Empirical Validation of an Automated Method of Knowledge Structure Creation from Single Documents

Hyung W. Kim
Knowledge Service Engineering Department
Korea Advanced Institution for Science and Technology
Daejeon, Republic of Korea
hw_kim@kaist.ac.kr

Mun Y. Yi
Knowledge Service Engineering Department
Korea Advanced Institution for Science and Technology
Daejeon, Republic of Korea
munyi@kaist.ac.kr

***Abstract*—The present study proposes a new method that automatically creates a knowledge structure from a single document. Knowledge structure refers to an organization of knowledge, represented through key concepts that make up the knowledge domain and their proximity relationships. We examine several alternative methods in inferring a knowledge structure from a document, and those methods are compared with the knowledge structures obtained from learners (before and after learning) and domain experts. The knowledge structure created by the method that utilizes paragraph co-occurrences with cosine similarity has been found to be the most successful in generating a knowledge structure from a document and its performance to be comparable to a human expert in terms of similarity and consistency. Findings of the study have the potential to significantly improve the current practice in information retrieval.**

***Keywords*—knowledge structure, knowledge organization, knowledge elicitation, concept extraction, co-occurrence analysis**

I. INTRODUCTION

We live in a knowledge driven society, in which knowledge constitutes the basis for individual competiveness and organizational performance. While it is still controversial what exactly the term knowledge means, many people agree that knowledge organization is an important aspect of knowledge. Being knowledgeable means not only knowing the concepts, ideas, terms, and rules that make up the knowledge domain, but also understanding their relationships correctly. Goldsmith and Kraiger [5] assert that "facts and rules are only meaningful or accessible in a domain because of their underlying knowledge structure".

Through the process of externalization, knowledge is compiled into tangible products such as documents, articles, books, and Web pages. In addition to knowledge, knowledge structure plays a key role in this externalization process as the key elements of knowledge in a person's mind (i.e., concepts) need to be searched, compared to other elements, and presented in a well-organized fashion. Although it has been widely recognized that knowledge structure is an important aspect of knowledge and knowledge structure is a key mechanism in relating knowledge elements, measuring knowledge structure has been a challenge. Prior research has proposed a variety of methods to obtain interrelationships between the key concepts that make up the knowledge domain and build knowledge structures, including word associations, ordered recall, card sorting, and pairwise rating [5]. These methods all rely on direct human inputs, requiring substantial time and effort for knowledge elicitation.

This study proposes a new method that automatically builds a knowledge structure from a single document. The method transforms the content of a document into a form of knowledge structure through the processes of extracting key concepts and extracting relational information between the key concepts. Furthermore, this study performs an empirical validation of the proposed method by comparing the automated knowledge structure with the knowledge structure of the actual document learners and domain experts obtained through a traditional knowledge elicitation approach. Multiple techniques are compared for effective building of the automated knowledge structure.

II. RELATED STUDY

A. Knowledge Structure

While it is possible to use knowledge structure as a more general term referring to an inner representation of the world, in a more specific, narrow sense, it refers to a model that organizes key concepts that make up the knowledge domain and interrelationships between those concepts [1,5]. Knowledge structure is different from declarative knowledge or procedural knowledge [2] and it plays a principal role in converting knowledge into tangible products through the externalization process [5]. The knowledge structure of the learner becomes more elaborate as the learning gets advanced. In this view, knowledge acquisition means not only simply learning about concepts, words, and rules, but establishing relationships between these elements. Learning can be seen as a series of procedures where it acquires and changes the knowledge structure of a specific domain.

