A KNOWLEDGE DOCUMENT STRUCTURED SUMMARIZATION MODEL

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

ABSTRACT

It is a common practice to acquire information and knowledge from the Internet; thus, keyword searching, document classification and other technologies have been developed to facilitate document searching. Although the search engines can narrow down the scope of search, knowledge demanders without domain knowledge in the specific fields need to continuously search and receive feedbacks. Hence, this paper develops a Knowledge Structured Document Summarization model to analyze the ergonomic technology reports from the website of "Institute of Occupational Safety and Health". Then the expressions and domain vocabulary of knowledge documents can be captured to develop the domain vocabulary database via Knowledge Document Analysis (KDA) module. Secondly, through the Conceptual Sentence Acquisition (CSA) module, the conceptual or representative sentences of domain documents can be derived and serve as candidate sentences for structured summarization. Finally, the Document Structured Summarization (DSS) module is used to calculate and retrieve representative sentences of the documents and integrate them into document abstract for knowledge demanders. That is, through this model, knowledge demanders can directly read the desired parts according to problems to ensure demanders can find document they want within a short time. In addition, a web-based system is developed based on the proposed model. Finally, the improvement reports (knowledge documents) collected from the "Institute of Occupational Safety and Health" are used for verification and the kernel modules of the system are applied to demonstrate feasibility of the proposed methodology and the developed system.

Keywords: Institute of Occupational Safety and Health, Knowledge Management, Data Mining, Document Summarization Technology

1. INTRODUCTION

Due to the development of Internet, it is a common practice to acquire information and knowledge online. This may easily lead to problems such as excessive amount of information that gives rise to keyword searching, document classification and other relevant technologies to facilitate searching. In addition, websites can be built to summarize and share documents in relevant fields. In other words, when the knowledge demanders want to inquire relevant information, they can intuitively search needed information on websites to save searching time. Although there are websites of specific fields as the search engines can narrow down the search scope, approaches such as document classification and keyword search to help getting the needed documents, knowledge demanders without domain expertise in specific fields may be unable to search by keywords but generalization. In addition, most of the document abstracts have no uniform format or the control of word number, forcing users to read each document abstract resulting in poor effects in sharing the knowledge of the websites. In summary of the above, the existing operational model is as shown in Figure 1.

In order to understand the user needs and provide real feedbacks regarding the problem, this paper believes that user selection thinking model should be strengthened regarding the document presentation in addition to the strengthening of the keyword searching and subject classification technologies to help knowledge demanders to rapidly and actually get the documents they need. Based on the website of Institute of Occupational Safety and Health, this paper analyzes the ergonomic improvement reports to understand the expression, presentation contents and related domain vocabulary

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

المنارة للاستشارات
www.manaraa.com

of knowledge documents. On the basis of expression items and representative vocabulary, this paper develops a Knowledge Document Structured Summarization model to help knowledge demanders for determination and selection of document contents.

Therefore, the proposed model can enhance keyword semantic determination by representative vocabulary of documents and help knowledge demanders to read document abstract in focused direction with the structured summarization concept, and thereby increasing the knowledge sharing effectiveness of the website of Institute of Occupational Safety and Health. The To-Be model proposed in this study is shown in Figure 2.

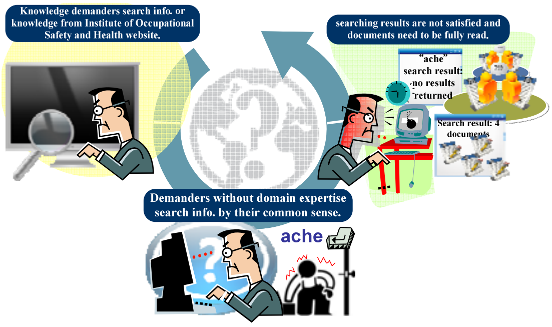
Knowledge demanders search info. or knowledge from Institute of Occupational Safety and Health website.
searching results are not satisfied and documents need to be fully read.
"ache" search result: no results returned
Search result: 4 documents
Demanders without domain expertise search info. by their common sense.
ache

