A KNOWLEDGE DOCUMENT SEMANTIC ANALYSIS TECHNOLOGY

Shih-Ting Yang[*] and Yu-Ting Gong
Department of Information Management
Nanhua University
Chia-Yi (622), Taiwan

ABSTRACT

Although websites at present allow users to obtain information by topic or keywords, the search may not be successfully if users lack domain knowledge in the specified field. For example, in the domain website of "Institute of Occupational Safety and Health" (IOSH, Http://www.iosh.gov.tw), the most demanders cannot obtain the needed knowledge documents using the colloquial query string without domain keywords respect to knowledge documents and thus reducing the knowledge sharing efficient of this domain website. To solve above problem, firstly, this paper analyzes the ergonomic technology reports from the website to capture the expressions and related vocabulary of domain knowledge documents to develop the knowledge vocabulary database. Secondly, through the knowledge document semantic analysis technology, the correlations between domain vocabulary and colloquial query string can be obtained. It is expected that knowledge demanders can directly read the desired parts according to problems to ensure they can find document they want within a short time. In order to demonstrate applicability of the proposed methodology, a web-based knowledge document retrieval system is also established based on the proposed model. Furthermore, the knowledge documents (i.e., ergonomic technology reports) from the website of "Institute of Occupational Safety and Health" are applied as examples to demonstrate the proposed model and system.

Keyword: Institute of Occupational Safety and Health, Knowledge Management, Data Mining, Semantic Analysis

1. INTRODUCTION

As the Internet becomes a popular source of information acquisition, many papers have been conducted and technologies have emerged to help users searching for and accessing information quickly and efficiently. To help users obtaining information conveniently, websites have been established to collect literature of related domain fields; these websites are representative in the specific domain fields. In other words, when users are searching for information, they conduct searches on the website intuitively. Although websites at present allow users to obtain information by topic or keywords, the search may not be successfully if the users lack domain knowledge in the specified field. For example, in the domain website of "Institute of Occupational Safety and Health" (IOSH, Http://www.iosh.gov.tw), the most demander cannot obtain the needed knowledge documents using the colloquial search phrases without domain keywords respect to knowledge documents and thus reducing the knowledge sharing efficient of domain website.

To solve above problem, this paper proposes a knowledge document semantic analysis technology between domain vocabulary and colloquial query string to help users rapidly and efficiently obtain the documents needed. Based on the domain website of IOSH, this paper analyzes the knowledge document expressions and related domain vocabulary regarding the research reports or technical books on the knowledge websites, and establishes the knowledge vocabulary library. In this way, semantic association with the search phrase can thus be established to enhance the keyword semantic search technology. The proposed technology can strengthen the search phrase in semantic determination by using the representative vocabulary of the document and thus enhancing the knowledge sharing effectiveness of the website of IOSH. To sum up, this paper develops a semantic analysis technology between domain vocabulary and colloquial search phrase for domain knowledge document.

[*] **Corresponding author: stingyang@nhu.edu.tw**

2. LITERATURE REVIEW

Regarding the topic of Q&A (Question and Answer) application and technology, in this study, this study conducted literature review relating to Q&A application types and Q&A technology.

The types of Q&A application can be divided into the Q&A system and retrieve system. Regarding Q&A system, Oh et al. [9] proposed a compositional Q&A system, using criteria judgment for question analysis. The question format (single or multiple question items), subject, question limitations (time or location) are used as the judgment criteria to learn about the types and formats of feedback sentences. Cao et al. [1] established an online Q&A system (AskHERMES) for medical clinical reports to capture the key points of the complex clinical reports without fixed format. Regarding information retrieval, Huang et al. [5] proposed a composite relational model to capture biomedical literature by using the shallow parsing to develop the grammatical and semantic structure, and using the greedy method for matching to acquire the theme of the biomedical literature through the training mode. According to the document association and common features' BE (Basic Element), Teng et al. [10] established a user-oriented document abstract retrieval system.

In terms of Q&A technology, this study categorized various considerations. The factors for consideration may be based on subject, the document characteristics or the semantics for analysis. In the case of using subject as the main reference, Oh et al. [9] proposed a Q&A system learning mechanism to analyze the structure through the existing Q&A documents, and used the word meaning disambiguation for semantic analysis to obtain the combinations of questions and answers (answer format, answer subject, target and expected answer content). Han et al. [5] determined question types to establish various types of relevant vocabulary, allowing the users to determine the problem targets and analyze the question retrieval category, in order to expand the question. Jones and Love [6] argued that if the relationships of the documents are more similar, it means that there is a common role in between the two documents. Through the background environment, with the relationship as the matching criteria, the common relationship of the documents can be obtained. Ko et al. [7] used the important sentences as the basis for document classification in order to enhance document classification technology. Using semantics as the main reference, Dorr and Gaasterland [2] proposed a composite model considering tense and semantic relationship to associate relevant events based on time sequence relationship and event viewpoints. Dunlavy et al. [3] proposed an integrated information question system to conduct the relevant question analysis according to main sentences with characteristic market documents, such as the sentence location and document content, by the potential semantic index technology.

