Master of Philosophy

Designing a System Prototype for Construction Document Management Using Automated Tagging and Visualization

August 2015

Department of Civil & Environmental Engineering
The Graduate School
Seoul National University

Shin, Yoonjung

Designing a System Prototype for Construction Document Management Using Automated Tagging and Visualization

지도교수 지 석 호

이 논문을 공학석사학위논문으로 제출함

2015년 8월

서울대학교 대학원

건설환경공학부

신 윤 정

신윤정의 석사학위논문을 인준함

2015년 8월

위 원 장 _____ (인)

부위원장 _____ (인)

위 원 _____ (인)

Abstract

Designing a System Prototype for Construction Document Management Using Automated Tagging and Visualization

Shin, Yoonjung

Department of Civil & Environmental Engineering

The Graduate School

Seoul National University

A large amount of text data have been accumulated over time in the construction industry. Important and useful information collected from previous construction projects as experience is mainly recorded in document form. Such information can be used as best practice for upcoming projects by delivering lessons learned for better risk management and project control. Thus, text-based information plays an important role in business strategy development in the highly competitive construction industry. To experience benefits from this text-based information, practical and usable text data management systems are vital.

A significant amount of construction text data are rarely utilized for new construction projects because of the difficulty in accessing them. As the technology that can handle text data has been developed a number of document management systems have been proposed based on text mining techniques. However, most of them focus on classifying documents and they are unable to deal with construction documents' complex and diverse features. In addition, unnecessary time and energy is still wasted to skim the whole database in order to uncover data of interest. Lastly, because the majority of research has focused on English data – with only a few studies using Korean data – there are plenty of constraints to applying existing English-based systems to Korea's domestic construction industry.

Thus, a construction document management system was designed to manage text data effectively and efficiently, and to activate data and information transfer among system users in the domestic construction industry. To achieve this a system prototype was developed. The proposed construction document management system comprises data collection, data processing, and automated document tagging and dataset visualization.

About 25,000 Korean Internet documents were collected to develop the system prototype using a web crawler. Collected data were processed by using text mining techniques, including POS tagging, to calculate the weight of each term in a document. Each term was clarified using a construction corpus which was also developed in this study. Five keywords were automatically extracted and tagged for each document and a tag's sub-dataset was visualized as a form of wordcloud based on the processed data.

The proposed system prototype was evaluated both qualitatively and

quantitatively by surveying ten experts. Questionnaire scores on the significance of the system's results, the usability of and the need for the proposed system design were all above four on a five-point scale. Moreover, on the quantitative evaluation, estimating the accuracy of the system's results, the accuracy of the proposed system prototype was 84 percent on average. Thus the evaluation results confirm the potential for and feasibility of the proposed system.

Keywords: Construction Document, Document Management System, Automated Tagging, Automated Visualization, Text Mining

Student Number: 2013-23150

Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Research Background

A large amount of text data have been accumulated over time in the construction industry (Hjelt and Björk, 2006; Ma et al., 2011). For example, in the case of design change/variation documents, about 7,200 documents are distributed during any single project (Craig and Sommerville, 2006). Text data generated in the construction industry can be categorized into two groups based on their sources: construction sites and websites (the Internet). Text data from construction sites are collected and stored on the organization's database, and mainly consist of periodic progress reports, quality and safety reports, claim documents, contract documents, design change documents, tender documents, meeting minutes, e-mails, specifications, and others (Rubin et al., 1999; Chun, 2001). Construction-related documents housed on the Internet can include sporadic documents, such as news articles, editorials, interviews, white papers, academic papers, and case study analyses.

Text data are very important in the construction industry for project-related decision-making. Figure 1.1 illustrates the massive growth worldwide of corporate data, showing that nearly 80% of information generated by companies is in document form (Cleveland, 1995; Gantz and Reinsel, 2012). Furthermore, important and useful information collected from previous projects

as experience is mainly held in document form in the construction industry (Pathirage et al., 2007; Soibelman et al., 2008). Such information can be used as best practice for upcoming projects, delivering lessons learned for better risk management and project control. Thus, text-based information plays an important role in business strategy development in the highly competitive construction industry (Song et al., 2009; Qady and Kandil, 2010).

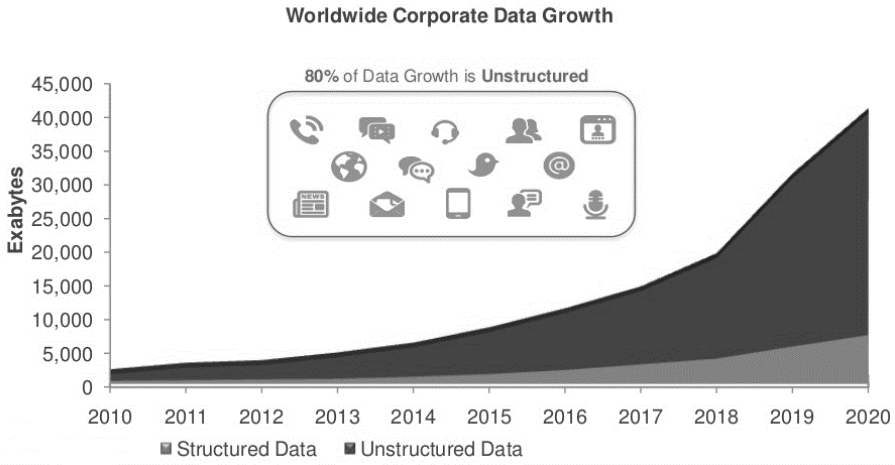

Figure 1.1 Massive Growth of Data (Gantz and Reinsel, 2012)

To experience benefits from text-based information, practical and usable text data management systems are needed. The required capability for text data management, which includes both generating and collecting data capacities, has rapidly increased while keeping the pace with the increased number of information technology (IT) applications in the construction industry (Soibelman et al., 2008). This growth implies that more and more valuable text

data will be generated. Moreover, it has been demonstrated that the need for an effective and efficient construction document management system will continue to increase, in order to analyze the data and provide insights for project-related decision-making (Qady and Kandil., 2013; Chi et at., 2014).

1.2 Problem Statement

Both the need for document management systems and related research thrive in the construction industry. However, there are still several constraints on the proper use of document management systems in the field. First of all, a significant amount of construction text data is rarely utilized for new construction projects. This problem exists because of the limited accessibility of documents in the database and the low opportunity for information reuse (Forcada et al., 2007). Furthermore, text data sources present some additional challenges such as: the lack of infrastructure for text data management systems (Go, 2013); and the properties of construction documents, including unstructured data form, complexity, and numerous types of documents (Soibelman, 2008). Thus, there is a high demand for text data management systems that can improve the utilization of construction data.

As the technology that can handle text data has developed, more and more studies on construction document management systems have focused on a document's contents rather than a document's metadata, such as filename, extension, file directory, category, and date. A number of document

management systems have been proposed based on text mining techniques over the last 10 years, and a majority of studies have applied data mining techniques such as classification or clustering (Qady and Kandil, 2014). However, classes and clusters are mutually exclusive, which means classes or clusters cannot represent the whole construction document's characteristics. To be specific, if a claim document is generated as a result of a fire accident during a construction project (e.g. Figure 1.2) this document will be classified into just one class (e.g. Figure 1.3, Class B, Claim title: Not foreseeable physical obstructions) or one cluster (e.g. Figure 1.4, Cluster 7) when applying a classification or clustering technique. However, document management systems need to reflect other terms (e.g. insulation, fire, insurance, etc.) because they are also key elements of the document.

◆ 이행항변 사례 연구 : 물리적 불가항력

- 사건 번호 : 서울민사지방법원 93 나 17668(1993. 11. 16)
  원심) 서울민사지법 92 가단 98232
- 원고 : (주) 안국화재해상보험(보험회사)
- 피고 : (주) 극동단열(하수급인)

(주) 대도건설이 (주) 경인냉동실업으로부터 인천시 중구 항동 7가 71 의 1번지 지상에 10 층 규모의 냉동 창고 신축 공사를 도급받고, 이중 단열 공사 부분을 803,000,000 원에 피고회사에게 하도급 주었다. 한편 (주) 대도건설은 보험업자인 원고와의 사이에 이 공사에 대한 화재 보험 계약(계약금 : 1,500,000,000 원) 을 체결하였다.

한편 피고회사가 단열공사를 시공하던 중 피고회사 직원의 중대한 과실로 인해 화재가 발생하였고, 그 결과 302822 17 원 상당의 손실이 발생하였다. 보험계약에 의거 원고는 대도건설에게 보험금으로 208222 76 원을 지급하였다. 위 보험금 지급 후 원고는 위 화재가 피고회사의 사무집행상의 중대한 과실행위에 의한 것으로서, 특별한 사정이 없는 한 피고회사에 대해 불법행위 또는 채무불이행으로 인한 손해배상채권을 취득하였다고 주장하며, 보상금 208222 76 원의 지급을 청구하였다.

이에 대해 피고는 대도건설과의 하도급 계약시 공동으로 화재 보험에 가입하고 화재 발생시 복구와 손해 부담 등의 사항에 대해 공동으로 책임지기로 약정하였으므로 피고는 보험 계약상의 제3자가 아닌 피보험자임을 주장한다.

이 사건에서 법원은 비록 그러한 약정이 존재하긴하나 보험계약서 상에는 영백히 대도건설만이 피보험자임으로 피고의 항변을 기각하고, 원고에게 위 보상금의 지급을 판결하였다.

- 항변 당사자 : 피고
- 항변 수용 여부 : 기각

Figure 1.2 Example of Construction Document: Claim Document

(CERIK, 2001)

Class	Claim title	Number of docs
A	Suspension of the works	22
B	Not foreseeable physical obstructions	13
C	ADP required changes	7
D	Remobilization of site office variation	5
E	Explosive detection system variation	10
F	Handrail height variation	7
G	Site fence variation	7
H	Custom exempted material delays	6

Figure 1.3 Example of Construction Document Classification

(Qady and Kandil, 2013)

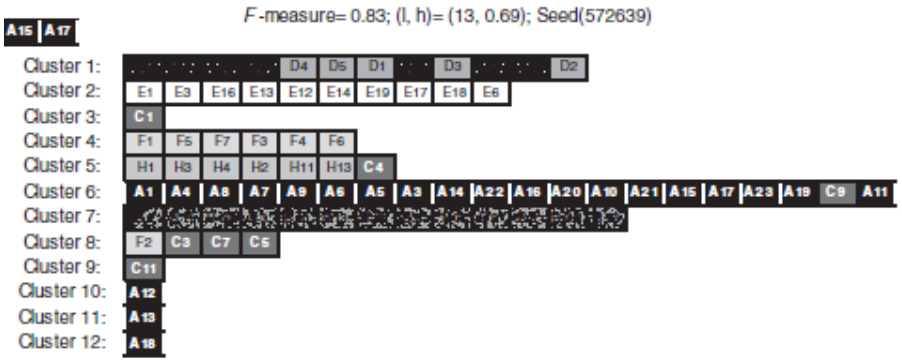

Figure 1.4 Example of Construction Document Clustering

(Qady and Kandil, 2014)

In addition, efficient data organization prior to data and information system service is extremely important because it creates a base for good performance data and the information service system. Data organization is a task which is usually time consuming and requires specific manual effort. It

always takes place during the database skimming phase in order to identify data of interest (Adeva et al., 2014). Thus, an improved methodology for efficiently organizing text data and reducing unnecessary work is needed for document management systems.

Lastly, the majority of research related to document management systems using text mining techniques use English data – only a few studies have been carried out using Korean data. Because Korean differs a lot from English there are plenty of constraints when applying existing English-based systems to Korea's domestic construction industry. Thus, research needs to be carried out on Korean-based document management systems.

1.3 Research Objectives

The primary objective of this study is designing a construction document management system to manage text data effectively and efficiently, and to facilitate the transfer of data and information among system users in the construction industry. Thus, the construction document management system proposed in this study is based on the concept of automated tagging and visualization. Automatically attached tags not only represent a document's specifications but also play a primary role in the proposed tag-based system, contributing to managing construction documents effectively and efficiently. The visualized part of a tag's sub-dataset helps to provide a summary of and insight into a dataset, facilitating the exchange of information.

The specific objectives to achieve the primary objective are as follows:

1. Investigate text data collection approaches to utilize dead data.
2. Develop an automated tagging methodology to represent a document's specifications; this is going to be a novel methodology in the document management system.
3. Propose a preliminary implementation platform using the application of visualization to provide a summary of and insight into a dataset.

1.4 Research Scope

This study focuses on designing and developing a construction document management system prototype and examining the feasibility of the proposed system. Thus, the system prototype was designed and developed keeping expandability of the system in view. More specifically, the system prototype was developed as a web-based system utilizing online Korean text data for a pilot case study prior to developing the whole system.