Figure 1: AS-IS model of knowledge document search

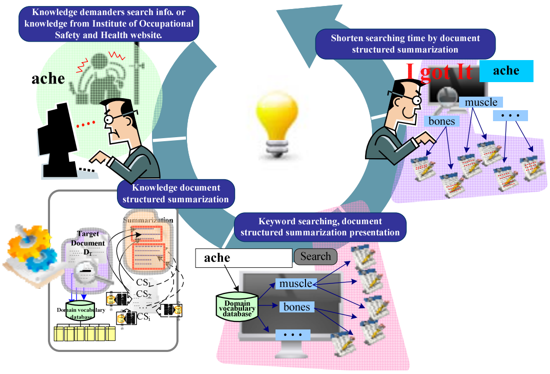
Knowledge demanders search info. or knowledge from Institute of Occupational Safety and Health website.
ache
Shorten searching time by document structured summarization
I got It
ache
muscle
bones
• • •
Knowledge document structured summarization
Summarization
Target Document D_T
CS_1
CS_2
CS_i
Domain vocabulary database
Keyword searching, document structured summarization presentation
ache
Search
Domain vocabulary database
muscle
bones
• • •

Figure 2: TO-BE model of knowledge document search

2. LITERATURE REVIEW

This study involves in two major topics of "Document Summarization Technology Application", and "Document Summarization Technology Development". Literature review and discussions regarding the two major topics are as illustrated below.

2.1 Document Summarization Technology Application

Document summarization is mainly applied in retrieval system and Q&A system; the summarization can strengthen the selection by information demander to get valid information. In the retrieval system, Lorch et al. [8] classified by problems (e.g., classification by What and How questions) and proposed MedQA (Medical Definitional Question Answering System) for medical field to generate the multi-document summarization for the information demander to get valid information from the problems. In addition, some studies established summaries with the search inquiry as the conditions. Sweeney et al. [14] created abstracts based on inquiry conditions and avoid limitations by giving more information through the "Show Me More" approach to represent the documents. Li and Chen [7] proposed an individualized fragment acquisition technology based on semantic analysis technology by getting the opening and ending of text fragments through the statistical linguistic model. Among the Q&A system applications, Cao et al. [2] proposed the AskHERMES for the medical field, which can acquire key points from the complex, unformatted clinical medical reports, and summarize the key points to get answers to clinical medical problems.

As shown above, document summarization technology is mostly applied in fields of excessive data amount or excessive documents of similar themes. Moen [9] proposed an automated summarization and retrieval system. In this system, users can find viewpoints and consultancy according to correlated cases. Elhadad et al. [5] proposed a unified summarization model for the medical field to solve the problem of huge data of medical literature. The retrieval results are summarized to help users more effectively browse the literature. Uzuner et al. [16] used the UMLS (Unified Medical Language System) to define disease types and symptoms to establish the medical report correlated by semantic relational classification, which can also serve as an index to medical records.

In addition to addressing the problem of information overloading, summarization technology can also be applied in news and forums fields. Based on the limitation on the textual length of the news title, Zajic et al. [20] applied document compression technology in multi-document summarization. Yang and Wang [19] and Bouras et al. [1] developed the automated text summarization of news documents for mobile device users to address the problem of difficulty in presenting huge documents in handheld mobile devices.

The document summarization approaches are different according to document types. It can be divided by document structure into "regularization document" and the "free-form document". Xie and Liu [17] obtained the regularization document summarization rules from the meeting minutes for the integration of the minutes of conference of multiple attendants and speakers. Regarding the no-title formatted documents, Pattern [12] acquired important sentences to establish the document abstract by

المنارة للاستشارات
www.manaraa.com

contextual information and mixed statistics. In addition, Tao et al. [15] conducted the automated text summarization by identifying specific themes from the forum frequently asked questions.

2.2 Document Summarization Technology Development

Topics relating to summarization technology development can be discussed in natural language analysis, lexical chain, latent semantic analysis and subject tree technologies.

Natural language analysis calculates the importance and representativeness of sentences by using document's grammatical structure, position, and vocabulary, and acquires the critical information to build the document abstract. Zhou [21] used the natural processing language in the medical field to establish the semantic structure and proposed an automated medical terminology acquisition model to build visualized summaries. Legara et al. [6] generated four rules based on the writing styles of the authors: syntax, structure, vocabulary and content features, and proposed a summarization method of column articles by automated classification based on authorship.