3. A KNOWLEDGE DOCUMENT SEMANTIC ANALYSIS TECHNOLOGY

The proposed a Knowledge Document Semantic Analysis Technology between domain vocabulary and colloquial query string used the technical books and research reports on the website of "Institute of Occupational Safety and Health" (IOSH, Http://www.iosh.gov.tw) as the basis for analysis. The corresponding knowledge document can be found to enhance retrieval accuracy. Therefore, the main research procedures can be divided into the following parts including Part1 "Knowledge Document Expression Item Analysis Module", Part2 "Conceptual Sentence Acquisition (CSA) Module", and Part3 "Question and Answer Analysis (QAA) module".

3.1 Knowledge Document Expression Item Analysis Module

This paper consulted ergonomics staffs and summarized the repetitive or important descriptive words in the improvement reports and technical books for the establishment of expression items of the knowledge documents. After the analysis of the knowledge documents, the establishment of expression items and the capturing of the conceptual sentences, the expression items and the details of the detailed expression items are illustrated as below.

3.1.1 Establishment of the expression items of the knowledge documents

The analysis of the content of the ergonomics workplace research reports can be divided into 8 expression items, and 19 detailed expression items. To strengthen the sentence smoothness, nine detailed expression items are added for sentence assistance. Hence, knowledge vocabulary database consists of 28 vocabulary sets can be shown in Table 1 and Table 2.

3.2 Conceptual Sentence Acquisition (CSA) Module

Since the ergonomic technology reports (target document) are written by experts in the domain field, the expression methods are not consistent with each other. The CSA module acquires the complete sentence SD_i by segmenting the target document D_T, and conducts vocabulary comparison rules on the basis of domain vocabulary set created by domain experts. Then, this module compares the complete sentence SD_i and vocabulary comparison rules to extract the conceptual sentences and attribute them to corresponding sets.

Table 1: Detailed expression items and corresponding contents and means of expression

Expression Items	**Detailed Expression Items**	**Descriptions of Expression**
Operation Field	Operation field	Set of industrial category with contents including “agriculture, forestry, fishery and animal husbandry”, “mining and quarrying”, “food manufacturing”, and “textiles and clothing industry”.
Operation Name	Operation name	Set of name of the operation in this industry with contents including “mode replacement operation”, “packaging operation”, and “transportation operation”.
Operation Identity	Operator gender	Set of vocabulary describing the gender of the operations with contents including “male” “female” or “male or female”.
	Operator age	Set of vocabulary describing the operator age, for example “20~30 years old”; with contents including “the middle aged”, “the youth”, and “no age limit”.
	Operator title	Set of vocabulary describing the operator title with contents including “nurse”, “technician”, and “operator”.
Operation Environment	Equipment vocabulary	Set of the equipment vocabulary for the operation such as “blood bed”, “thermotank”, and “thermal forging machine”.
	Facilities layout	Set of the vocabulary of facilities layout and placement such as “the height of the transmission belt is 75 cm”.
	Tool introduction	Set of the vocabulary describing tools used in operation or improvement process with contents including “butterfly cage”, “arm support rest”, “lift vehicle”, and “cart”.
Operation Behavior	Operation goal	Set of the vocabulary describing the name of the operation corresponding to the operation target with contents including “major job”, “main function” and “main points”.
	Operation description	Set of vocabularies corresponding to the operation name, operation goal and operation tool.
	Professional verbs	Set of vocabulary describing the corresponding operations of the operator with contents including “standing posture”, “bending”, “force application” and “lifting”.
Operation Hour	Operation hours/day	To describe the times of repetitive actions of the operator including “daily requirements”, and “daily necessity”.
	Operation hour/times	Set of vocabulary describing the time for the job of the operator with contents including “time for one times”, “one times requirement”, and “one times necessity”.
	Operation distance/times	Set of vocabulary describing the operation distance of the operator with contents including “distance”, and “shortest distance”.
Causes of Injury	Factors of injury	Set of vocabulary describing causes of injury with contents including “excessive force”, “highly repetitive actions”, “vibration”, “low temperature”, and “poor working posture”.
	Pain parts	Set of vocabulary describing the posture and pain parts of the operator with contents including “neck”, “torso”, “hand”, “wrist”, and “leg”.
Improvement Method	Improvement goal	Set of vocabulary describing the improvement goal of the causes of injury with contents including “main improvement”, “effective improvement”, “considerably”, and “significant reduce”.
	Improvement procedure	Set of vocabulary describing procedural improvement with contents including “consider”, “use”, “suggestion”, and “as long as”.
	Improvement review	Set of vocabulary describing the review after operation assessment with contents including “action level”, “grading points”, “total inspection score”, “risk level”, and “significant reduction of load”.