This paper is organized as follows. Related literature, including document management and principal technologies, is reviewed in Chapter 2. A system prototype design for construction document management is proposed in Chapter 3, and Chapter 4 describes the implementation process of the system

prototype design along with the prototype's evaluation. The paper concludes with a summary of the study's contributions and future research opportunities in Chapter 5.

Chapter 2. Literature Review

2.1 Document Management

As the proposed system prototype is for efficient and effective construction document management, studies on document management were reviewed. General document management systems were covered on the part of document management system. Then, document management systems in the construction industry were focused on the next part.

2.1.1 Document Management System

As about 70% of the whole business time is consumed for document management related tasks, the work efficiency is highly decreased if document related tasks are not carried out smoothly (Cho, 2002). Thus, document management systems (DMS) have been developed for document management lifecycle including document generation, obtainment, storage, utilization, and disposal. DMS has evolved into electronic document management systems (EDMS) by the development of IT which changes paper-based documents into electronic documents such as image file, MS-word file, and PDF file. Not only conglomerate but also small and medium enterprises are mostly managing documents by using commercial or in-house EDMS (Park and Kook, 2014).

An EDMS can be defined as a set of components including documents'

collection, processing, storing, and distribution to support the business activities and project-related decision making (Ahmad, 2000), and main functions are composed of version control, access control level, document management, history management, document retrieval, and viewer (Kim, 2004). Detailed descriptions of each component are as below.

1. Version control: All updated documents are saved in a new version.
2. Access control level: EDMS controls modification and access authorities of each user
3. Document management: Document management function consists of document classification system and life-cycle managements of documents.
4. History management: All documents' tracking from generation to disposal are possible.
5. Document retrieval: Document retrieval function consists of conditional search, contents search, multiple servers search, and Internet/Intranet search.
6. Viewer: Viewer function provides a view of retrieved documents.

As the primary objective of EDMS is to search document of interest with speed and accuracy, the component of document retrieval is most important. However, there are some essential prerequisites for a good-performance document retrieval such as access control level, document management, and viewer.

A number of researches especially focused on the function of document

management have studied, because improvement of document management affects to the other components of EDMS. In addition not only classification but also the other data mining algorithms such as clustering, keyword extraction and tagging were applied and proposed for document management in increasing number of studies (Caldas and Soibelman, 2003; Mao et al., 2007; Zhu et al, 2007; Colace et al., 2014; Qady and Kandil, 2014).

2.1.2 Document Management in the Construction Industry

In the 1980s and early 1990s, fundamental researches such as construction project work structure analysis for computerization, EDI (Enterprise Data Model) application for the efficiency of the construction industry, and materials management computer system development were mainly studied in order to build EDMS. From the late 1990s to early 2000s, Internet was also a main issue of studies on construction document management, such as data and information sharing on the web, and web-based XML/EDI utilization. Recent studies mainly focus on how to find and search construction data and information efficiently and effectively on the web (Park et al., 2013).

To keep pace with the rapidly changing trends in the construction document management, there have been some attempts to manage the construction documents efficiently and systematically in Korea. KICT (Korea Institute of Civil Engineering and Building Technology) has been operating CALS (Continuous Acquisition & Life-cycle Support) which had been developed to exchange and share data and information of the construction project during overall process including planning, design, construction, and

maintenance with the owner and associated companies. CITIS (Contractor Integrated Technical Information Service) is an online network system that provides the digitalized information related to the construction procurement. KISCON (Knowledge Information System of Construction Industry) is established by MOLIT (Ministry of Land, Infrastructure, and Transport), because there had been a need for the methodical construction information management system and timely policy that can cope with the rapid increase in construction projects and documents. In the private sector, PMIS (Project Management Information System) have been developed mainly by conglomerates and information-solution companies, and PMIS contains the function of sharing construction project documents by registering them via online (Park et al., 2013).

2.2 Crawling

Web crawler is a robot which automatically finds out resources of a specific topic by navigating the web on demand and maintaining indexes of the web (Pinkerton, 1994). Web crawlers are used for various purposes these days. First of all, web search engines use web crawlers to assemble a corpus of web pages, index them, and allow users to issue queries against the index and find the web pages that match queries. Web archiving is the other usage of web crawlers where large sets of web pages are collected periodically and archived for posterity. The third use is web data mining, where web pages' log data are

collected and analyzed with statistical properties, or the other data mining algorithms are applied to the collected data. Lastly, web monitoring services let the clients to submit queries, or triggers, and they continuously crawl the web and notify clients of pages that match those queries (Olston and Najork, 2010). The basic architecture of web crawling is detailed below (Alag, 2008) and illustrated as Figure 2.1.

1. A seed URL is given to a web crawler. (seed URL : a site URL that the web crawler is scheduled to visit for the first time)
2. The web crawler requests the other URLs to crawl for the next time to the queue, and checks termination conditions which is contained in crawling code. (Termination condition depends on the purpose of crawling, for example the crawler stops when the number of crawled page reaches the target value, the crawling run time exceeds the designated time, or there is no more URL to crawl.)
3. The web crawler receives the next URL to visit from the queue.
4. The web crawler accesses to the target URL and downloads contents by target page's HTML investigation, and a list of visited URLs is stored separately.
5. An intentional delay time is added to avoid the site's access prohibition caused by overload due to the crawling execution.

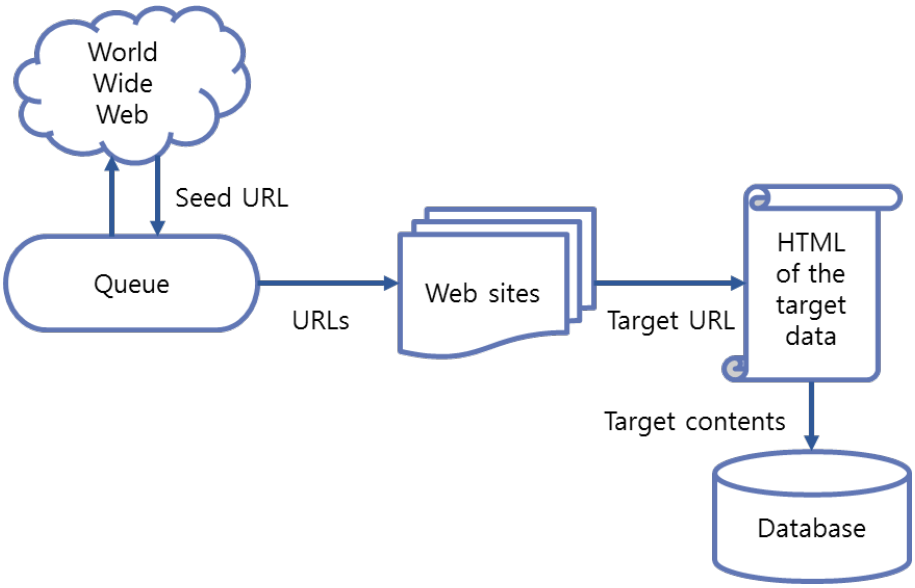

Figure 2.1 Basic Architecture of Web Crawling

2.3 Text Mining

Text mining is a technique based on NLP (Natural Language Processing). Natural language processing is the process of converting natural language text data to the form that computers can understand. However, natural language is lexically and grammatically unique, and also the expression form of natural language is diverse and complex. Thus it is difficult to make a uniform rule which can be applied regardless of language. Therefore, although NLP has been studied since computer was developed, many challenges still remain in the field of NLP technology because of the complexity of natural language (Park, 2004).

Natural language processing is already being used in various fields of

study and industry. First of all, information retrieval is a typical example of NLP. On the process of indexing documents to be searched, search engines mostly do the morphological analysis which is one of the basic NLP technologies. Automatic translation also utilizes NLP technology on the process of analyzing the structure and meaning of sentences and generating a sentence in the other language by converting the syntax of the sentence.

There are some pre-processing procedures commonly applied when executing natural language processing to make computer understand natural languages better. Key pre-processing procedures are tokening, Part-of-Speech tagging, stemming, and stopword removal.

1. Tokening: Separating the boundaries of the words in a given sentence
2. Part-of-Speech (POS) Tagging: Judging each word's part of speech from a given string of words
3. Stemming: Finding the root of the word
4. Stopword removal: Removing unnecessary words for future analysis

Text mining is a kind of technology that finds out valuable and meaningful information from unstructured text data. Users usually obtain more information than the result of information retrieval, because meaningful information is extracted from a vast amount of documents, or linkages with other documents are identified when using text mining. The performance of text mining will improve to discover hidden information in the documents with a massive language resources such as a high quality dictionary, and well-designed algorithm.

Text mining regarding to document management is mainly used for document classification, document clustering, information extraction, and document summarization. As in case of construction document management system, document classification and document clustering accounts for a great part of related researches, classification and clustering also has been reviewed. The basic concepts and algorithms of these applications are same as data mining's.

2.3.1 Classification

Classification is mapping attribute set x to class attribute y based on a function and this function becomes a model called classifier. Such classifiers, predict categorical class labels. The classification approach is generally described as two-step process, consisting of a learning step and an execution (classification) step. In the learning step as the first step, a classifier model is built reflecting a predetermined set of data classes or concepts. In the execution step as the second step, the model is used for classifying test dataset and the designer determines if the model's accuracy is acceptable, and if so, the classification model is accepted for classifying a new dataset (Han et al., 2012).

Applications of general classification includes fraud detection, target marketing, performance prediction, manufacturing, and medical diagnosis. One of the oldest discipline that has used documents classification is bibliography. Librarians have manually classified books according to the established classification system by grasping the contents of each book to classify and manage numerous books in a library. As development of IT had enabled the

production and distribution of a vast amount of documents, it became impossible to classify and manage documents in a conventional manner. However, automation of document classification has been recognized as a very difficult task because most documents are intricately related to each other, and each have heterogeneous content and format.

2.3.2 Clustering

Clustering is the process of finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups. Dissimilarities and similarities are calculated based on the attribute values describing the documents and often involve distance measures such as minimum distance, maximum distance, and group average. As the partitioning is automatically by clustering algorithms, clustering is useful in that it can lead to discover previously unknown groups (clusters) within the data. Clustering is applied in many areas such as biology, security, business intelligence, and web search (Han et al., 2012).

Clustering in text mining is gathering documents which are highly correlated or have similar contents or format based on contents investigation through NLP. Cluster enables users to explore a dataset of documents efficiently as documents have been clustered by its relevance order based on clustering algorithm. To be specific, similarity is calculated through the extraction of key features of documents and comparing them to the other documents', then documents are generally clustered with high similarity documents. Various statistic-based and rule-based clustering algorithms have been studied for

accurate similarity calculation and clustering.

2.4 Visualization

Visualization in these days includes information visualization, scientific visualization, visual design, and information graphics (Friendly, 2008). The basic concept of visualization, creating images, diagrams, or animations to communicate, has been always same, but naming is slightly differently depending on the purpose or application fields.

Visualization through visual imagery has been an effective way to communicate both abstract and concrete information. In addition, visualization is spotlighted in a flood of information, because visualization can provide a number of data discriminately in a limited space. It also helps the readers to understand the information intuitively, and this enabled intuitive reasoning through visualization helps readers to develop a story that would have been hard to find if given in text or numerical values (Oh and Kang, 2008).

Statistical graphics, thematic cartography, and wordclouds are typical examples of visualization. Specifically, wordclouds are a way to visualize words in a document by calculating the frequency of the words appeared in the document. As the words that come out frequently are displayed larger, readers are able to grasp the important part of the document at a glance (Jung, 2013).

2.5 Summary

As numerous text data have been accumulated over time in the construction industry and text data management systems have played an important role in successful performance of construction projects, a number of researches to improve construction document management systems have been studied.

Specifically, related studies were mainly focused on the document management, a component of electronic document management system (EDMS), because it is one of the most important components of EDMS and is a fundamental function which affects to the other functions' performance.

Thus, this study also focused on how to manage construction documents efficiently and effectively, by overcoming existing limitations using techniques of crawling, text mining and visualization. Crawling is to utilize accumulated dead data in the construction industry, text mining is to reflect and utilize the content of the document automatically, and visualization is to give an intuitive insight to the system users.

Chapter 3. A System Prototype Design for Construction Document Management

The construction document management system presented in this study has been designed to manage construction documents effectively and efficiently, and to facilitate data and information exchange among system users. To achieve the established objective, the proposed system design focuses on three major problems of existing construction document management systems. These problems and solutions are described below.

Firstly, there is still a large amount of data which is hard to use for other projects, despite related technologies such as data mining having rapidly improved. Moreover, it is much more complicated in the case of Korean data because to date little research has used Korean data. Thus, by exploring existing data and collecting data in a processed form, we can attempt to utilize the dead data for a new or different construction project.

Construction document management systems have evolved with the development of IT and related research; however, some limitations still remain. To overcome some of the limitations of existing construction document management systems tagging and visualization is proposed in this study.

