decision tree is composed of tree structure, and root node is the starting node. As a rule induction algorithm is to visualize the data processing into tree structure, it can easily be understood and it can easily be implemented [11].

II. MATERIALS AND METHODS

Data collection and variable description are introduced in this part. A modern library has diverse shops and it also provides convenience to its users. Copy center, coffee shop, banks, post office, and restaurants can be found in the library such as a Congress Library. The total number of variables, representing service shops or facilities in the libraries, is 15in this study. These 15 services or facilities are also found from survey and social data. 500,000 social messages containing the key word, 'library' in social media are collected for the study. A data set on the movement among major facilities inside the library is collected from survey with 200 participants.

2.1. *Methods and Tools*

This paper suggests an approach using both social data and survey data on people's movement insides the library with text mining and data mining technique. Research procedures are in the following order: ①Data collection, ②Analysis with text mining, ③Analysis with data mining, ④Generating concept for design.

2.1.1. Modeling and Analysis

Frequently mentioned words regarding library are collected from Twitter and blogs. In addition, feelings on library expressed in Twitter and blogs are analyzed with their polarity of positive, negative, and neutral. These words expressing feeling can also be regarded as the attitude toward consumers' keywords (emotional factors) on library. These words of feeling represent positive and negative opinions on the library by users in the library. The study extracts keywords using information retrieval, atypical big data analysis and semantic technology based on the natural language processing technology, which is implemented with R program. These words related to library, which is positive, negative or neutral words, shows users' opinions on library, which also can be applied to designing more efficient and convenient libraries.

Related words, especially nouns related to library are collected from social data. When considering a future library, there are services or facilities that are considered to be the most necessary services for future library. Some are already in the library and some are currently not in services.

From the surveys on what services and facilities users normally used or visited, this study finds the optimal arrangement of services and facilities in the library.

Combining the information from the analysis on the social data and survey data, this study suggests a framework that can be used to build an optimal design of libraries.

III. RESULTS AND DISCUSSION

This study finds meaningful and relevant words regarding library from the social data as in Figure 4 and Figure5. The number below the word means frequency of this word and % represent portion of this frequency compared to that of other word. Figure 4 shows services and facilities better be included inside a library. They are children's room, reading room, coffee shop, restroom, art gallery, lecture room, cultural space, changing room, health center, playroom, computer room, language lab, rest room, post office, exhibition hall, landscape theme room, and study room. This study also shows that the library users' diverse needs are shown not only from the existing functional services in the traditional library, but also from the fact that people want more new functional services those not available in the traditional library space, as mentioned in the previous research. The study confirms diverse functions, services and facilities that users in the library want. It gives guideline for designing new libraries or future libraries. The decision on design concept on a building is very important. Most design concepts, however, are decided by intuition. In addition, there are not many systematic approaches which help establish design concept.

This study provides an approach that uses the big data to derive a design concept that can be applied to library design. People's opinion represented with affirmative/negative words related to the library and list of the high frequency words give guideline to designing libraries. These provide the design concept with keywords that represent the needs of users. Key words regarding a topic are helpful in interpreting a given topic, and visual representation of these keywords using photographs and images provides interface to users to understand the problems easier[5].

A recent study shows that language-based stimuli and keywords are helpful in the development of creative concept such as design[12]. Also, using verbs and adjectives, new concepts can be promoted into a higher level more than when using nouns only. Considering this theory, this study uses not only nouns, but it also uses adjectives related to library from social data. The adjective keywords in Figure 5 induced from social data are as follows: clean, beautiful, happy, precious, cool, fascinating, cute, new, and blue. These adjectives are positive/negative words associated with the library and are the result of many people's opinion in relation to the library. They provide a desirable concept in designing a library[13], [14].A library should be designed to be a beautiful and happy place based on the derived adjectives.

Figure 6 shows graph showing relationship among services and facilities in a library. It is an output from association rule analysis on a survey from library users. The survey includes questions on what services and facilities library users used on a regular day. The number of derived rules from association rule analysis is 80, and relations or association among them