A study by Goldsmith, Johnson, and Acton [3] presents a method to extract individual's knowledge structures directly from people. The Goldsmith et al.'s study estimates and examines individual's knowledge structures using the three steps of knowledge elicitation, knowledge representations, and

978-1-4577-2162-5/11/$26.00 ©2011 IEEE

evaluation. In the knowledge elicitation step, an individual is asked about relatedness between every pair of the concepts selected to represent the domain. In the knowledge representation step, the numerical scores given by an individual are expressed as a network graph of concepts using the pathfinder algorithm [4]. In this step, the original data captured in the form of a proximity matrix goes through the pathfinder scaling procedure so that the noise associated with the raw data can be effectively removed [3]. In the final step of evaluation, the derived knowledge structures are compared to some standard, such as expert's knowledge structure, using the measure of closeness [6]. Closeness is a measure of similarity between two network structures. It ranges from zero to one, where one indicates that the two networks have identical configurations. The learners with closer knowledge structures to that of domain experts are considered to have achieved better results in the learning process [3]. In addition, a measure of coherence, which indicates the internal consistency of a set of similarity relationships, can be used to assess how well an individual's knowledge is organized. The coherence measure is based on the idea that if concepts A and B are considered as similar, and B and C are perceived as similar, then A and C should be also perceived similar.

B. Key Concept Extraction

The need for automatic keyword extraction research has emerged from the information explosion in the Internet environment, and is being actively researched on. Automatically extracted keywords can be used in various fields such as document searching, classification, clustering, browsing, translation, creating automatic thesaurus and so on.

For automatic key word extraction, a statistical method and a machine learning method are generally used. Since the study that assigns additional weights to words that frequently appear was proposed [7,8], other statistical methods of key word extraction have been proposed as well. Choosing the candidates of key words using TF (Term Frequency) and extracting the key words from the candidates based on co-occurrences information between the words in a document is typical in statistical methods [9]. In machine learning methods, proposing GenEx algorithm based on generic algorithms [10], using TF*IDF with Naïve Bayes algorithm based on position information of the words in documents or sentences [11], and using neural network based on TF*IDF, ITF (Inverted Term Frequency), T (Title), FS (First Sentences), LS (Last Sentence) information for extracting key words [12] have been studied.

C. Relation Extraction between Words

For automated knowledge structure, which is a core element of this study, automatic extraction of relationships between words is essential. Relation extraction between words is generally performed for two alternative types of relations: semantic relation and association relation. A semantic relation represents semantic relationships between concepts using a verb phrase as a linking mechanism between two noun phrases. An association relation represents definitional and psychological relatedness between two words that are conceptually related yet not equivalent such as a synonym or near-synonym. Knowledge structures are created from the proximity data, instead of semantic data [5].

A traditional association relation extraction method includes a method using co-occurrence of words. Co-occurrence represents simultaneous occurrence of two words in the same document, sentence, phrase and so on. It is based on a theory that a higher frequency of co-occurrence implies closer relationship between two words [13]. A method of estimating co-occurrence was proposed by Salton [7], from whom keyword extraction is originated. Although the techniques for searching closeness between words using extra external thesaurus or corpus such as WordNet and Wikipedia [14] have been made afterward, it has difficulty in defining unregistered words. It was also found ineffective to extract co-relations from the documents where proper nouns, newly coined terms, and technical terms are largely used.

III. Process of Automated Knowledge Structure from Single Document

To extract a knowledge structure from single document, three steps are needed: (1) key concept extraction from the document, (2) co-relation extraction between key concepts, and (3) knowledge structure creation using the extracted concepts and their relationships.

A. Key Concept Extraction

Morphological analysis of the document is required first to extract key concepts from a single document. This study extracts only nouns from the words in a document, using KoreanAnalyzer, which was developed by a Lucene Korean Analysis Open Source Project [15]. Not all extracted nouns that are frequently used in the document are key concepts of that document. Usually we lower the weighted value of those words using TF*IDF, yet IDF information cannot be used if it is a single document. Thus, we used Korean Wikipedia [16] to identify non-key concept words and mark them as stopwords. Korean Wikipedia is a web based encyclopedia that is open to the public so that all the visitors are able to edit any information over 28 million concepts of various domains in it. Among the words that match outside corpus like the Korean Wikipedia, we extracted the most frequently used words, number of N, as key concepts of that document. In addition, for proper recognition of compound words, terminology matching of successive words (ex: Trojan Horse) was also performed.