Secondly, the tag-based system structure is designed to cope with the multi-attributes of construction documents, which have been managed mutually exclusively. The tagging functionality is expected to significantly improve the

effectiveness of the document management system by covering more of a document's attributes. Figure 3.1 shows that processing data using tagging increases data accessibility and reuse than simply using raw data.

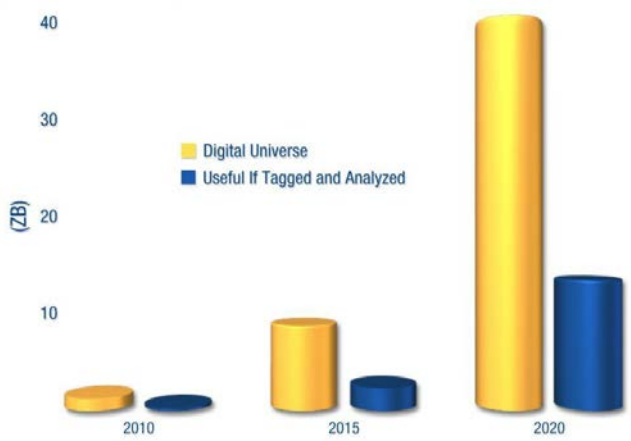

Figure 3.1 Opportunity for Data Processing and Tagging

(Gantz and Reinsel, 2012)

Lastly, it is hard to understand the documents' contents in a short period. Visualization improves the flow of information among the system users by presenting summaries of each tag's sub-dataset contents. This functionality reduces unnecessary time and energy consumption by helping users to capture essential information quickly and efficiently.

The proposed construction document management system is designed to overcome specific problems, as illustrated in Figure 3.2. Each box is a module which constitutes the whole system, and each module operates within the system in the designed sequence. The three study objectives are related to each

module. After investigating text data collection approaches, text data were collected through the selected data collection method. Following this, a tagging methodology to represent each document's specifications was developed, which included two modules: data processing and automated document tagging. Lastly, with two modules – dataset visualization and service – a preliminary implementation platform using visualization is proposed.

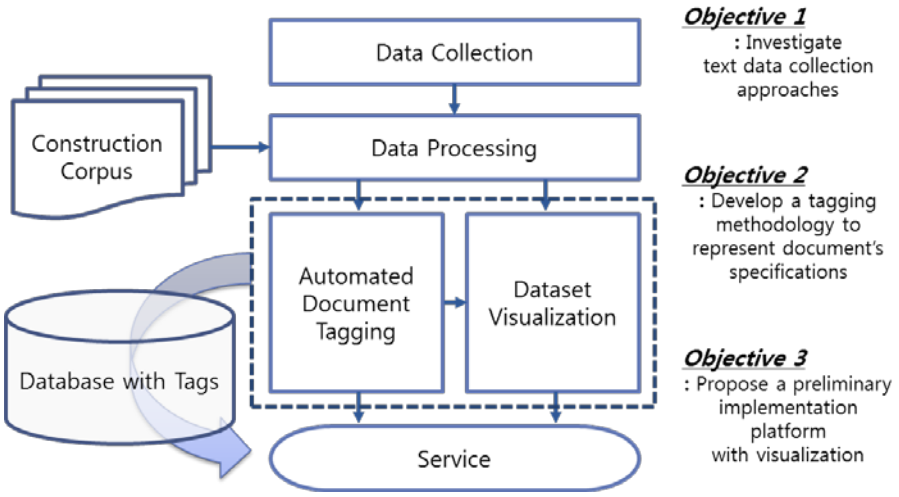

Figure 3.2 Proposed System Design for Construction Document Management

The construction corpus used in this study is a type of dictionary based on knowledge within the construction domain. Words that are emphasized in the construction industry are given a higher weight, while relatively less important words receive a lower weight. The construction corpus was added to improve the performance of the data processing module. Processed documents with tags are stored on the database and the system services from the database with tags.

Boxes within the dashed line (automated document tagging and dataset visualization) are automatic modules, while the others surrounded by the solid line (data collection, data processing and construction corpus) are semi-automatic modules.

The system design development process follows four main steps: 1) data investigation and collection, 2) data processing for tagging and visualization, 3) automated document tagging, and 4) dataset visualization.

3.1 Data Investigation and Collection

To build the construction document management system's database the construction text data source needs to be investigated and the data must briefly be explored. Specifically, the system designer needs to know what data are available, which means investigating the existing data including: data type, data form, contents, and data management condition. Some of the available data is explored in order to capture the data's features prior to further analysis. Through these processes the data collection approaches were selected according to data characteristics.

Text data generated in the construction industry can be divided into two groups according to its source: construction sites and websites (the Internet). Text data from construction sites were analyzed through the literature review, while construction-related text data on the Internet were explored directly. This study focused on Korean text data not only to verify the application of Korean

text mining techniques but also in order to contribute to the domestic construction industry.

3.1.1 Text Data from Construction Sites

A huge number of documents are generated on a construction site and these document are not usually available to the public for reasons such as security. Thus, text data from construction sites were explored through the literature review.

Most construction site documents are text data and include: periodic progress reports, quality and safety reports, claim documents, contract documents, design change documents, tender documents, minutes of meetings, e-mail, appendix, specifications, and weather records (Rubin et al., 1999; Chun, 2001).

However, text data from construction sites are not user-oriented, which means that it is difficult to understand or capture the document's information without processing raw data. This occurs because the data type, data form and the content differ from project to project (An, 2003).

3.1.2 Construction Related Text Data on the Web

With the growth of the Internet the volume of online data has increased dramatically. Construction-related data has also grown, with a considerable number of construction-oriented websites.

Construction-related data is generated on a diverse number of websites, mainly on construction-oriented sites including KISCON and ICAK

(International Contractors Association of Korea). The type of data available on the Internet mainly consist of: news reports, editorials, interviews, reports, official documents, papers, and case study documents. A summary of the construction-related documents found on the Internet is shown in Table 3.1, specifically it shows the relationship between the websites and data type. The same type of websites generate documents in the same data form (for example Internet community websites generate interviews and reports); in other words, each type of website has a different data form. Similarly, the same data types mostly deal with similar content themes, meaning the same data types have a similar content structure (for example case study documents have a similar content structure regardless of the type of website they originate from).

Table 3.1 Summary of Construction Related Documents on the Web

Data Type / Web sites	news	editorials	interviews	reports	official documents	papers	case study documents
Construction oriented site	●	●	●	●	●	●	
Access to journals platform site			●			●	●
Government-affiliated organization site	●			●	●		
Internet community			●	●			

Even though construction-related text data on the Internet provide sufficient information and are easy to access, these data have not been utilized on actual construction sites. In particular, the volume of the data available on the Internet is similar to that of a construction site's data, and the velocity and variety of the data on the Internet is much more diverse. Thus, if such data were fully utilized it is expected it would have a significant impact on the

construction industry. Above all, as this type of open source data is suitable to use to build the prototype system's database we decided to use web crawler to collect the data.

3.1.3 Text Data Collection Approaches

Internet data were determined to be suitable for the system prototype development because of their characteristics, such as data availability, low data usability, and data type variety. Thus, construction-related text data on the Internet were chosen for the target dataset and crawling was selected as the data collection method used to develop the system prototype.

Crawling was executed using the following five steps (see Figure 3.3). At first, web pages containing construction-related text data were selected, and their URLs were collected by investigating the web page's structure (usually HTML). Following this, the coded crawler accessed each data page via the URL (obtained in the previous step). Prior to saving the data on the accessed page the coder investigated the page structure (also usually HTML) to 'crawl' the exact data required. Finally, the targeted data were saved into a '.json' format by the crawler based on python.

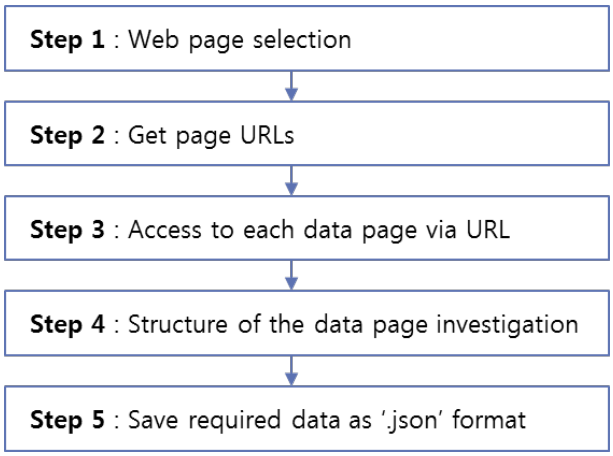

Figure 3.3 Web Crawling Steps

3.2 Data Processing for Tagging and Visualization

The data processing module is designed to represent a document's or a sub-dataset's specifications as well as the designed algorithm can. The designed data processing module includes three automated steps and four manual input steps that need to be followed (see Figure 3.4). Raw data is input data (crawled data in the case of the system prototype), and keyword extraction is the result of the data processing module. POS tagging, term frequency calculation, and filtering are automated processes, and dictionary selection, optimum morpheme selection, filter selection, and the keyword selection are manual processes. From the viewpoint of coding based on python, manual inputs are applied on the neighboring automated step.

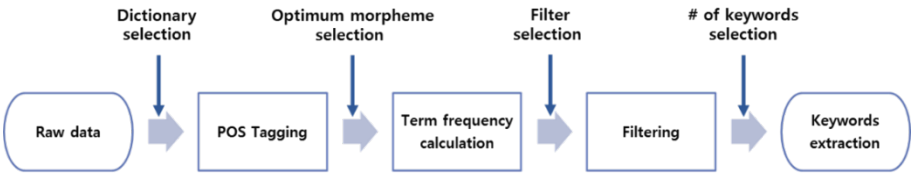

Figure 3.4 Data Processing Steps Using Text Mining Techniques

The detailed process is described as follows. First of all, a dictionary has to be selected considering the project's purpose before POS tagging, which results in considerably different performance at the keyword extraction stage. POS tagging is followed and executed based on the determined dictionary. Then the optimum morpheme – noun, verb, adjective – is chosen prior to calculating the term frequency. This process can significantly reduce calculation time by decreasing the document term vector dimensions. This is because unnecessary morphemes that they appear from frequently throughout the document when compared to the significant morphemes. Term frequency is calculated only for the selected morphemes. The filtering process, which not only reduces the calculation time but also helps to extract keywords to represent the characteristics of the document, is followed up with a manually selected filter. Finally, the number of keywords to be extracted is decided and that number of keywords are tagged or visualized in the connected modules.

3.2.1 Weight Calculation of Each Term in a Document

(1) POS Tagging

In the case of Korean data, the POS tagging process is usually combined with the morphological analysis, which identifies the structure of morphemes and other linguistic units in a phrase. Thus, POS tagging is the same as the process of marking-up morphemes in a sentence based on their definitions and context (Park, 2015). Thus POS tagging needs to be executed prior to carrying out the weight calculation of each term.

There are several commonly used Korean dictionaries, such as Hannanum, Kkma, and Twitter. The performance of POS tagging differs according to the dictionary used, as shown in Figure 3.5, which describes the POS tagging results for a sentence, for example: “아버지가방에들어가신다”. To be specific, “아버지가방에들어가신다” is a sentence, “DREAMISNOWHERE”, which can be read both as “DREAM IS NO WHERE” and “DREAM IS NOW HERE” as a result of word spacing. “아버지가 방에 들어가신다” means “Father is entering a room”, and “아버지 가방에 들어가신다” means “Father is entering a bag”. Thus, the example shows that word spacing plays a very important role in Korean natural language processing because there are a lot more word spacing issues in Korean writing. Therefore, the dictionary needs to be selected carefully, considering the project’s purpose and the quality and characteristics of the database.

Hannanum	Kkma	Komoran	Mecab	Twitter
아버지가방에들어가 / N 이 / J 시ㄴ다 / E	아버지 / NNG 가방 / NNG 에 / JKM 들어가 / VV 시 / EPH ㄴ다 / EFN	아버지가방에들어가신다 / NNP	아버지 / NNG 가 / JKS 방 / NNG 에 / JKB 들어가 / VV 신다 / EP+EC	아버지 / Noun 가방 / Noun 에 / Josa 들어가신 / Verb 다 / Eomi

Figure 3.5 Example of POS Tagging Results According to Dictionaries (Park, 2015)

The Twitter dictionary was selected for the proposed system prototype, taking into consideration the system's purpose and the feature of the database. Firstly, as the proposed system prototype needed to reduce its time consumption the Twitter dictionary's relatively short loading and execution time was an advantage (Park, 2015). In addition, the Twitter dictionary performs well and is suitable for online text data which sometimes includes typing or word spacing errors.

(2) Term Frequency Calculation

Term frequency is calculated based on the POS tagged data. The term frequency calculation is one of the most important processes because all the other calculations build on it. Moreover, the calculated term frequency becomes the weight of each term. The higher the weight of a term, the higher the possibility of the term becomes, because the term is regarded to be important based on its weight.