Unsupervised learning is one of the machines learning methods. The learning patterns are mainly of latent semantic analysis and spatial vector models. Based on latent semantic analysis, Chan [3] proposed a quantified model for the acquisition of the most representative sentences, which can strengthen the continuity and semantic relevance of text summarization by latent semantic strengthening of human understanding model and organizing representative vocabulary network diagram. Dahab et al. [4] proposed the semantic analysis model by shallow layer semantic analysis, and proposed a natural ontological model (TextOntoEx) based on semantics. The semantic models include abstract, verb groups and other major elements and match the model with each document to acquire the unclassified relationships in the documents [4]. Nomoto and Matsumoto [10] used the spatial vector model to propose an unsupervised diversified multiple-document summarization technique to find relevant thematic documents by K-means and MDLP (Minimum Description Length Principle) and acquiring most representative sentences to automatically form the diversified multiple-document text abstracts.

In automated text summarization, most studies use vocabulary features, document structure and other rules as the training basis for training through the application of SVM, latent Markov chain, N conjunctions and subject tree. Ruiz-Casado et al. [13] used vocabulary models to automatically identify semantic relationships to form the whole-part relationships in specific fields based on documents. Pattern [12] also used the bi-gram searching technique to acquire critical sentences to form the document abstracts through contextual information and mixed statistics. Regarding applications, Ouyang et al. [11] pre-defined seven review criteria based on document structure (e.g., semantics, relevance to the theme, frequency of wording, ending meaning, and position of the sentence), and implemented sentence relevance by SVR model learning. Xie and Liu [17] used the regression supervised learning to acquire the most representative sentences to summarize meeting minutes through the SVM (Support Vector Machines) supervised learning method.

3. KNOWLEDGE DOCUMENT STRUCTURED SUMMARIZATION MODEL

The proposed "knowledge document structured summarization model" is based on the ergonomic technology reports (i.e., knowledge document) provided on the website of Institute of Occupational Safety and Health This model first analyzes the structural features of the knowledge document and figures out eight expression items, 20 detailed expression items to build the domain vocabulary database. Based on the domain vocabulary database, the vocabulary comparison rules corresponding to knowledge document are built to obtain the conceptual sentences for belonging to each set. Finally, in accordance with the rules of structured summarization, the representative sentences from the sets according to the structured summarization rules can be acquired and integrate them into document abstract for knowledge demanders. Hence, the architecture of this model can be divided into three parts as shown in Figure 3 including Part1 Knowledge Document Analysis (KDA) module, Part2 Conceptual Sentence Acquisition (CSA) module and Part3 Document Structured Summarization (DSS) module.

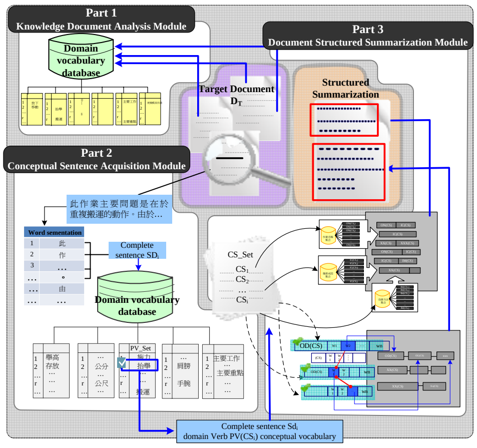

Figure 3: Architecture of knowledge document structured summarization model

www.manaraa.com

3.1 Knowledge Document Analysis (KDA) Module

The analysis of the contents of the ergonomic technology reports can be divided into eight expression items, and 20 expression sub-items. This paper builds a domain vocabulary set on the basis of the 20 expression sub-items. To judge the conceptual sentences contained in the knowledge documents, “conjunctive vocabulary”, “general verb vocabulary” should be added to structuralize the expression sub-items such as the operation definition, operation goal and improvement purpose. The expression of the job extent, frequency, and appearance, numerical vocabulary and unit measurement vocabularies including “age unit vocabulary”, “length unit vocabulary” and “money unit vocabulary” should also be added. Therefore, domain vocabulary database contains 29 vocabulary sets (e.g., Conjunctions Set, General Verb Set and Professional Pose Set) as illustrated below:

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.

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)).

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

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

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

www.manaraa.com

✓ 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)).

$$\text{IF}\,\mathrm{SD}_{i,j}\ \text{exist in}\begin{pmatrix}\text{ORT(CS) and N(CS)}\\ \text{and AU(CS)}\end{pmatrix}\forall j \tag{7}$$

$$\text{Then}\ \ \mathrm{SD}_i\ \in \mathrm{OR\_Set}$$

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.

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

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

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

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

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).

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

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

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

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

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

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

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.