B. Co-relation Extraction between Concepts

Co-occurrence of two words is used to extract co-relation between the key concepts of the document. This study classifies co-occurrence into sentence co-occurrence, which represents the frequency of two co-occurrent concepts in the same sentence and paragraph co-occurrence, which represents the frequency of two co-occurrent concepts in the same paragraph.

The following is a set of equations used to compute the proximity relationships between the key words. In (1) and (2), N_s and N_p are the number of sentence and the number of paragraph in the document, respectively. SC_{ij} represents the sum of W_i and W_j co-occurrences in sentences and PC_{ij} represents the sum of W_i and W_j co-occurrences in paragraphs. Equation (3) is used to compute the proximity

(similarity) between words using sentence co-occurrence (Sentence co-occurrences Similarity: SS) and (4) is the proximity (similarity) between words using paragraph co-occurrence (Paragraph co-occurrences Similarity: PS).

$$SC_{ij} = \{\sum_{L}^{Sn} a(W_i \cap W_j)\} \quad (1)$$

$$PC_{ij} = \{\sum_{L}^{Pn} a(W_i \cap W_j)\} \quad (2)$$

$$SS_{ij} = SC_{ij} / Max(SC),\ (0 \le SS \le 1) \quad (3)$$

$$PS_{ij} = PC_{ij} / Max(PC),\ (0 \le PS \le 1) \quad (4)$$

Although the equations above can easily estimate similarity between words by using co-occurrence information, it has a problem that frequent use of words leads to higher similarity between other words. To solve this problem, this study uses a modified cosine similarity estimating method which is widely used for document clustering.

TABLE I. EXAMPLE OF INVERTED SENTENCE VECTOR

	1st Sentence	2nd Sentence	3rd Sentence	…	Nth Sentence
W_i	3	0	1	…	1
W_j	2	1	0	…	2

Table I explains how ISV (Inverted Sentences Vector), which consists of frequency of concepts appeared on each sentence, is created.

$$SCS_{ij} = \frac{\sum_{k=1}^{Sn} W_{ik} W_{jk}}{\sqrt{\sum_{k=1}^{Sn} W_{ik}^2 \sum_{k=1}^{Sn} W_{jk}^2}},\quad (0 \le SCS \le 1) \quad (5)$$

$$PCS_{ij} = \frac{\sum_{k=1}^{Pn} W_{ik} W_{jk}}{\sqrt{\sum_{k=1}^{Pn} W_{ik}^2 \sum_{k=1}^{Pn} W_{jk}^2}},\quad (0 \le PCS \le 1) \quad (6)$$

Then using (5), we can estimate Sentence co-occurrences Cosine Similarity (SCS) between each concept from a single document. By modifying a sentence number in Table I to paragraph number (Inverted Paragraph Vector: IPV), we can also estimate Paragraph co-occurrences Cosine Similarity (PSC) between each concept with the same method. These methods might be better than SS or PS for estimating relatedness between concepts in a single document because similarity is estimated based on the word's co-occurrence rate regardless of its frequency.

C. *Knowledge Structure Formation Process from Co-relation Information*

To place the computed similarity scores into the same scale range used for knowledge structure assessment method involving human judges [1,2,3,5], co-relation between each concept is modified into a scale of 7 using (7) (1: Strongly related, 7: Not related at all).

$$D_{ij} = 7 - S_{ij} \times 6,\quad (1 \le D_{ij} \le 7) \quad (7)$$

These distance scores were organized in a table, representing similarity ratings between each concept. Next, we applied the pathfinder algorithm [4] to the table and derived a knowledge structure, based on the shortest paths among the concepts [3,6].

IV. EXPERIMENT AND RESULT

We compared the validity of the proposed methods using a document selected from Navercast [17]. Navercast contains 5,707 specialized articles written from journalists or professors of science, liberal arts, and so on. A document from the field of biology has been used as a sample document for this study. We recruited 106 people from a research university in South Korea as experiment subjects. The age of the subjects ranged from 19 to 32.