Table 2: Auxiliary detailed expression items and corresponding contents

Detailed Expression Items	Descriptions of Expression
Linking vocabulary	Including “and”, “but”, “as well as”.
General verbs	Set of the vocabulary of general verbs including “is”, “mainly is”, “as”, “raise”, “put down”, “move”, and “store”.
Numerical vocabulary	Vocabulary recording numbers from “0” to “9” and their combinations
Monetary unit vocabulary	Including “RMB”, “NTD”, etc.
Age unit vocabulary	Including “years old”.
Length unit vocabulary	including “distance”, “cm”, “meter”, “length”, etc.
Time unit vocabulary	including “time”, “minute”, “hour”, etc.
Wight unit vocabulary	Including “kg”, “g”, “ton”, etc.
Frequency vocabulary	including “times”, etc.

Step (A1): Target Document Sentence Acquisition

This step first builds the punctuation marks set (for example:. !,;, etc) to obtain the sentences of the target document D_T.

(A1.1): Subsection of Target Document:

According to the table of punctuation symbols (for example:. !,;), sub-sections of the target document are worked out. After this step, the complete sentences of the target document D_T including SD_1, SD_2, SD_3, …, SD_i, $SD_{N(DT)}$ can be obtained.

(A1.2): Word Dismantling of the Complete Sentences:

After getting the complete sentence SD_i, the word series are dismantled into word groups ranging from 2 to 6 words to form the vocabulary set. $SD_{i,j}$ represents the j‘th word of the i‘th sentence after dismantling, consisting of a number of words as shown in Equation (1).

$$SD_i = \left\{SD_{i,1}, SD_{i,2}, SD_{i,3} \cdots, SD_{i,j}, \cdots\right\} \qquad (1)$$

Step (A2): Establishment of Structured Vocabulary Comparison Rules

After the formation of the complete sentences SD_1, SD_2, SD_3, …, SD_i, …, $SD_{N(DT)}$, the conceptual sentences can be judged. This paper establishes eight selection rules regarding the vocabulary comparison rules to obtain the representative sentences of the vocabularies.

1. Operation Field Vocabulary Rule (R_OF): This rule is to express the industrial classification, and the rule is as shown in Equation (2). If the complete sentence S is in the operation field conceptual vocabulary, then the complete sentence SD_i is the operation field conceptual sentence OF_Set.

$$\text{IF } SD_{i,j} \text{ exist in } OF(CS)\ \forall j \text{ Then } SD_i \in OF\_Set \qquad (2)$$

2. Operation Name Vocabulary Rule (R_ON): This rule is to express the name of the action, and hence the rule is as shown in Equation (3). If the complete sentence is in the operation name conceptual vocabulary, then the complete sentence SD_i is the operation name conceptual sentence ON_Set.

$$\text{IF } SD_{i,j} \text{ exist in } ON(CS)\ \forall j \text{ Then } SD_i \in ON\_Set \qquad (3)$$

3. Operator Identity Rule (R_OR): This rule’s expressions include the operator’s gender vocabulary, age vocabulary, and title vocabulary. To strengthen the accuracy of the judgment rules in this paper, at this step, a strict rule to ensure accurate acquisition is built. The method is to acquire the set of words of the complete sentence by Equation (1) and select the vocabulary by rules. The loose and strict rules are defined as follows:

✓ The loose rule: This rule is to express the concept of operation by one to two words, for example, as shown in Equation (4), using operator title ORT (CS) to represent the operator identity, or using the operator title ORT(CS) coupled with operator age ORA(CS) to represent the age of the operator, using the combination of the operator age, operator title ORT(CS) and operator gender ORS(CS) to express the gender of the operator (Equations (5) and (6)).