The term frequency calculation's algorithm is simple: 1) make a term list

composed of all the terms that appear at least once in the document; and 2) sum the terms which turn up more than twice, where the same term occurs in different places.

Figure 3.6 is an example of a calculated term frequency which includes all kinds of morphemes and shows the top 20 terms as the format of ((term, morpheme), term frequency) in regards to frequency. However, as Figure 3.6 shows, not all the morphemes are meaningful at representing the whole document's contents. For that reason, optimum morpheme selection is needed to improve the term frequency calculation step.

```
Top 20 frequent morphemes:
[((의, J), 398),
 ((., S), 340),
 ((하, X), 297),
 ((에, J), 283),
 ((ㄴ다, E), 242),
 ((ㄴ, E), 226),
 ((이, J), 218),
 ((을, J), 211),
 ((은, J), 184),
 ((어, E), 177),
 ((를, J), 148),
 ((ㄹ, E), 135),
 ((/, S), 131),
 ((하, P), 124),
 ((는, J), 117),
 ((법률, N), 115),
 ((,, S), 100),
 ((는, E), 97),
 ((있, P), 96),
 ((되, X), 95)]
```

Figure 3.6 Example of Term Frequency Calculation (Park, 2015)

Nouns are chosen for the optimum morpheme as the other morphemes have vague explanations when considered alone, while nouns cover most of the documents' contents in the database. Taking into account only the nouns hereafter helps to reduce computation time and helps make the document concise and clear. Term frequency is then calculated using the noun-tagged terms based on the result of POS tagging process.

3.2.2 Clarification with Construction Corpus

More work is needed to clarify the meaning of extracted keywords. Specifically, extracted keywords cannot represent a document's specification if a term only turns up once in that document despite being a real keyword or vice versa (a keyword which appears repeatedly but which is not a keyword at all). To solve these problems the construction corpus is used for clarification in this paper.

This process includes using a manual filter to build the construction corpus and using automated filtering for keyword extraction. Construction domain knowledge is necessary to build the construction corpus because the same terms may be accepted and used differently depending on the domain.

In the case of the system prototype the construction corpus can be divided into two parts: low frequency but of prime importance and high frequency but of little importance. The construction corpus in this study was built using a global country list as the former case and three lists including stopword as the latter case. The global country list plays a part in weighting the country name terms higher, and the three stopword lists also play a part in weighting stopword

terms lower.

As the construction industry is a field-oriented industry, information about the construction site is very important. Furthermore, overseas construction projects are increasing and related data and information are needed (Choi, et al., 2008). Thus, country names were concluded to be a significant clue to identify information related to a specific country even though the frequency is low.

On the other hand, there are meaningless terms such as 1-character terms and excessively general terms in the construction industry, such as '건설(construction)' and '사업(business)'. These terms can therefore be excluded from the keywords. Thus, three filters were used. The first one filters 1-character terms, the second one filters excessively general terms within the whole data, and the third one filters excessively general terms within the country-related data such as '해외(overseas)', '공사(project)'. The cut line number from the term frequency list was determined to be 30 and 35 for each filter, and was based on the domain knowledge.

Filtering using the construction corpus significantly improves the quality of keywords and, with the selected number of keywords, final keywords are extracted. The data processed in this way were then utilized for automated document tagging and dataset visualization.

3.3 Automated Document Tagging

Both automated document tagging and dataset visualization are automatic processes which do not need any manual inputs. Automated document tagging focuses on how to attach effective tags, which represent a document's specifications automatically, and how to build the tag-based system.

A tag on the Internet commonly means metadata that the user directly creates about some information. Metadata in this context refers to a set of user-determined keywords or terms that were concluded to be relevant and appropriate for the information (Lee et al., 2007). The basic concept of a tag is the same as mentioned earlier; however, tags in this study are extracted and attached automatically.

High-performing automation plays a very critical role in the success of a computerized system. In the case of this module, automated document tagging is expected to enable the utilization of dead data which have been accumulated for decades in the construction industry over a short period time, and also reduce unnecessary tasks such as manually classifying or tagging each document to manage the volume of documents that mount up in a day. Furthermore it will contribute to objective information management regardless of the document's generator.

3.3.1 Tags Representing Documents' Specifications

The optimal number of tags needs to be discreetly decided prior to attaching tags to each document. There are both advantages and disadvantages

to having either a number of tags or a few tags. The more tags a document has the more representative the tags will be; however, the poorer the readability of the tags the less representative the tags will be and vice versa.

A total number of five keywords was selected as the optimal number of keywords for each document's tags, considering both the tag's representability and conciseness. Thus five extracted keywords from the data processing module were tagged to each document.

3.3.2 Tag Based System

There are some advantages to using a tag-based system. The first advantage is the representativeness of the tags, which means that the tags attached to a document represent the document's contents effectively and efficiently.

The second advantage is the generation of various classification schemes that reflect each document's specifications. In essence, each tag can become the main content of the documents' group, with documents managed based on their tags, not on a class, cluster, or any specific hierarchies.

The other advantages are: the utilization of the vocabulary of all the documents without prejudice; the reflection of the dynamic nature of terms, which means that the latest term usage is reflected in the tags; and the provision of opportunities for new discoveries with free navigation of information (Lee et al., 2007). An example of a tag-based system from Collavate is illustrated in Figure 3.7.

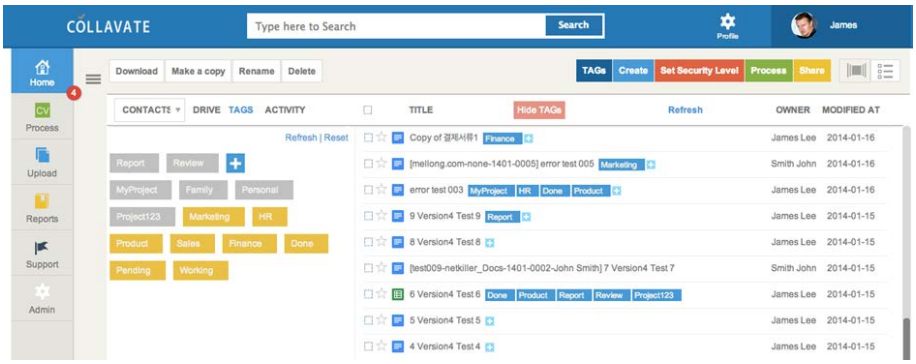

Figure 3.7 Example of Tag Based System (Collavate, 2015)

3.4 Dataset Visualization

Dataset visualization focuses on how to draw a wordcloud, which represents a tag's sub-dataset and how to build a visualization-based system. Visualization, which is often used interchangeably with the term 'infographics', is a rapidly growing research area. Good data visualization provides information about a large amount of data in a single view (Choi and Kim, 2012). Thus, the visualization function is presented in this paper to help users absorb information effectively and efficiently.

Using the processed data in the data processing step it is easy to extract keywords representing each document, thus they are suitable input data for visualization. To visualize a tag's sub-dataset the top 20 keywords are extracted for the first step. All 20 keywords from each sub-dataset are gathered and each term frequency is added up. A wordcloud of each tag is visualized based on the

term frequency summation data. However, as the term frequency varies from dozens to hundreds, the calculated weight (term frequency) of each term is normalized. The normalized weight of each term helps to make the wordcloud clearer because each word's size and thickness becomes relative compared to the others. Figure 3.8 describes the wordcloud drawing steps.

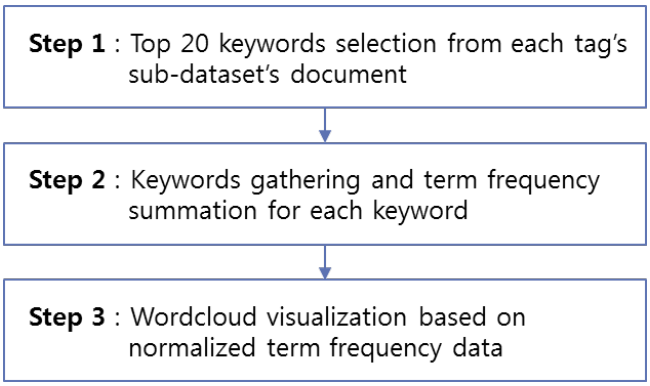

Figure 3.8 Wordcloud Drawing Steps

3.4.1 Wordcloud Representing a Tag's Sub-dataset

Attributes such as size, thickness, and colors are used to represent the characteristics of the words when drawing a wordcloud (Halvey and Keane, 2007). In most cases high-term frequency terms are presented in larger and bolder words (Lee et al., 2007). Figure 3.9 is an example of a wordcloud drawn with the database of Tweets containing 'Apple'. Naturally, 'apple' was the most frequent term and 'iphone', 'store', 'jobs', 'app', and 'macbook' follow.

The wordcloud drawing algorithm of this study is the same as most cases; if a term has appeared foremost, that term is drawn as the biggest and boldest

in the wordcloud.

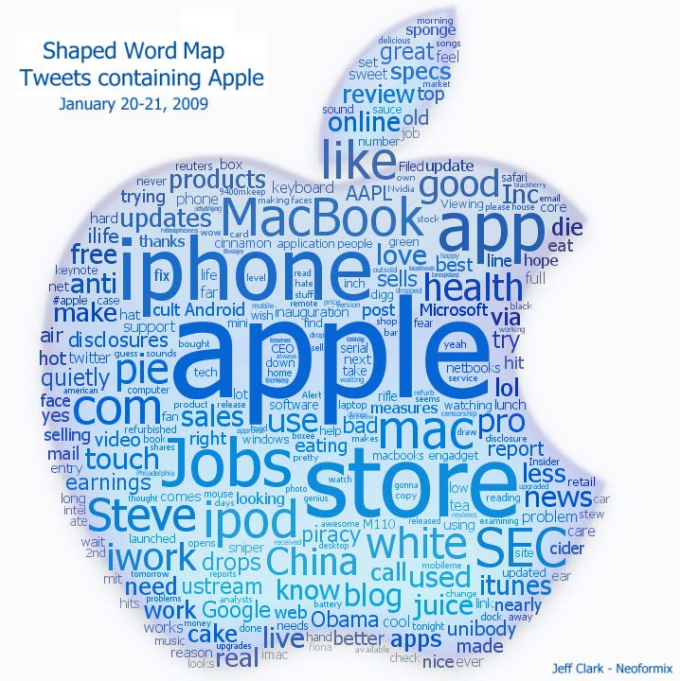

Figure 3.9 Wordcloud Example – Tag: Apple (Clark, 2009)

3.4.2 Visualization-based System

The purpose of visualization is to transmit information effectively and efficiently. Above all, the visualized image is expected to reduce any unnecessary time taken to grasp the information when reading a number of documents, because it is the summary reflecting all the contents of the dataset. Thus, the visualization is a user-oriented system with simpler, clearer and faster data and information provision.

A system prototype was developed to implement the proposed design for the construction document management system. The prototype system architecture is illustrated in Figure 3.10. Crawler was used for data collection and database building, and a construction corpus was developed and applied for data processing for the system prototype design. The other procedures are the same as the proposed whole system design.

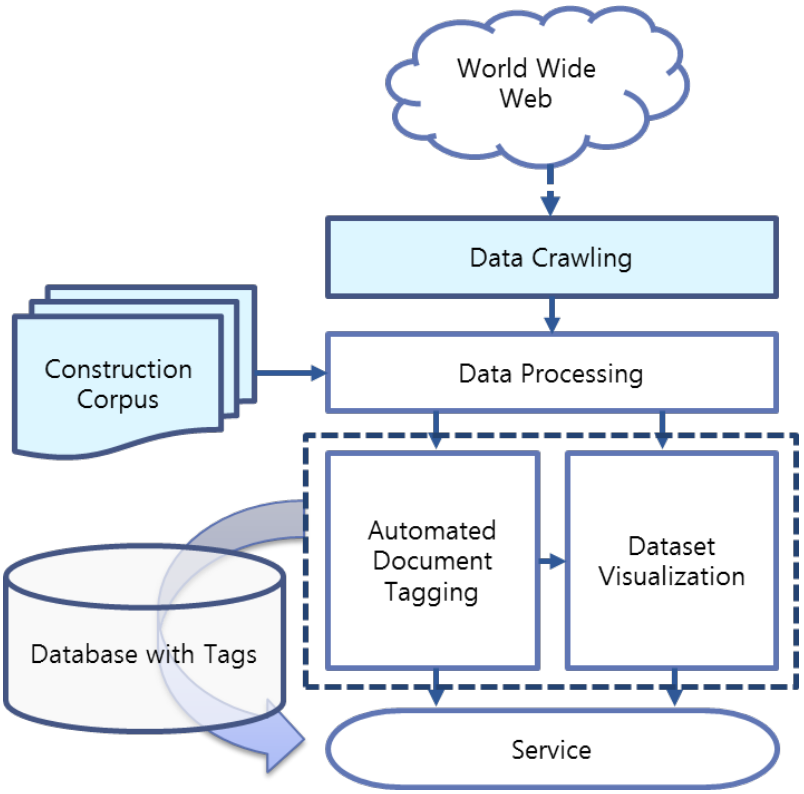

Figure 3.10 Proposed System Prototype Design

Chapter 4. System Implementation and Evaluation

The designed construction document management system consists of four modules, and each module's design was implemented by developing a system prototype. The implementation result of each module is described in this chapter. The system implementation and evaluation process follows four main steps: 1) database of system prototype, 2) implementation of data processing, 3) developed system prototype, and 4) evaluation.