As the subjects arrived at the experiment site, they were asked to fill out an initial survey in which demographics and the pre-existing knowledge structure with regard to the selected document were measured. Following prior research on knowledge structure assessment involving human subjects [1,2,3,5], the knowledge structure assessment was made by asking about the relatedness between every pair of eleven automatically extracted key concepts of document on a scale of 1 to 7 (1: Not related at all, 7: Strongly related). Then, the experimenters were given time to learn the assigned document, which was the same across the subjects. To improve the motivation of the study, the subjects were told that there would be a test after the learning period. No time limit was enforced and most subjects spent about 20 to 25 minutes for the learning of the assigned document. After the learning period, the subjects were asked again to assess the relatedness between every pair of the key concepts.

Based on the survey responses, for every subject, we created two instances of knowledge structures – knowledge structure before learning and knowledge structure after learning (see Fig. 1 and 2 for exemplary knowledge structures before and after learning). In addition, we recruited three graduate students (2 doctoral students and 1 master student) who majored in life sciences at the same university and created knowledge structures of experts on the assigned document based on their responses. These three experts shared their opinions while responding to the knowledge structure questionnaire and the average scores of closeness between words, calculated from three experts, were used to derive the expert knowledge structure (see Fig. 3). The measure of closeness between knowledge structures are shown in Table II.