$$\begin{array}{l}\text{IF } SD_{i,j} \text{ exist in } ORT(CS)\ \forall j \\ \text{Then } SD_i \in OR\_Set\end{array} \qquad (4)$$

$$\begin{array}{l}\text{IF } SD_{i,j} \text{ exist in } \left(\begin{array}{l}ORT(CS) \\ \text{ and } ORA(CS)\end{array}\right) \forall j \\ \text{Then } SD_i \in OR\_Set\end{array} \qquad (5)$$

$$\begin{array}{l}\text{IF } SD_{i,j} \text{ exist in } \left(\begin{array}{l}ORT(CS) \\ \text{and } ORS(CS)\end{array}\right) \forall j \\ \text{Then } SD_i \in OR\_Set\end{array} \qquad (6)$$

✓ The strict rule: This rule uses a couple of words to form the strict structure for the expression of the concept relating to the operator identity, uses the numerical vocabulary N(CS) and age unit vocabulary AU(CS) to expressly represent the operator’s age range (Equation (7)).

$$\begin{array}{l}\text{IF } SD_{i,j} \text{ exist in } \left(\begin{array}{l}ORT(CS) \text{ and } N(CS) \\ \text{ and } AU(CS)\end{array}\right) \forall j \\ \text{Then } SD_i \in OR\_Set\end{array} \qquad (7)$$

4. Operation environment vocabulary rule (R_OE): This rule is to express the facilities and tools of the operation environment with descriptions including the descriptions of length, width, height and other specifications. As shown in Equation (8), the description of the operation environment is realized by facility vocabulary F(CS), facility

layout vocabulary FL(CS), numerical vocabulary N(CS) and length unit vocabulary LU(CS) for definite expression of the operation facility's specifications. The description of the operation tools is as shown in Equation (9), the operation tool vocabulary OT(CS) is combined with the numerical vocabulary N(CS) and length unit vocabulary LU(CS) to definitely express the specifications of the operation tools.

$$\text{IF SD}_{i,j}\text{ exist in}\begin{pmatrix}\text{F(CS) and FL(CS)}\\ \text{and N(CS) and LU(CS)}\end{pmatrix}\forall j \qquad (8)$$

$$\text{Then SD}_i \in \text{OE\_Set}$$

$$\text{IF SD}_{i,j}\text{ exist in}\begin{pmatrix}\text{OT(CS) and N(CS)}\\ \text{and LU(CS)}\end{pmatrix}\forall j \qquad (9)$$

$$\text{Then SD}_i \in \text{OE\_Set}$$

5. Operation behavior vocabulary rule (R_OV): This rule is to express the description of the operation goals. The expressions include the including operation goal vocabulary, operation tool vocabulary and the domain verbs to express the operations and postures. According to Equation (10), the description of operation goal should be integrated with the operation goal OG (CS) and the general verb vocabulary GV (CS); or as shown in Equation (11), the operation goal vocabulary OG(CS) can be integrated with the general verb vocabulary GV(CS) and domain verb vocabulary PV(CS) to more strictly express the concepts. The expression for the operation definition vocabulary is as shown in Equation (12), the operation definition rule is to combine the operation name vocabulary ON(CS) with the general verb vocabulary GV(CS), domain verb vocabulary PV(CS) and operation tool vocabulary OT(CS).

$$\text{IF SD}_{i,j}\text{ exist in}\begin{pmatrix}\text{OG(CS)}\\ \text{and GV(CS)}\end{pmatrix}\forall j \qquad (10)$$

$$\text{Then SD}_i \in \text{OV\_Set}$$

$$\text{IF SD}_{i,j}\text{ exist in}\begin{pmatrix}\text{OG(CS) and GV(CS)}\\ \text{and PV(CS)}\end{pmatrix}\forall j \qquad (11)$$

$$\text{Then SD}_i \in \text{OV\_Set}$$

$$\text{IF SD}_{i,j}\text{ exist in}\begin{pmatrix}\text{ON(CS) and GV(CS)}\\ \text{and PV(CS) and OT(CS)}\end{pmatrix}\forall j \qquad (12)$$

$$\text{Then SD}_i \in \text{OV\_Set}$$

6. Operation hour vocabulary rule (R_OH): This rule is to represent the operation frequency and operation time. The descriptions include operation frequency (operation times/day) vocabulary OFQ (CS), operation hour (operation hour/times) vocabulary OH(CS), operation distance (operation distance/times) vocabulary ODT(CS). By the selection of Equations (13), (14), and (15), the sentences are listed in line with the standards as the set of the operation time vocabulary conceptual sentences OH_Set.