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

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

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

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

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

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

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).

$$\text{IF}\,\mathrm{SD}_{i,j}\ \text{exist in}\big(\text{IC(CS) and B(CS)}\ \big)\forall j \tag{16}$$

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

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

المنارة للاستشارات

www.manaraa.com

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.

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

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

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

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

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 structured summarization.

3.3 Document Structured Summarization (DSS) Module

For the completeness of the textual descriptions, the DSS module can be divided into two parts including the brief part and the detailed part.

3.3.1 Establishment of the Brief Part

The brief part of the structured summarization is mainly to calculate the centrality of the sentences before carrying out the selection by sentence structural integrity. If the sentence contains a variety of conceptual vocabularies, it means the sentence is representative of the text, and thus the sentence is listed as a candidate sentence for sentence structural strength calculation. The sentence structural strength calculation is to consider the readability of the abstract, hence, the sentence structural strength (namely, the relevance between the subjective, predictive, and verb of the sentence) is calculated to acquire sentences of completeness. The set of the acquired sentences include: the operation name set, the injury cause set and the improvement method set.

<u>Step (B1): Calculation of the Centrality of Conceptual Sentences</u>

The conceptual sentences of the expression items (namely, ON(CS$_i$), IC(CS$_i$) and IM(CS$_i$)) are segmented into word groups of 2 to 6 words to form the sets (namely, ON(CS$_{i,j}$), IC(CS$_{i,j}$) and IM(CS$_{i,j}$)). The judgment of the centrality of the expression items is as determined by Equations (21), (22) and (23). The comparison of vocabularies by words is conducted to accumulate centrality scores represented by 1 for existence and 0 for non-existence. Finally, Score ON(CS), Score IC(CS) and Score TM(CS) are used to present the centrality scores of the conceptual sentences, and the vocabularies of centrality are placed in the candidate vocabulary sets of various expression items.

- As show in Equation (21), if a conceptual sentence of operation name also contains conceptual vocabularies of operation goal or operation tool, it means the sentence is in line with the report subject with centrality.

$$\begin{aligned}
&\text{ON}(\text{CS}_i) = \left\{\text{CS}_{i,1}, \text{CS}_{i,2}, \text{CS}_{i,3}, \cdots, \text{CS}_{i,j}, \cdots\right\}\\
&\text{IF ON}(\text{CS}_{i,j})\ \text{exist in OG(CS)}\ \forall j\\
&\text{Then } W_1 = 1\ \text{, Otherwise } W_1 = 0\\
&\text{IF ON}(\text{CS}_{i,j})\ \text{exist in OT(CS)}\ \forall j\\
&\text{Then } W_2 = 1\ \text{, Otherwise } W_2 = 0\\
&\text{Score ON}(\text{CS}_i) = W_1 + W_2\\
&\text{IF } 1 < \text{ScoreON}(\text{CS}_i) \le 2 \text{ Then ON}(\text{CS}_i)\\
&\in \text{CandidateON}
\end{aligned} \qquad (21)$$

- As show in Equation (22), if a conceptual sentence of injury cause also contains conceptual vocabularies of domain verb, operation tool or operation hour, it means the sentence is in line with the report subject with centrality.

$$\begin{aligned}
&\text{IC}(\text{CS}_i) = \left\{\text{CS}_{i,1}, \text{CS}_{i,2}, \text{CS}_{i,3} \cdots, \text{CS}_{i,j}, \cdots\right\}\\
&\text{IF IC}(\text{CS}_{i,j})\ \text{exist in PV(CS)}\ \forall j\\
&\text{Then } W_3 = 1\ \text{, Otherwise } W_3 = 0\\
&\text{IF IC}(\text{CS}_{i,j})\ \text{exist in B(CS)}\ \forall j\\
&\text{Then } W_4 = 1\ \text{, Otherwise } W_4 = 0\\
&\text{Score IC}(\text{CS}_i) = W_3 + W_4\\
&\text{IF } 1 < \text{ScoreIC}(\text{CS}_i) \le 2 \text{ Then IC}(\text{CS}_i)\\
&\in \text{CandidateIC}
\end{aligned} \qquad (22)$$

- As show in Equation (23), if a conceptual sentence of improvement method also contains conceptual vocabularies of improvement purpose, operation tool vocabulary or review verb, it

المنارة للاستشارات
www.manaraa.com

means the sentence is in line with the report subject with centrality.