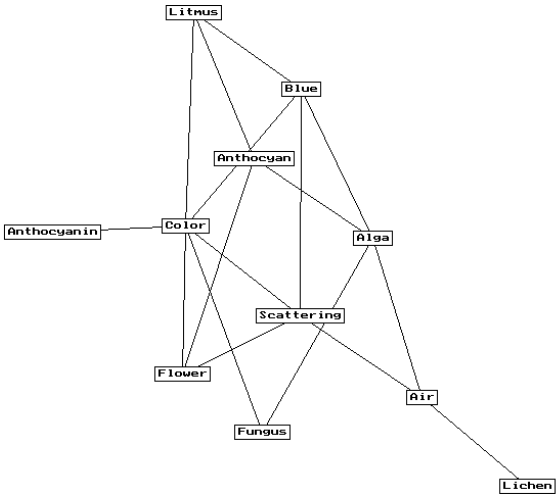

Figure 1. Knowledge structure of subject #68 before learning

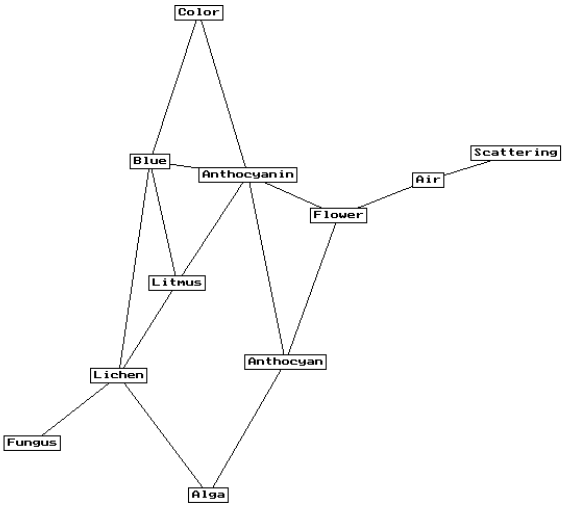

Figure 2. Knowledge structure of subject #68 after learning

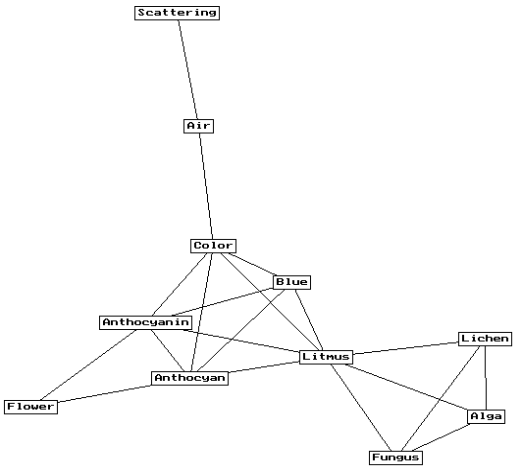

Figure 3. Knowledge structure based on expert ratings

TABLE II. CLOSENESS MEASURE BETWEEN KNOWELDGE STRUCTURES

	After learning	Experts	PCS	SCS	PS	SS
Before learning	**0.32**	**0.30**	0.30	0.25	0.32	0.27
After learning	-	**0.58**	**0.52**	0.42	0.51	0.44
Experts	-	-	**0.67**	0.48	0.57	0.48

In Table II, the closeness of knowledge structure before and after the learning period is low at 0.32, implying that the knowledge structure of subjects has changed through the learning period. Furthermore, the results show that the knowledge structure developed after the learning period is closer to that of the experts than before learning occurred (Closeness before learning: 0.30, Closeness after learning: 0.58). This shows learning did actually occur and it was in the right direction.

Among the four methods that measure similarities differently to produce a knowledge structure from a document, the study results show that the PCS method, which estimates relatedness between documents with cosine similarity using paragraph co-occurrence, succeeded in creating the most similar knowledge structure to that of the subjects' obtained after learning (Closeness: 0.52) and to that of the experts (0.67). The value of 0.52 is similar to the closeness value obtained between the knowledge structure of the experts and that of the learners (closeness: 0.58), indicating that the PCS method is comparable to the experts' ratings. In addition, the PCS method shows the closest result to experts' knowledge structure among the four methods (closeness: 0.67). Fig. 4 presents the knowledge structure obtained using the PCS method. Fig. 5 shows the summarization of the closeness analysis.

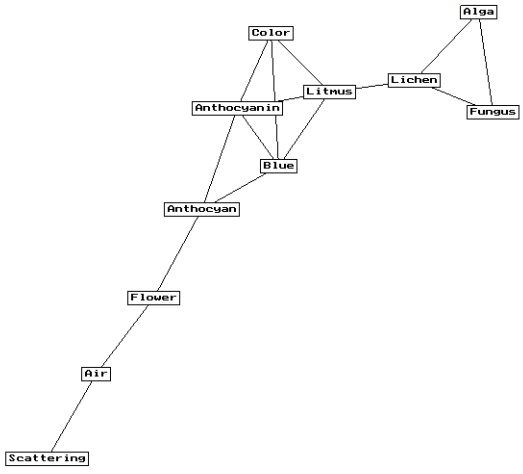

Figure 4. Knowledge structure based on the PCS method

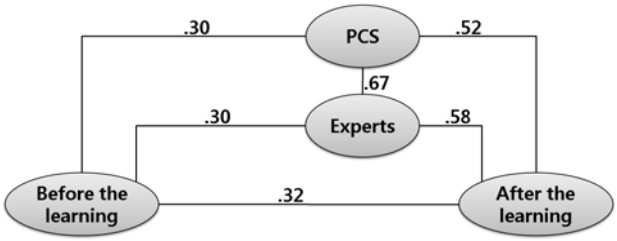

Figure 5. Summary of closeness between knowledge structures

Coherences scores are shown in Table III. In Table III, while knowledge structure before learning lacks coherence, knowledge structure after learning shows that coherence has been increased significantly (before the learning: 0.25, after the learning: 0.64). This again shows learning occurred successfully. Automated knowledge structure is as consistent as, or better than, that of experts. Particularly, knowledge structure created through PCS is more coherent than experts' knowledge structure. Additionally, the closeness of knowledge structures between three experts ranged from 0.65 to 0.73, and the closeness between PCS-automated knowledge structure and three experts' knowledge structures ranged from 0.50 to 0.68. The overall results show that PCS is the most effective in generating knowledge structures from documents and the knowledge structure based on PCS is comparable to a knowledge structure based on a human expert in terms of similarity and consistency.