$$\text{IF SD}_{i,j}\text{ exist in}\begin{pmatrix}\text{OFQ(CS) and PV(CS)}\\ \text{and N(CS)}\end{pmatrix}\forall j \qquad (13)$$

$$\text{Then SD}_i \in \text{OH\_Set}$$

$$\text{IF SD}_{i,j}\text{ exist in}\begin{pmatrix}\text{OH(CS) and N(CS)}\\ \text{and FU(CS)}\end{pmatrix}\forall j \qquad (14)$$

$$\text{Then SD}_i \in \text{OH\_Set}$$

$$\text{IF SD}_{i,j}\text{ exist in}\begin{pmatrix}\text{OT(CS) and ODT(CS)}\\ \text{and N(CS) and LU(CS)}\end{pmatrix}\forall j \qquad (15)$$

$$\text{Then SD}_i \in \text{OH\_Set}$$

7. Injury cause vocabulary rule (R_IC): This rule is to express the injuries caused by the operations. The expressions include injury cause vocabulary and body part vocabulary. As shown in Equation (16), expressions of injury cause can be realized by integrating the injury cause vocabulary IC(CS) with the operation body part vocabulary B(CS).

$$\text{IF SD}_{i,j}\text{ exist in}\left(\text{IC(CS) and B(CS)}\right)\forall j \qquad (16)$$

$$\text{Then SD}_i \in \text{IC\_Set}$$

8. Improvement method vocabulary rule (R_IM): This rule includes improvement purpose, improvement process, and improvement review. As shown in Equation (17), the expression and description of the improvement purpose should be combined the improvement purpose vocabulary IG(CS) and the general verb vocabulary GV(CS) and domain verb vocabulary PV(CS). The expression forms of the improvement process vocabulary are as shown in Equation (18). The description of the improvement process is expressed by the combination of the improvement process vocabulary IR(CS), the general verb vocabulary GV(CS) and operation tool vocabulary OT(CS). Regarding the expression of the review vocabulary is as shown in Equation (19) by improvement review vocabulary R(CS) directly or as shown in Equation (20) by the combination of the review verb vocabulary RV(CS), operation title vocabulary ORT(CS) and domain verb vocabulary PV(CS) in a strict way.

$$\text{IF SD}_{i,j}\text{ exist in}\begin{pmatrix}\text{IG(CS) and GV(CS)}\\ \text{and PV(CS)}\end{pmatrix}\forall j \qquad (17)$$

$$\text{Then SD}_i \in \text{IM\_Set}$$

$$\text{IF SD}_{i,j}\text{ exist in}\begin{pmatrix}\text{IR(CS) and GV(CS)}\\ \text{and OT(CS)}\end{pmatrix}\forall j \qquad (18)$$

$$\text{Then SD}_i \in \text{IM\_Set}$$

$$\text{IF SD}_{i,j}\text{ exist in R(CS)}\ \forall j \qquad (19)$$

$$\text{Then SD}_i \in \text{IM\_Set}$$

$$\text{IF SD}_{i,j}\text{ exist in}\begin{pmatrix}\text{RV(CS) and ORT(CS)}\\ \text{and PV(CS)}\end{pmatrix}\forall j \qquad (20)$$

$$\text{Then SD}_i \in \text{IM\_Set}$$

Finally, this module can obtain the sets of eight conceptual sentences including operation field, operation name, operation title, operation environment, operation, operation time, injury cause and improvement method. At the stage of the conceptual sentence acquisition module, the free-form documents are converted into structured expressions containing conceptual sentences for the question and answer analysis.

3.2 Question and Answer Analysis (QAA) Module

As most of the query strings input by the users are intuitive or colloquial query string words of the users and do not belong to the domain vocabulary of the knowledge document. The domain vocabulary search is used to find out the relevant knowledge document; on the contrary, the self-defined query strings (i.e., colloquial query string) may have no clear definitions, it may result in finding some irrelevant documents. Hence, to enhance the natural language search flexibility, this paper proposes a knowledge document Q&A Analysis (QAA) module to conduct the analysis of the main question words of the colloquial query strings of the users, and find out the semantic words of association by matching and parsing of question words and answer words, and thus capturing the corresponding knowledge documents and enhancing the retrieval accuracy.

