The 6th International Conference on Construction Engineering and Project Management (ICCEPM 2015)
Oct. 11 (Sun) ~ 14 (Wed) 2015 • Paradise Hotel Busan • Busan, Korea
www.iccepm2015.org

Creating Knowledge from Construction Documents Using Text Mining

Yoonjung Shin[1] and Seokho Chi[2]

Abstract: A number of documents containing important and useful knowledge have been generated over time in the construction industry. Such text-based knowledge plays an important role in the construction industry for decision-making and business strategy development by being used as best practice for upcoming projects, delivering lessons learned for better risk management and project control. Thus, practical and usable knowledge creation from construction documents is necessary to improve business efficiency. This study proposes a knowledge creating system from construction documents using text mining and the design comprises three main steps – text mining preprocessing, weight calculation of each term, and visualization. A system prototype was developed as a pilot study of the system design. This study is significant because it validates a knowledge creating system design based on text mining and visualization functionality through the developed system prototype. Automated visualization was 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.

Keywords: Knowledge Discovery, Construction Documents, Text Mining, Visualization

I. INTRODUCTION

A large amount of text data have been accumulated over time in the construction industry (Hjelt and Björk, 2006; Ma et al., 2011). Furthermore, important and useful information and knowledge collected from previous projects as experience are mainly held in a document form in the construction industry (Pathirage et al., 2007; Soibelman et al., 2008). Such information and knowledge can be used as best practices for upcoming projects, delivering lessons learned for better risk management and project control. Therefore, text-based information and knowledge play an important role for project-related decision-making and business strategy development in the highly competitive construction industry (Song et al., 2009; Qady and Kandil, 2010). Thus, to experience such benefits, a practical and usable knowledge creation system from construction text data is vital.

A significant amount of construction text data are rarely utilized for new construction projects because of the difficulty in accessing and reusing them. As the technology that can handle text data has been developed, there have been efforts to take advantages of documents based on text mining techniques. However, most of them focused on classifying documents and they were unable to deal with construction documents' complex and diverse features. In addition, unnecessary time and energy were wasted to skim the whole database in order to uncover data of interest and absorb information. Lastly, because the majority of research focused on English data, there have been plenty of constraints to applying existing English-based text mining techniques to Korean domestic construction industry. Thus, a knowledge creating system from construction documents was designed to discover knowledge and activate knowledge transfer among system users in the domestic construction industry.

II. A SYSTEM PROTOTYPE DESIGN

A. *Text Mining Preprocessing*

In the case of Korean data, the Part-of-Speech (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.

B. *Weight Calculation of Each Term*

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, in the case of proposed prototype, the calculated term frequency becomes the weight of each term. The higher the weight of a term, the higher the possibility to represent a document of the term becomes, because the term is regarded to be important based on its weight.The term frequency calculation 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.

C. *Visualization*

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 and discover

[1] Master course student, Civil & Environmental Engineering, Seoul National University 1 Gwanak-Ro, Gwanak-Ku, Seoul, Korea, nicky@snu.ac.kr (*Corresponding Author)

[2] Assistant professor, Civil & Environmental Engineering, Seoul National University 1 Gwanak-Ro, Gwanak-Ku, Seoul, Korea, shchi@snu.ac.kr

knowledge effectively and efficiently. Using the calculated weight, it is easy to extract keywords representing each document, thus they are suitable inputs for visualization. To visualize a dataset, the top 20 keywords are extracted for the first step. All 20 keywords from each dataset are gathered and each term frequency is added up. A wordcloud of each dataset is visualized based on the term frequency summation data. However, as the term frequency varies from dozens to hundreds, the calculated weight of each term needs to be 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.

III. RESULTS AND DISCUSSIONS

A. *Database of System Prototype*

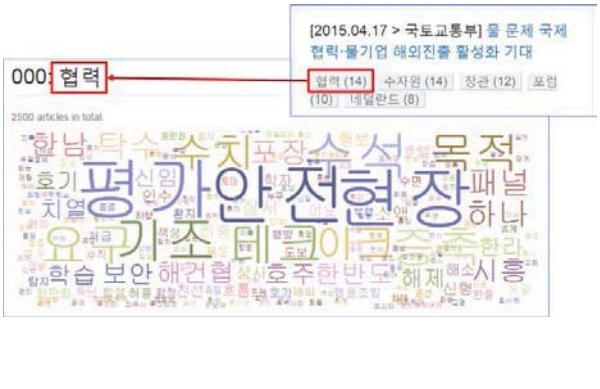
The targeted dataset of the proposed system prototype comprised construction-related documents scattered on the Internet. Specifically, construction-related news, editorials, interviews, reports, and official documents on the Internet were deemed to be appropriate for the dataset because these data types are prolific, frequently generated and accumulated in a variety of themes, and treated as being useless. The date of collected data ranged from the website's inception date to 21/05/2015. The total number of data files for the pilot study was 25,143, which was approximately 279MB as a unit of memory.

B. *Knowledge Discovery*

Collected data were preprocessed by using text mining techniques, including POS tagging, to calculate the weight of each term in a document. A dataset was visualized as a form of wordcloud based on the processed and weight calculated data summarizing the dataset.

Figure I is an example of the '협력(cooperation)' tag from a news article titled '물 문제 국제협력·물기업 해외진출 활성화 기대 (International cooperation on water issues · Expectation on activating water companies' international project awards))'. The keywords of '협력(cooperation)' seem to be '평가(evaluation)', '안전(safety)', and '현장(construction site)'. Thus, knowledge related to cooperation can be discovered such as "the Korean construction industry related to '협력(cooperation)' is mainly concerned with safety evaluation and cooperation on site".

FIGURE I
PROTOTYPE SERVICE OF A DATASET'S VISUALIZATION



C. *Evaluation*

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 Likert 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.

IV. CONCLUSIONS

The designed knowledge creating system from construction documents will increase reuse and sharing of data and information, improve reading efficiency, and ultimately, improve business efficiency. Automated visualization was 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.

ACKNOWLEDGEMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (grant #2015017605)

REFERENCES

[1] C.P. Pathirage, D.G. Amaratunga, and R.P. Haigh, "Tacit Knowledge and organisational performance: construction industry perspective", *Journal of Knowledge Management*, vol. 11, no. 1, pp. 115-126, 2007.

[2] G.M. Song, D.J. Kim, and Y.S. Yu, "Studies on utilization of lessons-learned system on intranet system in architectural design office" *Info Design Isuue*, vol. 16, pp. 59-73, 2009.

[3] J.W. Choi and L.Y. Kim, "A Study on Inforgraphic for Effective Visual Communication of the Big Data Era -Government Departments and Public Institutions", *Korea Science & Art Forum*, vol. 11, pp. 165-175, 2012.

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

[5] L. Soibelman, J. Wu, C. Caldas, I. Brilakis, and K.Y. Lin, "Management and analysis of unstructured construction data types", *Advanced Engineering Informatics*, vol. 22, no. 1, pp. 15-27, 2008.

[6] M. Hjelt and B.C. Björk, "Experiences of EDM usage in construction projects", *Journal of Information Technology in Construction*, vol. 11, pp. 113–125, 2006.

[7] M.A. Qady and A. Kandil, "Concept relation extraction from construction documents using natural language processing", *Journal of Construction Engineering and Management*, vol. 136, no.3, 294-302, 2010.

[8] Z. Ma, N. Lu, and S. Wu, "Identification and representation of information resources for construction firms", *Advanced Engineering Informatics*, vol. 25, no. 4, pp. 612-624, 2011.