4.1 Database of System Prototype

The targeted dataset of the proposed system prototype comprised construction-related documents scattered on the Internet. Specifically, construction-related news, editorials, interviews, reports, and official documents on the Internet were deemed to be appropriate for the dataset because these data types are prolific, are frequently generated and accumulated in a variety of themes and, above all, are treated as being useless.

Six websites were selected for the target web pages, which mostly deal with the mentioned data types identified through the brief data exploration phase. All of the selected websites are construction-oriented websites,

including one construction government-affiliated organization site. Data on these websites have the potential to become a new data source for a new construction project. These websites also satisfied some qualifications for the system prototype database, such as enough volume of the data, proper frequency of data generation, and a variety of data.

As each website has different a web page structure (in the form of HTML) and different data forms crawling codes needed to be differently designed for each website. The crawling procedure is explained in detail below and illustrated in Figure 4.1 with some examples.

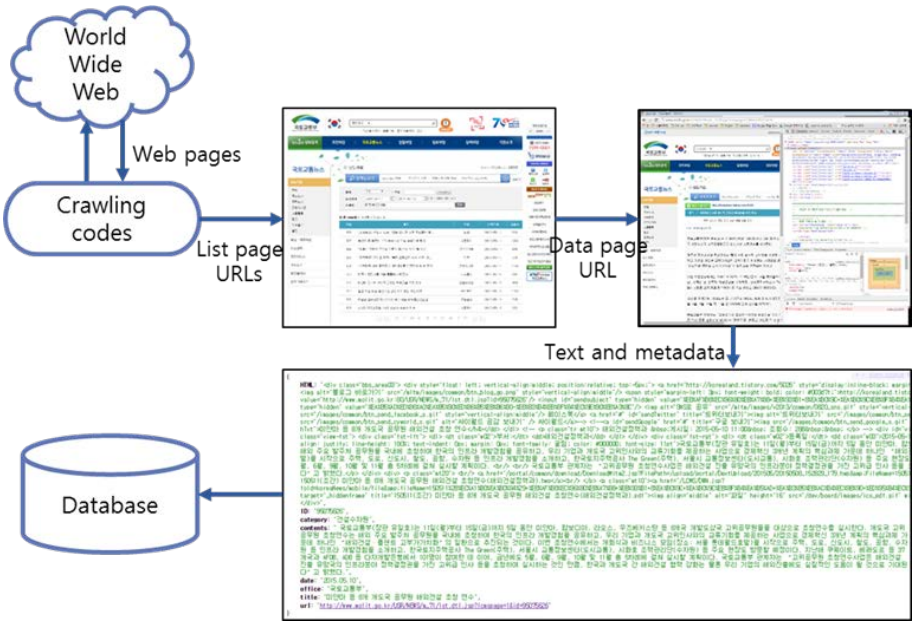

Figure 4.1 Detailed Crawling Procedure

The designed crawler collected URLs of data pages of interest in order to

provide access to each of them; that is, the crawler accesses the target web page, parses the page's HTML and obtains the data URLs of each page until the page ends. The crawler is then able to access the data page with the URL, and the accessed page's structure is then investigated. As the locations of data ID, category, contents, date, office, title, and URL in the web page structure (in a form of HTML) are built in the designed crawler all of these data are saved in a '.json' format and stored on the database. Having built the database in this way it can be fully utilized to build a new database with tags after the data have been processed.

Cdaily, Cnews, Fnnews, KISCON, MOLIT, and Ohmycon websites were selected as target web pages. Cdaily, Cnews, Fnnews, and Ohmycon sites mainly cover the data types of news, editorials, and interviews. KISCON mostly deals with editorials, reports, and official documents. MOLIT generates chiefly official documents and related news.

The date of collected data ranges from the website's inception date to 21/05/2015. The total number of data files for the pilot study is 25,143, which is approximately 279MB as a unit of memory. Detailed information on each data sources follows in Table 4.1.

Table 4.1 Crawled Data Overview Classified by Data Source

Data Source		Number of Data	Web Site Address
Korean	English		
건설일보	Cdaily	474	http://www.cdaily.kr/
건설경제	Cnews	4,242	http://www.cnews.co.kr/
파이낸셜	Fnnews	1,561	http://www.fnnews.com/
KISCON	KISCON	3,804	https://www.kiscon.net/
국토교통부	MOLIT	1,492	http://www.molit.go.kr/
오마이건설뉴스	Ohmycon	13,570	http://www.ohmycon.co.kr/
Total		**25,143**	

4.2 Implementation of Data Processing

The collected Korean text data were processed using text mining techniques, with the keywords of each document being extracted during this step. The data processing process was designed with three automated steps and four manual input steps. In short, manual inputs were selected during the system prototype design process. These are: Twitter dictionary, nouns, construction corpus-applied filters, and the number of keywords. Each manual step has an effect on the neighboring automated step. In addition, each automated step forms the basis for the following steps. Figure 4.2 illustrates the data processing steps with the applied manual inputs.

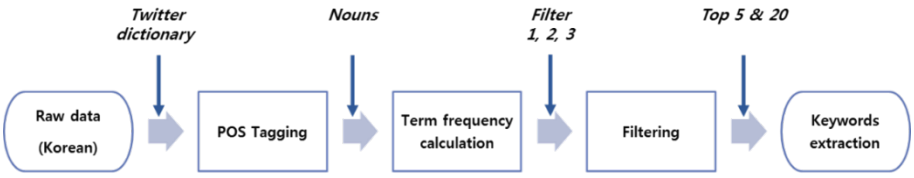

Figure 4.2 Data Processing Steps with Selected Manual Steps

The step involving the weight calculation of each term in a document includes two automated steps: POS tagging and term frequency calculation. The clarification using the construction corpus is executed by filtering using three different filters. Keywords are then derived from the filtered data.

4.2.1 Weight Calculation of Each Term in a Document

(1) POS Tagging

For the first step of weight calculation POS tagging was executed. As KoNLPy is flexible to choose any of the common Korean dictionaries and all have the same input-output structure the KoNLPy package based on Python was selected for POS tagging. The input is a phrase and the output is a list of tagged morphemes with the original terms.

Whole data were POS tagged in the same way and formatted as shown in Figure 4.3. The same document is used to provide an example of data processing and the example document's title is '미얀마 등 8개 개도국 공무원 해외건설 초청 연수 (Public Officials overseas construction invitation training of Myanmar, including the eight developing countries)'. As is show in Figure 4.3, each term was analyzed and appropriate morphemes were

tagged to each split term according to the Twitter dictionary. In addition, it was possible to confirm that the Twitter dictionary was suitable for this study.

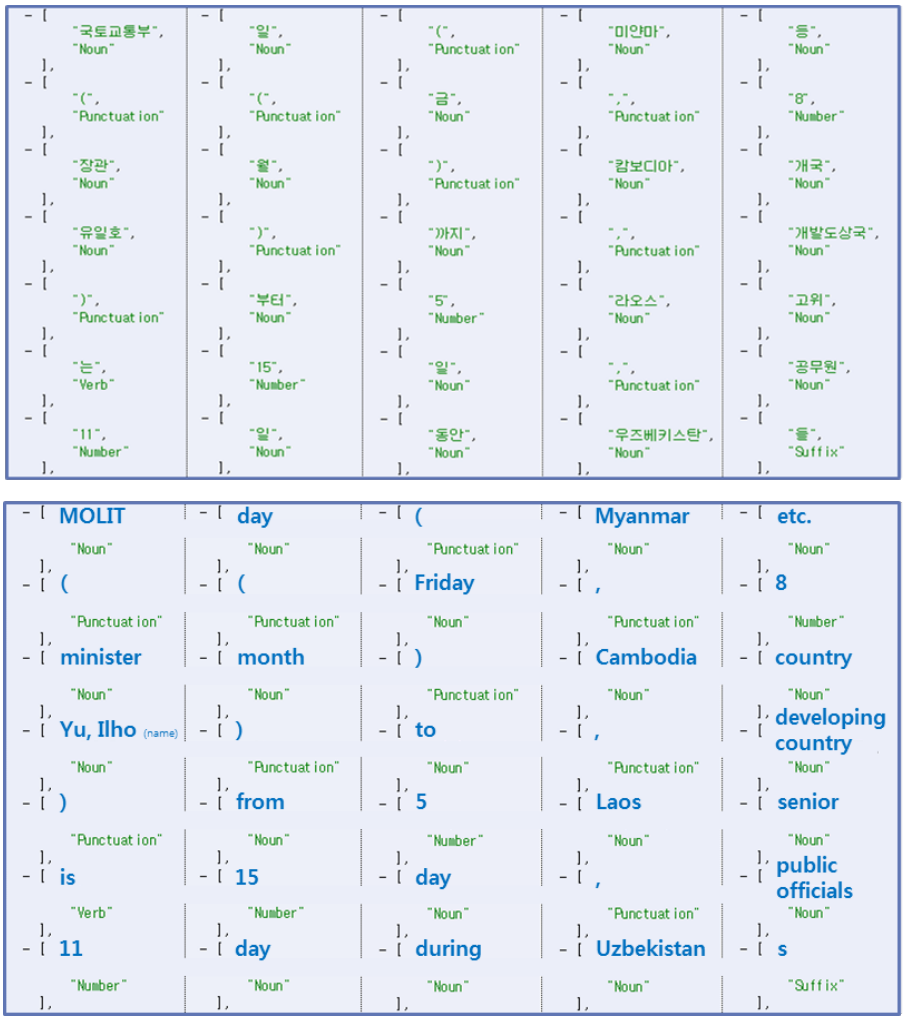

Figure 4.3 POS Tagged Data Example

(2) Term Frequency Calculation

For the next step, the term frequency of nouns was calculated. However, not all the terms showing a high term frequency were important or represent a document's specification, as can be seen in Figure 4.4. Thus a filtering step using construction domain knowledge was necessary.

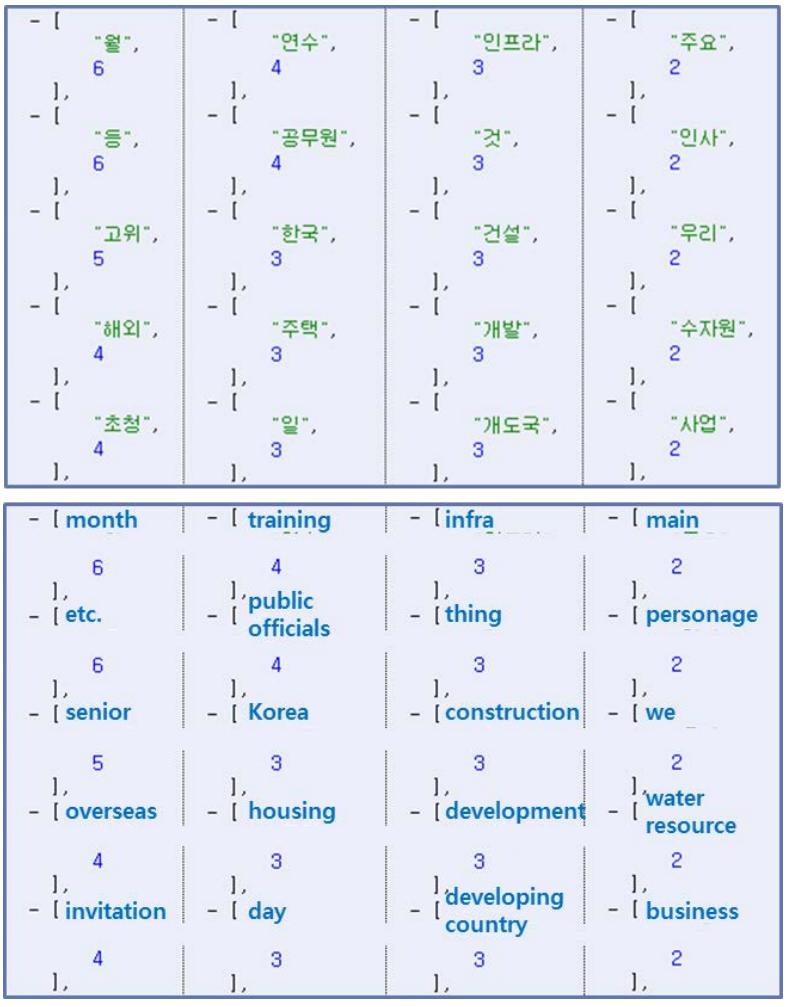

Figure 4.4 Term Frequency Calculated Data Example

4.2.2 Clarification with Construction Corpus

To clarify the meaning of keywords the construction corpus, including a country list and three filters, were applied to the processed data. First of all the country list obtained from Wikipedia was applied. A summary of the number of data including each country name is provided in Table 4.2. A total of 154 country names appeared from the database among 244 countries in the world; one-fifth of the whole data included a country name. If duplicated counting of documents were allowed a total of 10,291 documents related to a country appear in the database.