<u>Step (B1): Determination of the implied goals of the question words</u>

In this step, Q&A can be used to obtain the implied goals and relevant words of the question words to for the training of the similarity probability of question words and relevant semantic words. As shown in Equation (21), when the users inputs the question sentence (QW_i), after the segmentation of the sentence, the meaningful question word ($QW_{i,j}$) can be captured. In accordance with the bilingual vocabulary developed by the Institute of Occupational Health and Safety, the relevant semantic words can be obtained. According to the domain term “bruise”, the relevant semantic words “skin” or “membrane” can be captured from the notes “the superficial tissues (e.g., skin or membrane) covering the body has been scratched or torn off”. By using the semantic words as the question words, the answer sentence (AW_k) composing of multiple answer words ($AW_{k,m}$) can be obtained. Equation (22) computes the probability of the words in the question sentence ($QW_{i,j}$) and answer words ($AW_{k,m}$) to obtain relevant vocabulary to establish the set of question words and the corresponding sets of answer words to various question words. Finally, as shown in Equation (23), by using the semantic threshold value of the question word against answer words ω(AW,QW), the question words associated with the answer sentence can be filtered and selected to form the set.

$$\begin{aligned} QW_{\mathrm{i}} &= \left\{ \mathrm{QW}_{\mathrm{i,1}}, \mathrm{QW}_{\mathrm{i,2}}, \mathrm{QW}_{\mathrm{i,3}}, \cdots, \mathrm{QW}_{\mathrm{i,j}}, \cdots \right\} \\ AW_{\mathrm{k}} &= \left\{ \mathrm{AW}_{\mathrm{k,1}}, \mathrm{AW}_{\mathrm{k,2}}, \mathrm{AW}_{\mathrm{k,3}}, \cdots, \mathrm{AW}_{\mathrm{k,m}}, \cdots \right\} \end{aligned} \quad (21)$$

$$\begin{aligned} & \mathrm{P}(\mathrm{AW}_{\mathrm{k,m}} \mid \mathrm{QW}_{\mathrm{i,j}}) \\ & = \frac{\mathrm{C}(\mathrm{AW}_{\mathrm{k,1}} \mid \mathrm{QW}_{\mathrm{i,1}}) + \mathrm{C}(\mathrm{AW}_{\mathrm{k,1}} \mid \mathrm{QW}_{\mathrm{i,2}}) \cdots + \mathrm{C}(\mathrm{AW}_{\mathrm{k,1}} \mid \mathrm{QW}_{\mathrm{i,j}})}{\mathrm{C}(\mathrm{QW}_{\mathrm{i,j}})} \quad (22) \\ & \bullet \mathrm{C}(\mathrm{AW}_{\mathrm{k,1}}) \end{aligned}$$

$$\begin{aligned} & \mathrm{IF}\ \mathrm{P}(\mathrm{AW}_{\mathrm{k,m}}, \mathrm{QW}_{\mathrm{i,j}}) > \varpi(AW, QW) \\ & Then\ \mathrm{AW}_{\mathrm{k,m}}, \mathrm{QW}_{\mathrm{i,j}} \in QWAW^{\varpi}{}_{k} \end{aligned} \quad (23)$$

<u>Step (B2): Determination of the vocabulary category similarity</u>

Through the expression items of the target document analysis, this paper establishes the document keyword set ($D_{i,q}$). Moreover, as shown in Equation (24), the expression vocabulary of each expression item is defined as a domain vocabulary. In this step, the VSM (Vector Space Model) Cosine is used to compute the similarity of the document and the set of the answer words and the level of similarity of the answer word set and the document. Sim(D_q|$QWAW_k$), can be determined by Equation (25). If the similarity level is above the threshold value ω and is closer to 1, it means the set has more parsing meanings of the document.

$$\mathrm{D}_{\mathrm{i}} = \left\{ \mathrm{D}_{\mathrm{i,1}}, \mathrm{D}_{\mathrm{i,2}}, \mathrm{D}_{\mathrm{i,3}}, \cdots, \mathrm{D}_{\mathrm{i,q}}, \cdots \right\} \quad (24)$$

$$\begin{aligned} & \mathrm{D}_{\mathrm{q}}{}^{\varpi} = \left[w_1, w_2, \cdots, w_q \right]^T \\ & QWAW^{\varpi}{}_{k} = \left[w_1, w_2, \cdots, w_k \right]^T \quad (25) \\ & \mathrm{Sim}(\mathrm{D}_{\mathrm{q}} \mid \mathrm{QWAW}_{\mathrm{k}}) = \frac{D^{\varpi}{}_{q} \cdot QWAW_k^{\varpi}}{|D^{\varpi}{}_{q}| \cdot |QWAW_k^{\varpi}|} \end{aligned}$$

In addition, four ways of setting the threshold values are proposed for users in the selection of documents. The threshold value can be set as the average, the median, the quartile or the direct definition. If it is above the threshold value ω(D_q,$QWAW_k$), it means that the domain vocabulary is connected with the document and it is placed in the reserve document set ($ReserveDoc\_Set_d$).