are graphed as in Figure 6

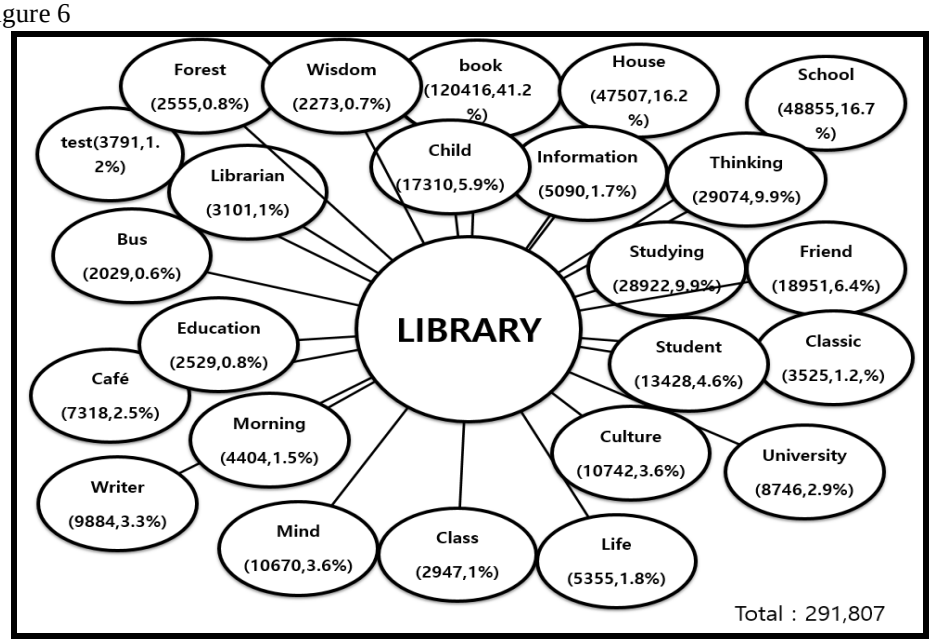

Figure 4. Words related to library – nouns

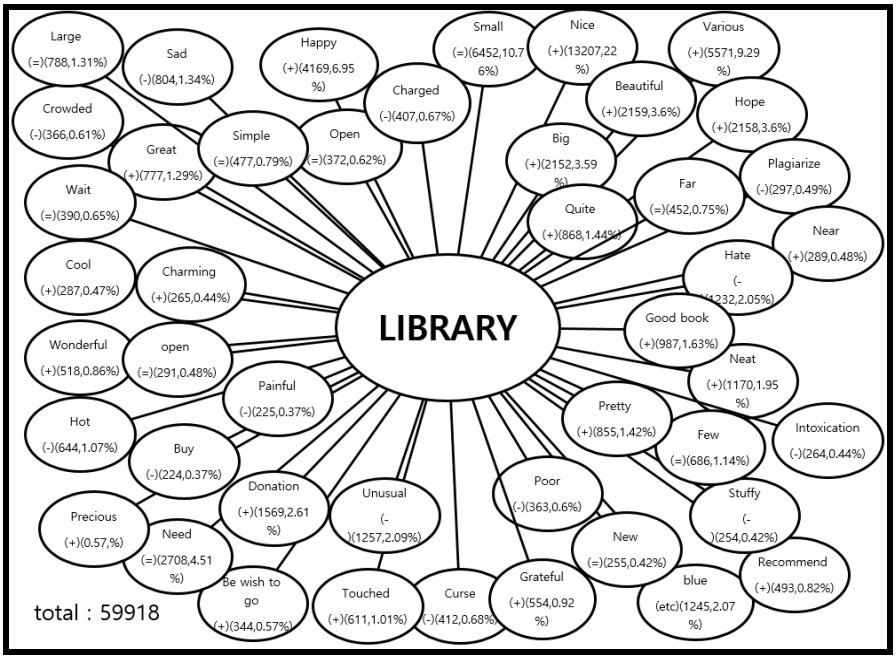

Figure 5. Words related to library – adjectives

. By analyzing the association rules, the study found that users who selected the internet and data retrieval are more likely to select the study room, with a reliability of 56% and a support of 18.67%. Also, users who selected music room, book & magazine are likely to choose study room with reliability is 50% and support of 15.77% as shown in Figure 6.

From Figure 6 we find that there is a very high correlation among the rest room, the study room, and the locker. Among them, the study room is closely related to other factors. The width of the line between two services or facilities represents the strength of their relationship between two objects. The thicker the line, the stronger their relationship is. In this case, it is effective to arrange the study room, the internet space, and the rest room to be adjacent together or close to each other. Among them, the study room has a high correlation with other elements, and it is a facility that should be in the center of the library. The study room, the Internet, and the rest room are shown in bold, and the study room is also linked to the coffee shop. These highly associated services are better to be in the close distance together in the center of the library. Post offices, souvenir shops alone with the others are better to be arranged in the outskirts.