$$\begin{aligned}
&\mathrm{IM}(CS_i) = \{CS_{i,1}, CS_{i,2}, CS_{i,3} \cdots, CS_{i,j}, \cdots\}\\
&\text{IF } \mathrm{IM}(CS_{i,j}) \text{ exist in } \mathrm{IG}(CS)\ \forall j\\
&\text{Then } W_5 = 1\ \text{, Otherwise } W_5 = 0\\
&\text{IF } \mathrm{IM}(CS_{i,j}) \text{ exist in } \mathrm{OT}(CS)\ \forall j\\
&\text{Then } W_6 = 1\ \text{, Otherwise } W_6 = 0\\
&\text{IF } \mathrm{IM}(CS_{i,j}) \text{ exist in } \mathrm{RV}(CS)\ \forall j\\
&\text{Then } W_7 = 1\ \text{, Otherwise } W_7 = 0\\
&\text{Score } \mathrm{IM}(CS_i) = W_5 + W_6 + W_7\\
&\text{IF } 2 < \text{ScoreIM}(CS_i) \le 3 \text{ Then } \mathrm{IM}(CS_i)\\
&\in \text{CandidateIM}
\end{aligned} \qquad (23)$$

Step (B2): Calculation of Structural Strength of Conceptual Sentences

According to Equations (21), (22) and (23), conceptual sentences with centrality have been acquired to various candidate vocabulary sets. Next, the structural strength scores of various sentences in the candidate vocabulary sets $\mathrm{Score}\,(TS, TV, TO)$ should be calculate. There are three types of sentence structures according to the different presentations of subject (TS), verb (TV) and objective (TO) of various expression items. First, the sentence of complete structure, namely, the sentence has the combinations of three structural elements, and the structural score of the sentence $\mathrm{Score}\,(TS, TV, TO)$ is 2. Second, semi-structured sentence, if a sentence has a verb integrated with the objective or the subject, then it means the sentence has to be linked with other sentences, and then the sentence structural $\mathrm{Score}\,(TS, TV, TO)$ is 1. Third, the sentence of incomplete structure, namely, the sentence has no verb and is hard to be linked with the preceding or following sentence, and then the sentence structural $\mathrm{Score}\,(TS, TV, TO)$ is 0. The structural strength calculation of various expression items is as shown in Equations (24) to (32).

✓ As shown in Equations (24) to (26), the subject parts (TS) of the conceptual sentences of operation name can be judged by the operation name vocabulary ON(CS); the verb part (TV) can be judged by the domain verb vocabulary PV(CS), general verb vocabulary GV(CS); the objective part (TO) can be judged by the operation tool vocabulary OT(CS). The sentence is judged by the sequence of having verb part structure (TV), followed by the objective part (TO) and the subject part (TS).

$$\begin{aligned}
&\mathrm{ONCS}_i = \{CS_{i,1}, CS_{i,2}, CS_{i,3}, \cdots, CS_{i,j}, \cdots\}\\
&\text{IF } \mathrm{ON}(CS_{i,j}) \text{ exist in } (\mathrm{PV}(CS)) \text{ OR } (\mathrm{GV}(CS)) \forall j\\
&\text{And } \mathrm{ON}(CS_{i,j}) \text{ exist in } (\mathrm{OT}(CS))\ \forall j\\
&\text{And IF } \mathrm{ON}(CS_{i,j}) \text{ exist in } (\mathrm{ON}(CS))\ \ \forall j\\
&\text{Then } (\mathrm{Score}(TS,TV,TO)) = 2
\end{aligned} \qquad (24)$$

$$\begin{aligned}
&\mathrm{ONCS}_i = \{CS_{i,1}, CS_{i,2}, CS_{i,3}, \cdots, CS_{i,j}, \cdots\}\\
&\text{IF } \mathrm{ON}(CS_{i,j}) \text{ exist in } (\mathrm{PV}(CS)) \text{ OR } (\mathrm{GV}(CS)) \forall j\\
&\text{And } \mathrm{ON}(CS_{i,j}) \text{ exist in } (\mathrm{OT}(CS))\ \forall j\\
&\text{Or IF } \mathrm{ON}(CS_{i,j}) \text{ exist in } (\mathrm{ON}(CS))\ \ \forall j\\
&\text{Then } (\mathrm{Score}(TS,TV,TO)) = 1
\end{aligned} \qquad (25)$$

$$\begin{aligned}
&\mathrm{ONCS}_i = \{CS_{i,1}, CS_{i,2}, CS_{i,3}, \cdots, CS_{i,j}, \cdots\}\\
&\text{IF } \mathrm{ON}(CS_{i,j}) \text{ not exist in } (\mathrm{PV}(CS))\\
&\text{OR } (\mathrm{GV}(CS)) \forall j\\
&\text{Then } (\mathrm{Score}(TS, TV, TO)) = 0
\end{aligned} \qquad (26)$$