4. KNOWLEDGE DOCUMENT RETRIEVAL SYSTEM

In this web-based system, common users can upload the knowledge document; then, the system administrator can set the system parameters and add new question word and answer words, and thus implement the analysis of the expression items and Q&A analysis of the knowledge document.

In order to verify the feasibility of the knowledge document retrieval system in the practical application, this study used ergonomic technology reports (i.e., knowledge documents) from Institute of Occupational Safety and Health website for verification and applies the kernel modules of the system (including “Q&A analysis”) to demonstrate feasibility of the proposed methodology and the developed system. The common users may realize the function of “knowledge document uploading” through “knowledge document management module”. After the knowledge document uploading, the system administrator can perform “knowledge document expression item parsing module” to capture the conceptual sentences of various expression items of the knowledge document (as shown in Figure 1 and Figure 2). According to the relevant conceptual sentences captured by the expression items, such as the conceptual sentences of the expression item of “operation identity” including “…the joint lifting by the operators…”, the system can obtain the keywords of the document such as “storage box” and “handling”, based on which the system can analyze the question goal and the relevant answer word combinations (as shown in Figure 3). According to the analysis of the question word, answer sentences and answer words matching, the system can obtain the relevant answer words of the question word “pain” such as “construction industry” and “age” (as shown in Figure 4).

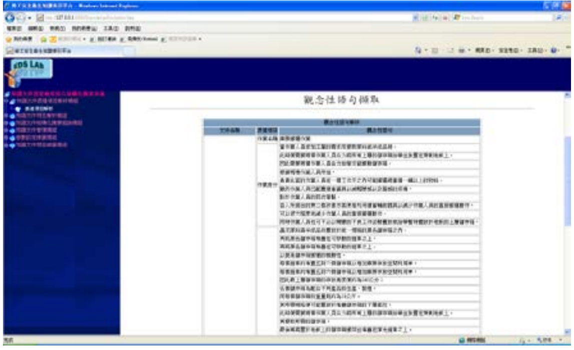

Figure 1: Document expression items analysis result(1)

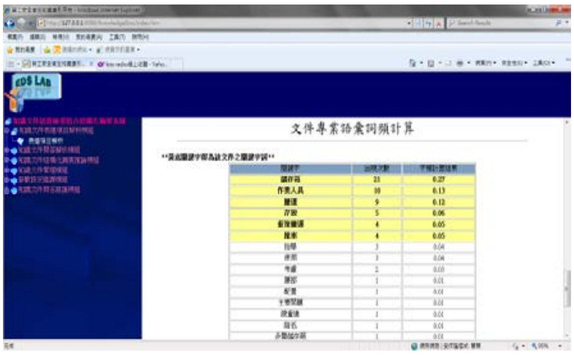

Figure 2: Document expression items analysis result(1)

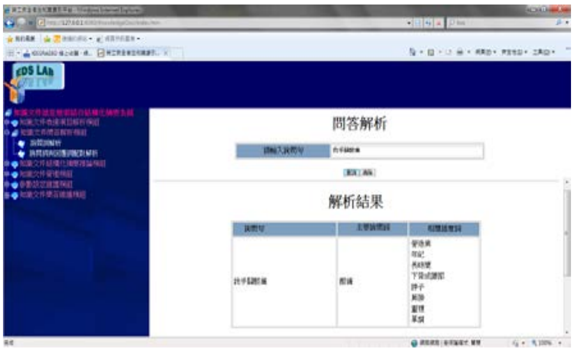

Figure 3: Q&A analysis result (1)

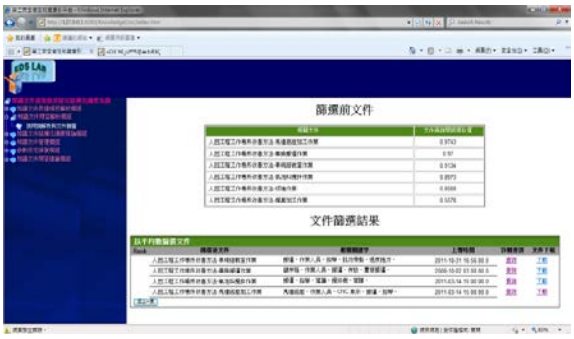

Figure 4: Q&A analysis result (2)

5. CONCLUSIONS

As the specific knowledge fields are too professional, common users can hardly know and define the key words. As a result, the document search will take more time and have more obstacles. Therefore, it can easily affect the selection of documents due to personal browsing and reading preferences, and thus reducing the knowledge sharing effectiveness of the knowledge websites. Hence, this study proposes a semantic analysis technology for knowledge document and establishes a web-based system to confirm the feasibility of the methodology and model. As the verification results have suggested, the system can process knowledge document semantic Q&A and realize the semantic association of general vocabulary and professional vocabulary through Q&A semantic analysis. Hence, common users can search domain knowledge documents by colloquial query string and get the relevant knowledge documents to enhance the reading and selection of users. In this way, it can help users to access to the information on the website of IOSH, and thus enhancing the domain knowledge document search effectiveness.