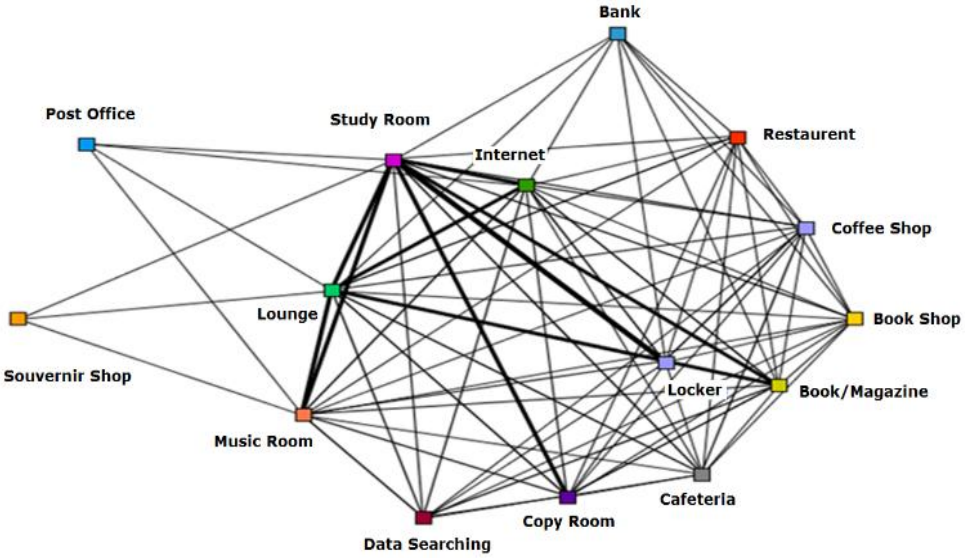

Figure 6. Graph showing relationships among services and facilities in a library

IV. CONCLUSION

Big data techniques became the important tools for researchers and for business applications. This is the beginning of a new era where big data mining will help our daily lives. Big data will continue to increase over time in the future, and researchers in data mining area are expected to manage much bigger data and information. The data they handle will be more diverse and more complex. This study suggests an approach that provides insights on the application of big data such as a social data and data mining techniques to designing building.

This study, specifically, suggests data mash-up approach to design. As one of this solution, the study suggests the features required for the library using text-mining. Also, by using the opinion mining technique, we find the design concept on building, which provides important design concept. Finally, using the data mining techniques such as an association rule analysis based on survey, the study presents a functional layout of services and facilities in the libraries. The study provides a framework to design building with big data and multiple sets of data. This study has several contributions and also limitations.

Firstly, by analyzing the preference of the new complex space demanded by the modern library, the users' preference in services and facilities, in addition to the traditional services and facilities, are found to be relaxation room, study room, and music room.

Secondly, from the adjectives for a library design concept, libraries should be neat, beautiful, happy, precious, wonderful, ecstatic, new, blue, and cute.

Thirdly, the analysis from the survey data shows a visualized graph that provides optimal layout of services and facilities in the library.

This study has limitations on its data. It uses two types of data: social data and survey data. Future study can use more diverse data for more complicated data mash-up. A data set showing users' voices is expected to be explored from diverse data sources. It is also more desirable to use new artificial intelligence techniques included in the future studies.

REFERENCES

1. Mohammed N, Fung B, Wang K, Hung PC. “Privacy-preserving data mash-up,” Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology. ACM. 2009, pp. 228-239.
2. Barhamgi M, Benslimane D, Ghedira C,Gancarski, AL. “Privacy-preserving data mash-up,” Advanced Information Networking and Applications (AINA), 2011 IEEE International Conference on. IEEE. 2011.
3. Rahman S, Masud M, Hossain M, Alelaiwi A, Hassan M, Alamri A. “Privacy preserving secure data exchange in mobile P2P cloud healthcare environment,” Peer-to-Peer Networking and Applications. 9(5), 2016, pp. 894-909.
4. Wang G, Ng TS, Shaikh A. “Programming your network at run-time for big data applications,” Proceedings of the first workshop on hot topics in software defined networks. ACM. 201, pp. 103-108.
5. Gantz J, Reinsel D. “Extracting Value from Chaos,” IDC IVIEW. 2011, Jun, pp. 1-12.
6. McAfee A,Brynjolfsson E. “Big-Data: The Management Revolution,” Harvard Business Review. 2012, 90(10), pp. 60-66.
7. Shadkam M, O'Hara J. “Social Commerce Dimensions: The Potential Leverage for Marketers,” Journal of Internet Banking and Commerce. 2013, 18(1), pp. 1-14.
8. Young M. The Technical Writer’s Handbook. Mill Valley. CA: University Science. 1989.
9. James M, Michael C. “Big data: The next frontier for innovation, competition and productivity,” Mckinsey Global Institute. 2011, pp. 1-36.
10. Egidio LT, Adilson AB. “A Strategy for Mining Association Rules Continuously in POS Scanner Data,” ECIS 2000 Proceedings. 2000, pp. 1-6.
11. Wang C, Zhang P. “The evolution of social commerce: an examination from the people, business, technology, and information perspective,” Communication of the Association for Information Systems. 2013, 31(5), pp. 105-127.
12. Chiu I,Shu LH. “Using language as related stimuli for concept generation.,” AI EDAM. 2007, 21(2), pp. 103-121.
13. Salehan M, Kim DJ. “Predicting the performance of online consumer reviews:A sentiment mining approach to big data analytics,” Decision Support Systems. 2016, 81, pp. 30-40.
14. Minanovic A, Hrvoje G, Zivko K. “Big data and sentiment analysis using KNIME: Online reviews vs. social media. Information and Communication Technology,” Electronics and Microelectronics (MIPRO). 37th International Convention on IEEE. 2014.