Table 4.2 Summary of the Number of Data Including Country Name

Data Source		Number of Data	Country (154)	Else
Korean	English			
건설일보	Cdaily	474	50	424
건설경제	Cnews	4,242	607	3,635
파이낸셜	Fnnews	1,561	137	1,424
KISCON	KISCON	3,804	879	2,925
국토교통부	MOLIT	1,492	365	1,127
오마이건설뉴스	Ohmycon	13,570	2,921	10,649
Total		**25,143**	**4,959**	**20,184**

(1) Filter 1

The results of each filter will be shown step-by-step. The first filter removed meaningless 1-character terms such as '월(month)', '등(etc.)', '일(day)', '것(thing)', and '조(trillion)' because most 1-character terms are ineffective to deliver the main contents of documents. Thus, 1- character terms were excluded from each data and Figure 4.5 shows the filter 1 applied result.

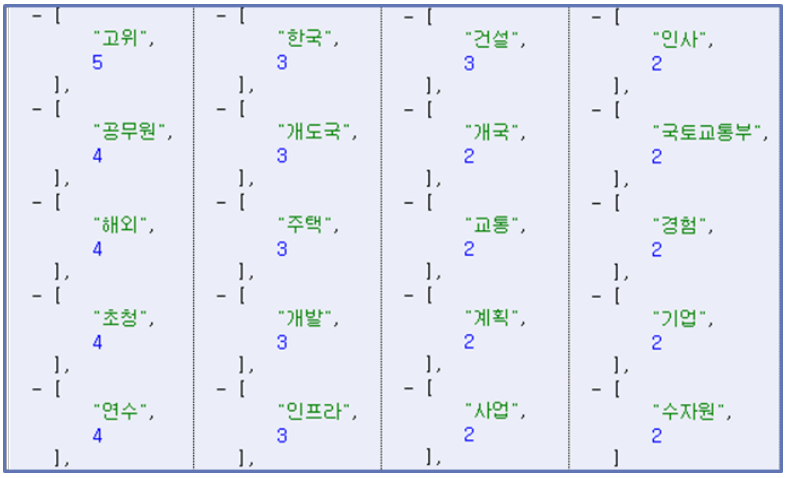

Figure 4.5 Filter 1 Applied Data Example

(2) Filter 2

The second filter removed excessively general terms found within the whole data. The top 30 frequent terms were concluded to be too general to represent a specific document. Thus, the list of the top 30 frequent terms from a total of 46,594 terms (excluding 1-character terms) were excluded from each document. Filter 2 consists of 30 terms including '건설(construction)', '사업(business)', '공사(project)', '주택(housing)', '억원(hundred million won)', '계획(plan)', '업체(enterprise)', '대한(Korean)', '기술(technology)', and '지역(area)'. In short, terms in the filter 2 were eliminated from the result of filter 1 applied data. The results of filter 2 applied data are shown in Figure 4.6.

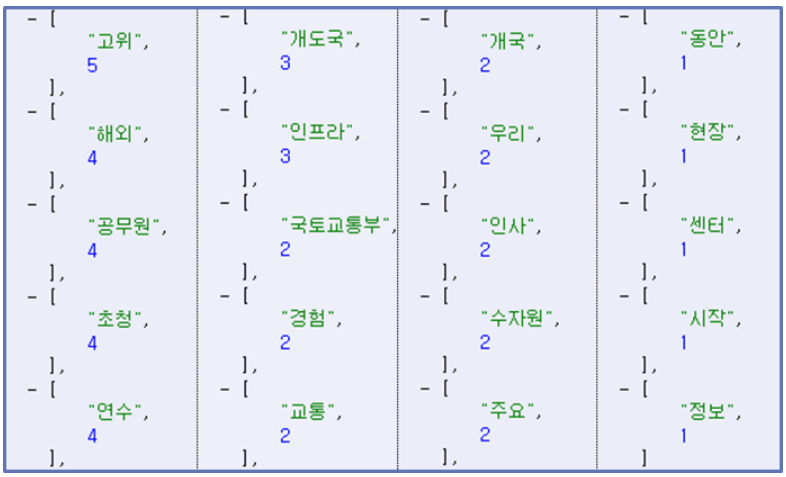

Figure 4.6 Filter 2 Applied Data Example

(3) Filter 3

The third filter removed excessively general terms found within the country-related data. The top 35 frequent terms were concluded to be too general to represent a specific document. Thus, the list of the top 35 frequent terms from a total of 29,254 terms (excluding 1-character terms) were excluded from each document. Filter 3 consists of 35terms including '건설(construction)', '사업(business)', '공사(project)', '기술(technology)', '시장(market)', '대한(Korean)', '산업(industry)', '수주(winning a contract)', '업체(enterprise)', and '주택(housing)'. In short, terms in the filter 3 were eliminated from the result of filter 1 applied data. The results of filter 3 applied data are shown in Figure 4.7.

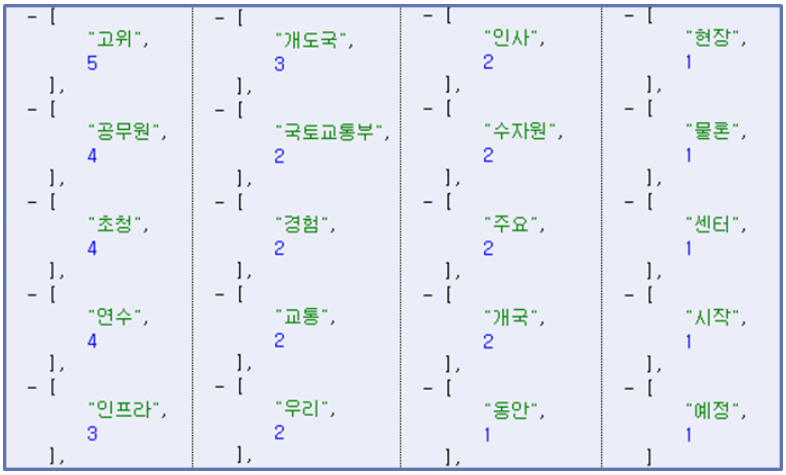

Figure 4.7 Filter 3 Applied Data Example

There was some overlap between filter 2 and filter 3, such as ‘건설(construction)’, ‘사업(business)’, and ‘공사(project)’, because each term frequency list was built independently. A few tests were done comparing filter 2 and filter 3 and it was concluded that filter 3 performed better for the system prototype. Thus keywords were extracted from the filter 3 applied results, five for document tagging and 20 for visualization for each data. Figure 4.8 illustrates the keyword extraction process, and keywords are the top five in the order of calculated weight.

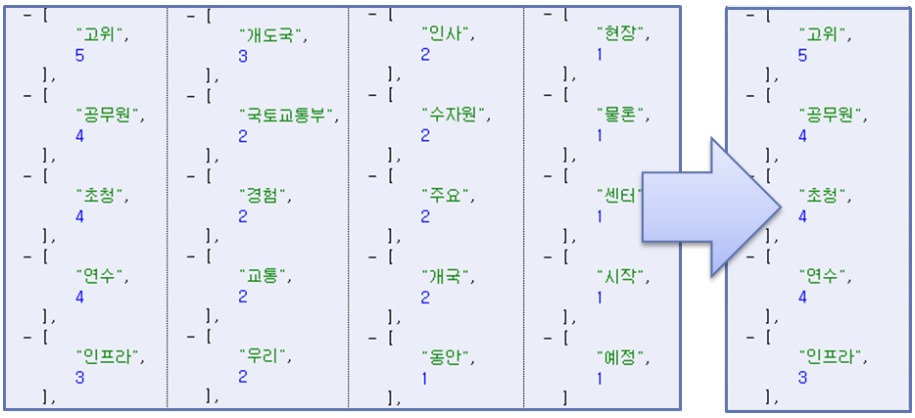

Figure 4.8 Extracted Keywords of Data Example

Figure 4.9 provides an example of a tag's sub-dataset's visualization and illuminates the function of filters. System users find it easy to capture the information of a tag – in this example the tag '미얀마 (Myanmar)' – through a visualized summary of the tag. The information depicted in the picture improves at each step.

From the first picture, the user is just able to capture the information that "the terms of '건설(construction)' and '수주(winning a contract)' are foremost because they are mentioned a lot with Myanmar". However, these terms are not specifically related to Myanmar at all. With the application of filter 1 the picture seems to improve without 1-character terms, but there is still no particularly useful information.

Filter 2 and filter 3 applied examples show much improved results. Users can get information such as "Korean construction industry cooperates with Myanmar a lot, Korean government tries to do ODA projects with Myanmar, or Myanmar construction industry is growing up" through the visualized summary of 'Myanmar' depicted by the largest words, such as '협력(cooperation)', '지원(assistance)', and '봉사(Volunteer)'. Above all, this implication is significant because the visualized picture provides only objective and accurate information.

<Filter 1>

<Filter 2>

<Filter 3>

Figure 4.9 Visualization of a Tag's Sub-dataset Example

4.3 Developed System Prototype

The system prototype using automated document tagging and visualization was developed based on the Internet. Each web page mainly consists of two types: one is a map (Figure 4.10) and the other is a dataset visualized image and tagged data lists (Figure 4.11).

As location-related information is important in the construction industry a country list was added to the construction corpus and is reflected in the system prototype main page. Users can get to a country's dataset by clicking a country on the map, as illustrated in Figure 4.10, or on the list which is located on the left side, as shown in Figure 4.11. In short, users are able to access the summary of a country's information and access each data list by clicking a country on the map or the list.

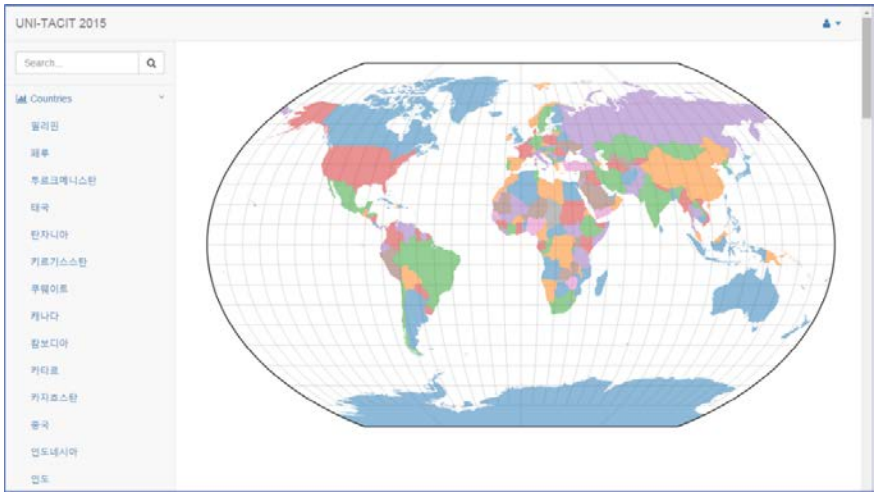

Figure 4.10 System Prototype Service – Main Page

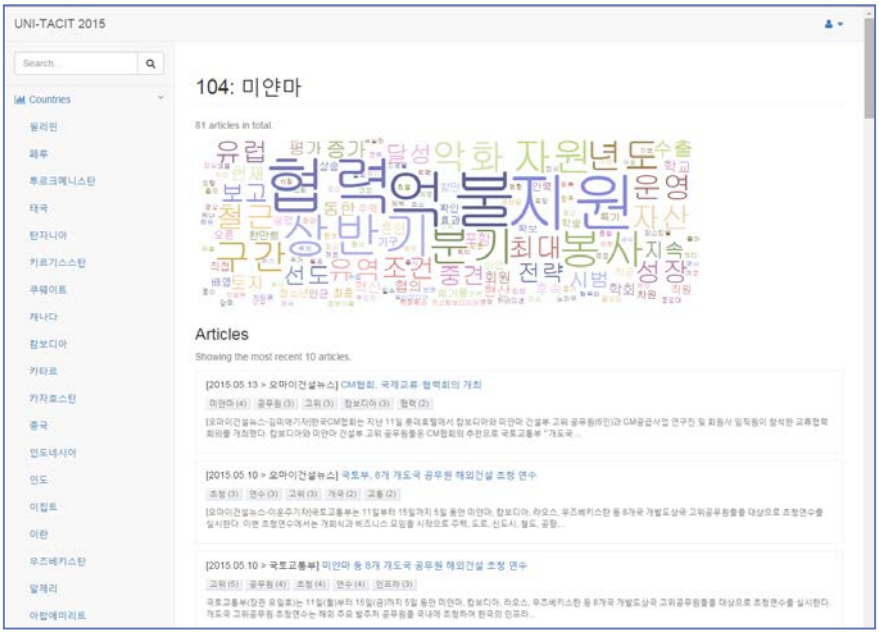

Figure 4.11 System Prototype Service – Tag Page

4.3.1 Implementation of Automated Document Tagging

Automated tagged data are shown below a tag's sub-dataset's visualized image. Five automatically extracted keywords are designated to be tags and they sit below the data title. Only 150 characters from the beginning of the document are available in the system prototype; the whole body of data is available at the original website, which is hyperlinked through the data's title due to the data's copyright. Also, by clicking one of tags users can get the tag's information from the whole database, including the tag-attached data and a visualized summary.