✓ As shown in Equations (27) to (29), the subject part (TS) of the conceptual sentences of the injury cause can be judged by the injury cause vocabulary IC(CS); the verb part (TV) can be determined by the domain verb vocabulary PV(CS), general verb vocabulary GV(CS); the objective part (TO) can be judged by the injury cause vocabulary IC(CS).

$$\begin{aligned}
&\mathrm{ICCS}_i = \{CS_{i,1}, CS_{i,2}, CS_{i,3}, \cdots, CS_{i,j}, \cdots\}\\
&\text{IF } \mathrm{IC}(CS_{i,j}) \text{ exist in } (\mathrm{PV}(CS)) \text{ OR } (\mathrm{GV}(CS))\ \forall j\\
&\text{And } \mathrm{IC}(CS_{i,j}) \text{ exist in } (\mathrm{IC}(CS))\ \forall j\\
&\text{And IF } \mathrm{IC}(CS_{i,j}) \text{ exist in } (\mathrm{IC}(CS))\ \ \forall j\\
&\text{Then } (\mathrm{Score}(TS,TV,TO)) = 2
\end{aligned} \qquad (27)$$

$$\begin{aligned}
&\mathrm{ICCS}_i = \{CS_{i,1}, CS_{i,2}, CS_{i,3}, \cdots, CS_{i,j}, \cdots\}\\
&\text{IF } \mathrm{IC}(CS_{i,j}) \text{ exist in } (\mathrm{PV}(CS)) \text{ OR } (\mathrm{GV}(CS))\ \forall j\\
&\text{And } \mathrm{IC}(CS_{i,j}) \text{ exist in } (\mathrm{IC}(CS))\ \forall j\\
&\text{OR IF } \mathrm{IC}(CS_{i,j}) \text{ exist in } (\mathrm{IC}(CS))\ \ \forall j\\
&\text{Then } (\mathrm{Score}(TS,TV,TO)) = 1
\end{aligned} \qquad (28)$$

$$\begin{aligned}
&\mathrm{ICCS}_i = \{CS_{i,1}, CS_{i,2}, CS_{i,3}, \cdots, CS_{i,j}, \cdots\}\\
&\text{IF } \mathrm{IC}(CS_{i,j}) \text{ not exist in } (\mathrm{PV}(CS))\\
&\qquad\quad \text{OR } (\mathrm{GV}(CS)) \forall j\\
&\text{Then } (\mathrm{Score}(TS, TV, TO)) = 0
\end{aligned} \qquad (29)$$

✓ As shown in Equations (30) to (32), the subject part of the conceptual sentence of improvement method (TS) can be judged by the operation title vocabulary ORT(CS); the verb part of (TV) can be judged by the domain verb vocabulary PV(CS), general verb vocabulary GV(CS), review verb vocabulary RV(CS); the objective part (TO) can be judged by the operation tool

المنارة للاستشارات
www.manaraa.com

vocabulary OT(CS).

$$\begin{aligned}
&\mathrm{IMCS_i} = \{\mathrm{CS_{i,1}}, \mathrm{CS_{i,2}}, \mathrm{CS_{i,3}}, \cdots, \mathrm{CS_{i,j}}, \cdots\}\\
&\text{IF IM}(\mathrm{CS_{i,j}})\\
&\ \text{exist in PV(CS) OR GV(CS) OR RV(CS)}\ \forall j\\
&\text{And IM}(\mathrm{CS_{i,j}})\ \text{exist in}\ (\mathrm{OT(CS)})\ \forall j\\
&\text{And IF IC}(\mathrm{CS_{i,j}})\ \text{exist in}\ (\mathrm{ORT(CS)})\ \forall j\\
&\text{Then}\ (\mathrm{Score(TS, TV, TO)}) = 2
\end{aligned} \tag{30}$$