REFERENCES

1. Cao, Y. G., Liu, F., Simpson, P., Antieau, L., Bennett, A. Cimino, J. J., Ely, J. and Yu, H., 2011, AskHERMES: An Online Question Answering System for Complex Clinical Questions, *Journal of Biomedical Informatics*, Vol. 44, No. 2, pp. 277-288.
2. Dorr, B. J. and Gaasterland, T., 2007, “Exploiting aspectual features and connecting words for summarization-inspired temporal-relation extraction,” *Information Processing and Management*, Vol. 43, No. 6, pp. 1681-1704.
3. Dunlavy, D. M., O’Leary, D. P., Conroy, J. M. and Schlesinger, J. D., 2007, “QCS: A system for querying, clustering and summarizing documents,” *Information Processing and Management*, Vol. 43, No. 6, pp. 1588-1605.
4. Han, K. S., Song, Y. I., Kim, S. B. and Rim, H.

C., 2007, “Answer extraction and ranking strategies for definitional question answering using linguistic features and definition terminology,” *Information Processing & Management*, Vol. 43, No. 2, pp. 353-364.

5. Huang, M., Zhu, X. and Li, M., 2006, “A hybrid method for relation extraction from biomedical literature,” *International Journal of Medical Informatics*, Vol. 75, No. 6, pp. 443-455.
6. Jones, M. and Love, B. C., 2007, “Beyond common features: The role of roles in determining similarity,” *Cognitive Psychology*, Vol. 55, No. 3, pp. 196-231.
7. Ko, Y., Park, J. and Seo, J., 2004, “Improving text categorization using the importance of sentences,” *Information Processing and Management*, Vol. 44, No. 1, pp. 65-79.
8. Oh, H. J., Myaeng, S. H. and Jang, M. G., 2012, “Effects of answer weight boosting in strategy-driven question answering,” *Information Processing and Management*, Vol. 48, No. 1, pp. 83-93.
9. Oh, H. J., Sung, K. Y., Jang, M. G. and Myaeng, S. H., 2011, “Compositional question answering: A divide and conquer approach, *Information Processing and Management*, Vol. 47, No. 6, pp. 808-824.
10. Teng, C., Xiong, N., He, Y., Yang, L. T. and Liu, D., 2010, “A behavioural mode research on user-focus summarization,” *Mathematical and Computer Modelling*, Vol. 51, No. 7-8, pp. 985-994.

ABOUT THE AUTHORS

Shih-Ting Yang is an assistant professor in the Department of Information Management at Nanhua University. Dr. Yang received his Ph.D. in Industrial Engineering and Engineering Management at National Tsing-Hua University and his research interests are knowledge management and mobile commerce.

Yu-Ting Gong is a graduate student in the Department of Information Management at Nanhua University. Her research interests are knowledge management and knowledge retrieval.

(Received October 2014, revised November 2014, accepted December 2014)

知識文件語意解析技術

楊士霆[*]、龔鈺婷
南華大學資訊管理學系
嘉義縣大林鎮南華路一段55號

摘要

人們已習慣透過網路搜尋方式取得資訊與知識，目前雖有關鍵字搜尋、文件分類等搜尋窗口以縮小搜尋範圍，但面對特定領域網站時，若無相關領域背景之知識搜尋者仍需不斷嘗試以取得回饋，關鍵字搜尋與文件分類乃缺乏有效地協助。因此，本研究乃以「勞工安全知識網」為基礎，針對特定領域知識文件先行解析知識文件之表達方式以及相關性語彙，進而建構知識語彙庫，並透過本研究所建構之「知識文件問答解析技術」進行知識文件之專有名詞與搜尋字詞之語意關聯並取得句有相關性文件，以避免個人閱讀偏好影響文件之篩選，進而加強知識網搜尋知識文件技術以提高專業知識網（如勞工安全知識網）知識分享之成效。

關鍵字：勞工安全知識網、知識管理、資料探勘、語意分析

（*聯絡人：stingyang@nhu.edu.tw）