An example of automatically tagged data presented in the system

prototype is shown in Figure 4.12. The example data's title is '미얀마 등 8개 개도국 공무원 해외건설 초청 연수 (Public officials overseas construction invitation training of Myanmar, including the eight developing countries)', and five tags are '고위(senior)', '공무원(public officials)', '초청(invitation)', '연수(training)', and '인프라(infrastructure)'. Users can get simpler, clearer, faster, and higher-quality information from tags, as can be seen in the example.

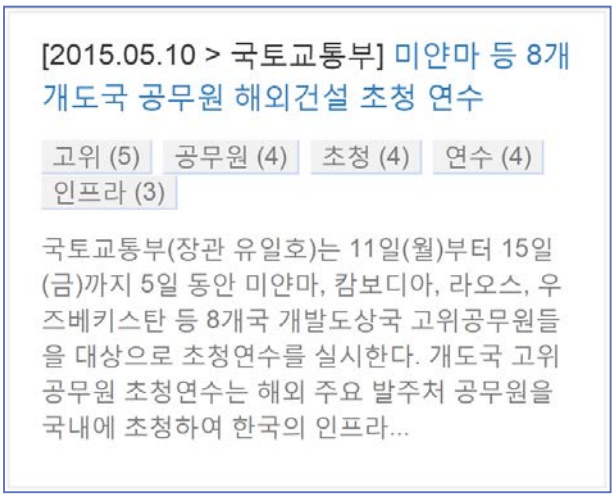

Figure 4.12 Automatically Tagged Data Example

4.3.2 Implementation of Dataset Visualization

A tag's sub-dataset is summarized using a wordcloud through visualization. The selected tag is presented in the left-upper location of the page, the wordcloud comes below the tag with a short description of the total number of articles within the tag, and the list of data with tags follows.

Both the tags in the wordcloud and attached to the data are able to be selected to see more information about the tag. Figure 4.13 illustrates the former case, and Figure 4.14 shows the latter case.

Figure 4.13 is an example of clicking the ‘철근(steel reinforcement)’ tag in the wordcloud of the tag ‘미얀마(Myanmar)’. Terms of note are ‘만원(ten thousand won)’, ‘가격(cost, price)’, and ‘현장(construction site)’. Thus, the Korean construction industry regarding ‘철근(steel reinforcement)’ is interested in its cost at the construction project site.

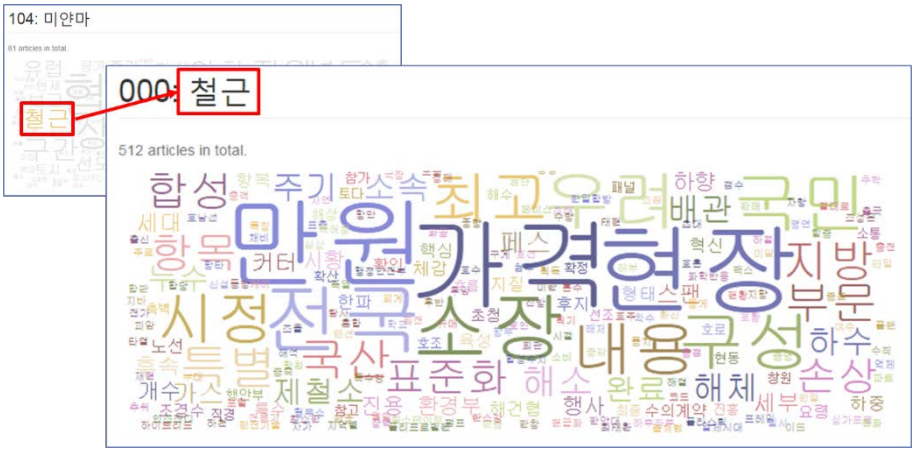

Figure 4.13 Prototype Service of a Tag’s Sub-dataset Visualization 1

Figure 4.14 is an example of the ‘협력(cooperation)’ tag from a news article tag titled ‘물 문제 국제협력·물기업 해외진출 활성화 기대 (Water issues international cooperation · Water related companies overseas market entrance activation expectation)’. The keywords of ‘협력(cooperation)’

seem to be ‘평가(evaluation)’, ‘안전(safety)’, and ‘현장(construction site)’. Thus, the Korean construction industry related to ‘협력(cooperation)’ is mainly concerned with safety evaluation and cooperation on site.

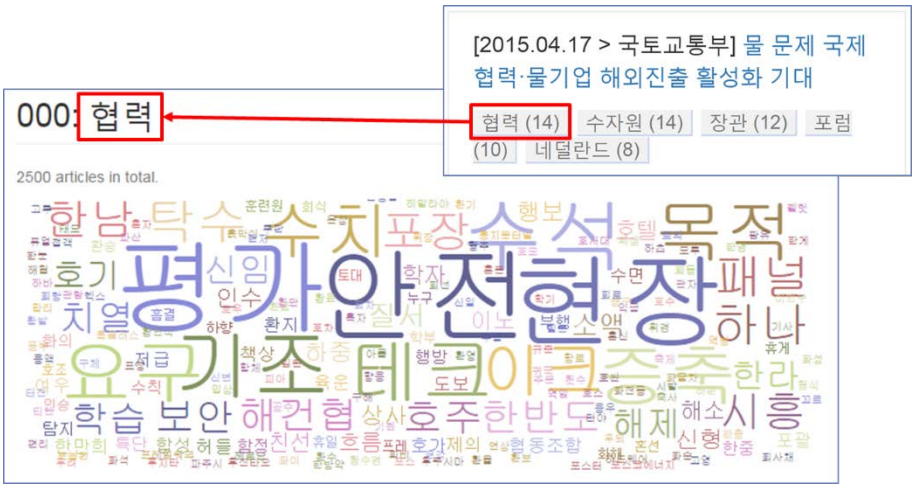

Figure 4.14 Prototype Service of a Tag’s Sub-dataset Visualization 2

4.4 Evaluation

Both qualitative and quantitative evaluation was required to verify the usefulness of the construction document management system prototype proposed in this study. Thus, the proposed system prototype was evaluated qualitatively and quantitatively by surveying a number of experts. The survey was carried out targeting construction experts, after providing a full description of the operation sequence and explaining the performance of the proposed

system prototype. Ten respondents were provided a three-item questionnaire for qualitative evaluation, a five-item questionnaire for quantitative evaluation and were asked for additional comments.

4.4.1 Qualitative Evaluation

Qualitative evaluation focuses on how well the system performs and how much this system is needed on an actual construction site. As system performance mainly depends on the results of two modules, automated document tagging and dataset visualization, the questionnaire comprised three items, with each module's typical example as shown in Figure 4.15. The three-item questionnaire is as follows and is rated on a five-point scale:

1. The significance of automatically attached tags
2. The usability of the system to be developed based on the proposed system prototype in the future.
3. Need for the proposed system.

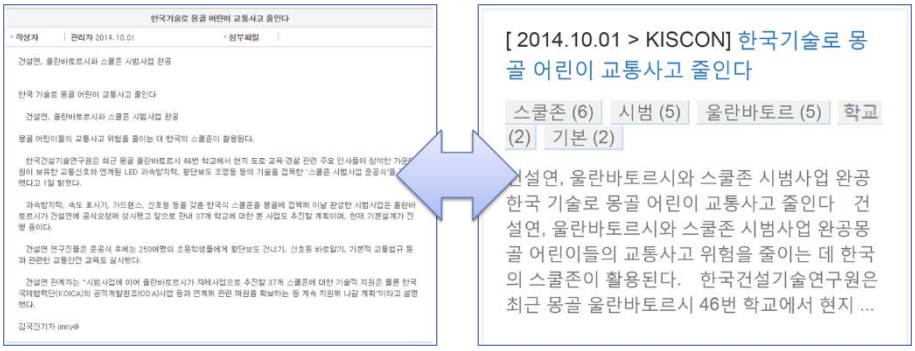

Figure 4.15 A Typical Example of the Automated Document Tagging Module

Scores for each questionnaire were 4, 4.3, and 4.4. The results show that providing processed data with tags and a visualized image is significant compared to providing just raw data. Furthermore, respondents evaluated that an improved system based on the developed system prototype in this study will be useful and, accordingly, the system will be in demand.

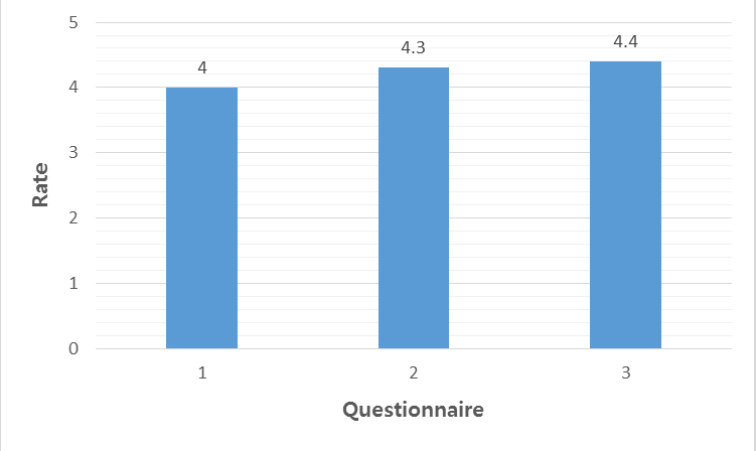

Figure 4.16 Result of the Qualitative Survey

4.4.2 Quantitative Evaluation

Quantitative evaluation was executed by comparing the extracted keywords through the system and the selected keywords by construction experts. A total of five documents were given to respondents and they marked five words which they decided to be the most representative of each document based on their judgment. The five most frequent terms among the gathered answers became manually selected tags (keywords) for each document. These

tags were compared to automatically attached tags by the proposed system prototype.

In the case of documents one, three, four, and five, the coverage of automatically attached keywords among manually selected keywords was 80%, which means that four out of five tags matched. Moreover, the tags for document two matched perfectly; that is, the result for the proposed system prototype is 84 percent accurate on average (see Figure 4.17). These results confirm that automated document tagging's performance is as good as manual tagging; accordingly, it is expected to play a major role in utilizing dead data.

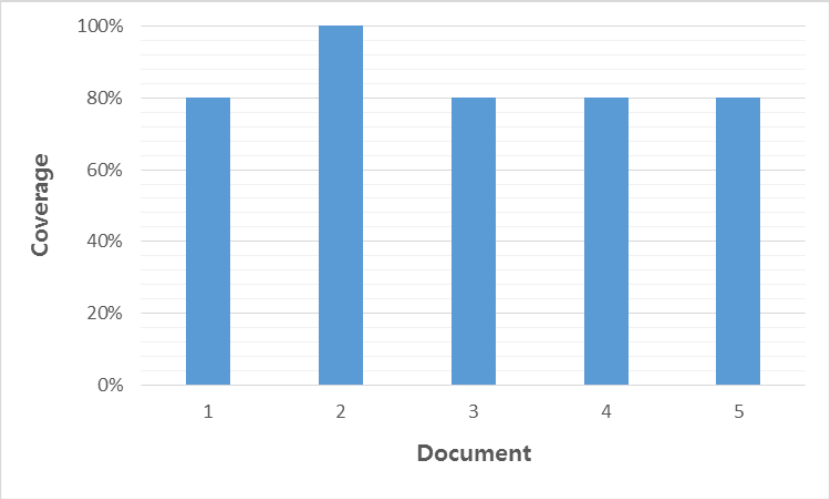

Figure 4.17 Result of the Quantitative Survey

The results of the evaluation confirm the possibility and feasibility of the proposed system. The proposed system was evaluated to be helpful to manage construction documents effectively and efficiently, and to facilitate data and information transfer among system users through both the qualitative and

quantitative questionnaire. In particular, the simplicity of the designed system architecture, the expandability of input data, its usability and convenient service, and ease of system maintenance were highly assessed.

The survey respondents also provided some comments in regards to the system's improvement. The processing of collocation will improve the performance of two main modules because there is a limit to grasping the meaning of the document's content based on the terms of morphological units. In addition, reflecting a tag's hierarchy would help to enhance users' understanding of a document's content.