$$\begin{aligned}
&\mathrm{IMCS_i} = \{\mathrm{CS_{i,1}}, \mathrm{CS_{i,2}}, \mathrm{CS_{i,3}}, \cdots, \mathrm{CS_{i,j}}, \cdots\}\\
&\text{IF IM}(\mathrm{CS_{i,j}})\ \text{exist in PV(CS)}\\
&\text{OR GV(CS) OR RV(CS)}\ \forall j\\
&\text{And IM}(\mathrm{CS_{i,j}})\ \text{exist in}\ (\mathrm{OT(CS)})\ \forall j\\
&\text{And IF IC}(\mathrm{CS_{i,j}})\ \text{exist in}\ (\mathrm{ORT(CS)})\ \forall j\\
&\text{Then}\ (\mathrm{Score(TS, TV, TO)}) = 1
\end{aligned} \tag{31}$$

$$\begin{aligned}
&\mathrm{IMCS_i} = \{\mathrm{CS_{i,1}}, \mathrm{CS_{i,2}}, \mathrm{CS_{i,3}}, \cdots, \mathrm{CS_{i,j}}, \cdots\}\\
&\text{IF IM}(\mathrm{CS_{i,j}})\ \text{not exist in PV(CS) OR GV(CS)}\\
&\qquad\quad \text{OR RV(CS)}\ \forall j\\
&\text{Then}\ (\mathrm{Score(TS, TV, TO)}) = 0
\end{aligned} \tag{32}$$

Step (B3): Calculation of the Weights of the Conceptual Sentences

To highlight the sentence centrality, this paper gives the centrality and structure scores weights WS_1 and WS_2. This indicates that the brief part of the summarization is to mainly consider conceptual centrality followed by the structure of the sentence to acquire the sentences with centrality scores and complete structure into the selected vocabulary sets of various expression items. The weights of the expression items are calculated as shown in Equations (33) to (35).

$$\begin{aligned}
&\text{TotalScore ON}(\mathrm{CS_i}) = (\text{Score ON}(\mathrm{CS_i}) * \mathrm{WS_1})\\
&+ ((\mathrm{Score(TS, TV, TO)}) * \mathrm{WS_2})\\
&\text{IF LowerLimit} < \text{TotalScoreON}(\mathrm{CS_i})\\
&\quad \le \text{UpperLimit Then ON}(\mathrm{CS_i}) \in \text{SelectON}
\end{aligned} \tag{33}$$

$$\begin{aligned}
&\text{TotalScore IC}(\mathrm{CS_i}) = (\text{Score IC}(\mathrm{CS_i}) * \mathrm{WS_1})\\
&+ (\mathrm{Score((TS, TV, TO))} * \mathrm{WS_2})\\
&\text{IF LowerLimit} < \text{TotalScoreIC}(\mathrm{CS_i})\\
&\le \text{UpperLimit Then IC}(\mathrm{CS_i}) \in \text{SelectON}
\end{aligned} \tag{34}$$

$$\begin{aligned}
&\text{TotalScore IM}(\mathrm{CS_i}) = (\text{Score IM}(\mathrm{CS_i}) * \mathrm{WS_1})\\
&+ ((\mathrm{Score(TS, TV, TO)}) * \mathrm{WS_2})\\
&\text{IF LowerLimit} < \text{TotalScoreIC}(\mathrm{CS_i})\\
&\le \text{UpperLimit Then IC}(\mathrm{CS_i}) \in \text{SelectON}
\end{aligned} \tag{35}$$

The above steps are to calculate and acquire the weights of the conceptual sentences of operation name, injury cause, improvement method. With sentence centrality and sentence structural completeness as the condition for the brief part, the sentences by the weights of the conditions are selected and acquired to form the structural summarization of the brief part.

3.3.2 Establishment of the Detailed Description Part

The contents of the detailed description part include the description part and the assessment part with detailed descriptions of operation, operation environment and operation hour. According to the above, the set of the sentences in the description part includes: operation set and operation environment set, and the assessment part acquires mainly the set of improvement methods. The description part is mainly of the operation and operation environment sets with implied vocabularies including: operation goal, domain verb, force application level, operation title and frequency; the assessment part is mainly of the improvement method followed by injury cause, operation tool, assessment verb vocabularies.