TABLE III. COHERENCE MEASUER OF KNOWELDGE STRUCTURES

Before learning	After learning	PCS	SCS	PS	SS	Experts
0.25	**0.64**	0.89	0.73	0.72	0.66	0.73

V. CONCLUSION AND FURTHER RESEARCH

This study proposed an automated method that creates cognitive knowledge structure from a single document. The automated knowledge structure created through co-location information between concepts in the paragraphs shows significant closeness with the knowledge structure of subjects who learned the knowledge in a document and that of domain experts. In terms of similarity and consistency, the automated knowledge structure was comparable to that of human experts. Our study demonstrates the effectiveness of an automated method of knowledge structure creation out of a single document. Automatically created knowledge structures can be applied to various areas such as document recommendation, learning guide, information search and retrieval, and so on. The proposed method of automated knowledge structure creation opens new opportunities for significant enhancement of current practice in information retrieval and a shift of a paradigm - from techniques based on words to techniques based on knowledge structures, which include words and their relationships. Further research needs to be done to develop and expand the current automated method of creating knowledge structure based on a single document to an automated method of domain knowledge structure using multiple documents. Follow-up research that can further optimize the knowledge structure creation also needs to be done continuously. For those studies, the current study serves as an important starting point.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0024560).

REFERENCES

[1] E. A. Day, W. J. Arthur, and D. Gettman,"Knowledge Structures and the Acquisition of a Complex Skill", Journal of Applied Psychology, vol.86, no.5, 2001.

[2] F. D. Davis, and M. Y. Yi, "Improving Computer Skill Training: Behavior Modeling, Symbolic Mental Rehearsal, and the Role of Knowledge Structures", Journal of Applied Psychology, vol.89, no.3, 2004.

[3] T. E. Goldsmith, P. J. Johnson, and W. H. Acton, "Assessing Structural Knowledge", Journal of Educational Psychology, vol.83, no.1, pp.88-96, 1991.

[4] R. W. Schvaneveldt, F.T. Durso, and D. W. Dearholt, "Network Structures in Proximity Data", The Psychology of Learning and Motivation , vol. 24, pp.249-294, 1989.

[5] T. Goldsmith and K. Kraiger, "Structural Knowledge Assessment and Training Evaluation", In J. K. Ford, S. W. J.Kozlowski, K. Kraiger, E. Salas, & M. S. Teachout (Eds.), Improving Training Effectiveness in Work Organizations, pp. 73–96, 1997.

[6] T. Goldsmith and D. M. Davenport, "Assessing Structural Similarity of Graphs. In R. W. Schavneveldt (Ed.), Pathfinder Associative Networks: Studies in Knowledge Organization (pp.75-87), 1990.

[7] G. Salton, "Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer",. Addison-Wesley, 1989.

[8] Salton, G., "Developments in Automatic Text Retrieval.", Science 253, pp.974 –980. 1991.

[9] Y. Matsuo and M. Ishizuka, "Keyword Extraction from a Single Document Using Word Co-occurrence Statistical Information", International Journal on Artificial Intelligence Tools, vol. 13, no. 1, pp. 157-169, 2004

[10] P. Turney, "Learning Algorithms for Keyphrase Extraction", Information Retrieval, vol. 2, no. 4, pp. 303–336, 2000

[11] Y. Uzun, "Keyword Extraction Using Naïve Bayes", Bilken University, Department of Computer Science, 2005

[12] T. Jo, M. Lee, and T. M. Gatton, "Keyword Extraction from Documents Using a Neural Network Model", ICHIT'06, vol. 2, pp. 194-197, 2006

[13] C. J. Van Rijsbergen, "A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval", Journal of Documentation, vol.33, no.2, pp.106-119, 1997

[14] F.M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A Core of Semantic Knowledge—Unifying WordNet and Wikipedia", in Proc. of the 16th International Conference on World Wide Web, WWW2006, Banff, Canada, 2007

[15] http://cafe.naver.com/korlucene/

[16] http://ko.wikipedia.org/

[17] http://navercast.naver.com/