Chapter 5. Conclusions

5.1 Summary

A construction document management system using automated tagging and visualization was designed, and a system prototype was developed to manage construction industry text data effectively and efficiently and to facilitate data and information transfer among system users in the construction industry.

To develop a system prototype online text data collection approaches were investigated, a tagging methodology to represent document specifications was developed and, finally, a preliminary implementation platform with visualization was proposed. Specifically, data investigation and collection, data processing for tagging and visualization, automated document tagging and dataset visualization were executed.

The developed system prototype overcame some of the existing limitations. The system prototype proposed a novel methodology for document management using a tag-based system design, provided a summary of and insight into the dataset by using visualization, utilized dead data by using online text data as input data, and contributed to Korean-based text mining research by developing a Korean-based system. In addition, the developed system prototype validated the feasibility of the proposed system design based on the

simplicity of the designed system architecture, the expandability of input data, usability and convenient service, and its ease of system maintenance. Thus designed construction document management system will increase reuse and sharing of data and information, improve reading efficiency, and ultimately, improve business efficiency

5.2 Contributions and Future Study

5.2.1 Contributions

This study is significant because it validates a novel construction document management system design based on tags and visualization functionality through the developed system prototype. Automated document tagging and visualization were found to significantly reduce unnecessary time consumption and energy for processing existing data and reading a range of documents to get to their core, and helped the system to provide an insight into the construction industry.

The developed system prototype also illustrated the possibility of applying text mining techniques to the Korean construction industry, as it used text mining techniques based on Korean input data.

Furthermore, the system prototype will provide a base for the information retrieval system's foundation. The proposed system design for construction document management, with automated tagging and visualization functions, will play an important part in the information retrieval system by improving the

effectiveness and efficiency of the system.

5.2.2 Future Study

The proposed construction document management system can be improved by developing each system module. The data processing module is expected to perform better with an expanded construction corpus. Expanding the construction corpus to contain construction-specific synonyms and a thesaurus will enable tags to be further clarified. In addition, if necessary hierarchies are applied to the construction corpus users will be able to understand the relationship between tags much easier and faster (Lee et al., 2007).

The visualization module can be further improved by adopting developed algorithms and additional functions, such as timeframe, keywords category coloring, and keyword relationship mapping. An advanced visualization module can be carried a step further to knowledge extraction.

Finally, an improved construction document management system with automated document tagging and dataset visualization will form part of an information retrieval system or search engine through building an unstructured database.

Bibliography

Adeva, J.J.G., Atxa, J.M.P., Carrillo, M.U., and Zengotitabengoa, E.A. (2014) "Automatic text classification to support systematic reviews in medicine", *Expert Systems with Applications*, 41(4), 1498-1508.

Ahmad, I. (2000) "Data warehousing in construction organizations", *Proceedings of the sixth construction congress*, Orlando, Florida.

Alag, S. (2008) *Collective intelligence in action*, 1st Ed., New York; Manning Publications.

An, S.J. (2003) "Development of document exchange model for construction participants using web-based XML/EDI", Master Dissertation, Seoul National University, Seoul, Korea.

Caldas, C.H. and Soibelman, L. (2003) "Automating hierarchical document classification for construction management information systems", *Automation in Construction*, 12(4), 395-406.

CERIK (2001) *Analysis of the dispute case of South Korea in the construction field*, CERIK.

Chi, N.W., Lin, K.Y., and Hsieh, S.H. (2014) "Using ontology-based text classification to assist job hazard analysis", *Advanced Engineering Informatics*, 28, 381-394.

Cho, H.C. (2002) "A study on the relationship of construction information in

documents control for automatic generation of construction documents", Master Dissertation, Ajou University, Suwon, Korea.

Choi, S.L., Kim, J.H., Jang, S.J., and Paek, J.H. (2008) "A study on the risk factors to strengthen the competitiveness in the overseas development projects – focused on new town development of developing country", *Journal of the Korea Institute of Building Construction*, 8(3), 59-67.

Choi, J.W. and Kim, L.Y. (2012) "A Study on Inforgraphic for Effective Visual Communication of the Big Data Era -Government Departments and Public Institutions", *KOREA SCIENCE & ART FORUM*, 11, 165-175.

Chun, M.J. (2001) "A method for establishing an electronic document management system of construction fields", Master Dissertation, Seoul National University, Seoul, Korea.

Clark, J. (2009) *Apple twitter word map* (http://www.neoformix.com/2009/AppleTwitterWordMap.html).

Cleveland, G. (1995) "Overview of document management technology", *IFLA UDT Core Programme Occasional Paper*, 2, National Library of Canada.

Colace, F., Santo, M.D., Greco, L., and Napoletano, P. (2014) "Text classification using a few labeled examples", *Computers in Human Behavior*, 30, 689-697.

Collavate (2015) *User guide*

(https://docs.google.com/document/d/1ZOO3jnp0ycFMidqQWbZKujbGjTJBGqAjyuSq1yVO7dw/edit?pli=1#).

Craig, N. and Sommerville, J. (2006) “Information management systems on construction projects: case reviews”, *Records Management Journal*, 16(3), 131-148.

Forcada, N., Casals, M., Roca, X., and Gangolells, M. (2007) “Adoption of web databases for document management in SMEs of the construction sector in Spain”, *Automation in Construction*, 16(4), 411-424.

Friendly, M. (2008) “Milestones in the history of thematic cartography, statistical graphics, and data visualization”, (www.datavis. ca/milestones).

Gantz, J. and Reinsel, D. (2012) *The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east*, IDC (www.emc.com/leadership/digital-universe/index.htm).

Go, H.S. (2013) “The science of winning with big data”, Seoul; Easyspub.

Halvey, M.J. and Keane, M.T. (2007) “An assessment of tag presentation techniques”, *Proceedings of the 16th international conference on World Wide Web*, Banff, Canada.

Han, J., Kamber, M., and Pei, J. (2012) *Data mining concepts and techniques*, 3rd Ed., Boston; Morgan Kaufmann.

Hjelt, M. and Björk, B.C. (2006) “Experiences of EDM usage in construction

projects", *Journal of Information Technology in Construction*, 11, 113–125.

Jung, Y.C. (2013) *Big data*, Seoul; Communicationbooks.

Kim, D.J. (2004) "Study on development direction of electronic document management system(EDMS)", Master Dissertation, Ulsan University, Ulsan, Korea.

Lee, K.P., Kim, D.N., and Kim, H.J. (2007) "Tagging technology trends in the Web 2.0 environment", *Journal of Korean Institute of Information Scientists and Engineers*, 25(10), 36-42.

Ma, Z., Lu, N., and Wu, S. (2011) "Identification and representation of information resources for construction firms", *Advanced Engineering Informatics*, 25(4), 612-624.

Mao, W., Zhu, Y., and Ahmad, I. (2007) "Applying metadata models to unstructured content of construction documents: A vie-based approach, automation in construction", *Automation in Construction*, 16(2), 242-252.

Oh, B.K. and Kang, S.J. (2008) *Text book of information design*, Seoul; Ahn Graphics.

Olston, C. and Najork, M. (2010) "Web crawling", *Foundation and Trends in Information Retrieval*, 4(3), 175-246.

Pathirage, C.P., Amaratunga, D.G., and Haigh, R.P. (2007) "Tacit Knowledge

and organisational performance: construction industry perspective", *Journal of Knowledge Management*, 11(1), 115-126.

Park, H.J., Yoon, H.R., and Kook, K.J. (2013) "Knowledge based Automatic Block Management Model for Long Construction Document", *Architectural Institute of Korea Spring Conference*, 33(1), 549-550.

Park, H.J. and Kook, K.J. (2014) "Metadata based information management prototype system of building material", *Journal of the Architectural Institute of Korea Structure & Construction*, 30(5), 109-116.

Park, J.J. (2004) "Development of text mining based clinical decision support system model for cancer staging" Master Dissertation, Yonsei University, Seoul, Korea.

Park, L. (2015) "KoNLPy documentation, release 0.4.3" (http://konlpy-ko.readthedocs.org/ko/v0.4.3/).

Pinkerton, B. (2000) "WebCrawler: finding what people want", Ph.D. Dissertation, University of Washington, Seattle, WA, USA.

Qady, M.A. and Kandil, A. (2010) "Concept relation extraction from construction documents using natural language processing", *Journal of Construction Engineering and Management*, 136(3), 294-302.

Qady, M.A. and Kandil, A. (2013) "Automatic classification of project documents on the basis of text content", *Journal of Computing in Civil Engineering*, 29(3), 04014043.

Qady, M.A. and Kandil, A. (2014) “Automatic clustering of construction project documents based on textual similarity”, *Automation in Construction*, 42, 36-49.

Rubin, R.A., Fairweather, V., and Guy, S.D. (1999) *Construction claims prevention and resolution*, 3rd Ed., New York; Wiley.

Soibelman, L., Wu, J., Caldas, C., Brilakis, I., and Lin, K.Y. (2008) “Management and analysis of unstructured construction data types”, *Advanced Engineering Informatics*, 22(1), 15-27.

Song, G.M., Kim, D.J., and Yu, Y.S. (2009) “Studies on utilization of lessons-learned system on intranet system in architectural design office” *Info Design Isuue*, 16, 59-73.

Zhu, Y., Mao, W., and Ahmad, I. (2007) “Capturing implicit structures in unstructured content of construction documents.” *Journal of Computing in Civil Engineering*, 21(3), 220–227.

초 록

건설 산업에는 건설 프로젝트 현장에서 발생하는 문서뿐만 아니라 웹 상에서 발생하는 건설 산업과 관련된 문서까지 셀 수 없이 많은 문서가 존재한다. 특히 건설 프로젝트가 수행되면서 생성되는 중요하고 유용한 정보는 대부분 문서에 남아있을 뿐만 아니라, 축적된 문서는 건설 프로젝트의 의사결정 및 건설 기업의 전략 수립 과정에서 중요한 역할을 하므로 그 가치가 매우 높다. 따라서 성공적인 건설 프로젝트의 수행 및 전략 수립을 위해 효과적이고 효율적인 건설 문서 관리 시스템은 반드시 필요하다.

그러나 문서의 대부분을 구성하고 있는 텍스트의 특성 상 문서에 대한 접근이 까다로워 기존에 존재하는 수많은 문서들이 제대로 활용되지 못하고 있는 실정이다. 최근 텍스트를 다룰 수 있는 기술이 발전함에 따라 텍스트마이닝 기술을 활용한 건설 문서 관리 시스템이 다양하게 제안되었지만, 대부분 문서 분류를 목적으로 하여 한 문서 내에 다양하고 복잡한 내용을 포함한 건설 문서의 특징을 모두 포괄하지는 못한다는 한계점이 있었다. 또한, 원하는 문서를 찾기 위한 검색어에 대한 아무런 정보가 주어지지 않기 때문에 검색 결과 전체를 훑어보는 데서 불필요한 시간과 노력의 소모가 있음을 확인할 수 있었다. 마지막으로, 관련 연구의 대부분은 영어를 기반으로 이루어져, 국내 산업에 적용이 어렵다는 단점이 있었다.

따라서 본 연구에서는 국내 건설 산업에서 발생하는 문서를 효율적으로 관리하고, 이를 사용자들이 효과적으로 사용할 수 있는

시스템을 설계하고, 가능성 및 타당성 검토를 위해 프로토타입을 개발하였다. 제안한 시스템은 데이터 수집, 데이터 가공, 그리고 자동화된 문서 태깅 및 데이터 집단 시각화로 이루어진다.

시스템 프로토타입 개발을 위해 웹 상에 존재하는 약 2만 5천여건의 한글 문서가 수집되었다. 수집된 문서는 각 단어의 가중치 계산을 위해 POS 태깅 등의 텍스트 전처리 과정을 거쳤으며, 문서 내의 각 단어는 건설 코퍼스를 기반으로 보다 정밀한 가중치를 갖게 되었다. 이렇게 가공된 데이터로부터 각 문서에는 5개의 키워드가 자동으로 태깅되었으며, 각 태그에 대한 문서 데이터 집단은 통합된 주요 키워드를 기반으로 워드클라우드 형태로 시각화되었다.

개발된 시스템 프로토타입은 열 명의 건설 산업 전문가에게 정성적 및 정량적 평가를 받았다. 정성적 평가에서 본 시스템은 시스템 결과에 대한 유용성, 향후 개발될 시스템의 활용도 및 필요성을 묻는 모든 항목에서 5점 척도 4점 이상의 점수를 기록하였다. 또한, 결과물의 정량적인 정확도 파악을 위한 평가에서 본 시스템의 결과물은 평균 84%의 정확도를 나타내어 제안한 시스템 설계는 본격적인 개발을 위한 충분한 가능성 및 타당성이 있는 것으로 판단되었다

주요어: 건설 문서, 문서 관리 시스템, 태깅 자동화, 시각화 자동화, 텍스트 마이닝

학 번: 2013-23150