Step (C1): Acquisition of Conceptual Sentences with Operation Definition Vocabulary

As shown in Equation (36), the method is to use the operation definition (OD) as the judgment vocabulary to conduct the word frequency calculation in comparison to other latent vocabularies to derive similarities of operation definition with various latent vocabularies, and the vocabulary of relatively higher score of $WOD_{i,j}$ as the candidate sentence $Res(S_i)$ to be stored in the set of the candidate sentence set ResSet. The candidate sentences in the set are segmented to expressions with 2 to 6 words. As shown in Equation (37), the sentences are segmented into word groups to form the set represented by Res $(S_{i,j})$ for the vocabulary judgment and comparison.

$$WOD_{i,j} = TFOD_{i,j} \cdot \log \frac{NumSet}{ODF_i} \tag{36}$$

$$\mathrm{Res}(\mathrm{S_i}) = \{\mathrm{S_{i,1}}, \mathrm{S_{i,2}}, \mathrm{S_{i,3}}, \cdots, \mathrm{S_{i,j}}, \cdots\} \tag{37}$$

Step (C2): Calculation of Mutual Dependence of Conceptual Vocabularies

To avoid the semantic conflict of the context, this paper uses the relationship of the occurrence sequence of vocabularies to predict the following sentence. Namely, the linkage probability of the candidate sentence Res (S_i) and the following sentence Res (S_{i-1}) is represented by $P(Res(S_i)|Res(S_{i-1}))$. If the two sentences are correlated, the linkage probability will be higher. Therefore, this step first deduces the sequence of the vocabulary occurrence and then calculate the occurrence of vocabulary and the sequence of occurrence of vocabularies in a paragraph P_i (namely, $CW_k[P_i]$).

www.manaraa.com

If the the k'th word in the Paragraph i $CW_k[P_i]$ contains concepts such as operation goal, domain verb, force application level (as represented by A(CS)), and the next word $CS_{k+1}[P_i]$ also has the same concepts (represented by B(CS)), then it is labeled I[i,k] as 1 to suggest the sequence relationship, otherwise set it as 0 to suggest no relationship in existence (see Equation (38)). After that, the sequence of the occurrence of the words is added up and represented by $F[CW_k, CWk_{+1}]$. According to Equation (39), the occurrence frequency of the k'th word to the k+1'th word can be calculated.

$$\mathrm{I}\left[\mathrm{i,k}\right]=\begin{cases}1\,,\ IF\ CW_k[P_i]\in A(CS)\ \ ,CS_{k+1}[P_i]\in B(CS)\\ 0\,,Otherwise\end{cases} \tag{38}$$

$$F\left[CW_k\,,\mathrm{CW}_{\mathrm{k+1}}\right]=\sum_{all\ i,\,all\ k} I\ \ [i,k] \tag{39}$$

By the above, the occurrence frequency from the k'th word to the k+1'th word can be obtained. The frequency of the word occurring at the beginning and ending of the paragraph is to judge the overall sequence of the occurrence of the word as represented by the ratio of $P^{From}(k)$ and $P^{To}(k)$ $R[CW_k]$ as shown in Equations (40) to (42). According to the ranking of the ratio $R[CW_k]$, the occurrence sequence from the k'th word to the k+1'th word can be obtained, and the overall sequence in accordance with ratio, SCSW can be represented as shown in Equation (43).

$$P^{From}(k)=\sum_{all\ k+n} F\,[CW_k\,,CW_{k+n}\,] \tag{40}$$

$$P^{To}(k)=\sum_{all\ k} F\,[CW_{k+n}\,,CW_k\,] \tag{41}$$

$$R[CW_k\,]=\frac{P^{From}(k)}{P^{To}(k)} \tag{42}$$

$$\begin{aligned}SCSW=&\mathrm{CW}'_{\mathrm{k\text{-}1}}\rightarrow\mathrm{CW}'_{\mathrm{k}}\rightarrow\cdots\rightarrow\mathrm{CW}'_{\mathrm{k+n}}\rightarrow\cdots\\ &Where\,\mathrm{R}[\mathrm{CW}'_{\mathrm{k\text{-}1}}]\geq\mathrm{R}[\mathrm{CW}'_{\mathrm{k}}]\geq\cdots\geq\mathrm{R}[\mathrm{CW}'_{\mathrm{k+n}}]\geq\cdots\end{aligned} \tag{43}$$

Step (C3): Definition of Conceptual Sentences with Conjunctive Vocabulary

Regarding the acquisition of conceptual sentences with conjunctive vocabulary, at this step, the conjunctive vocabulary (CS) of the domain vocabulary are used for selection. By conjunctive vocabulary, whether the sentence $Res(S_i)$ has the function of connecting sentences can be determined. In the conjunctive vocabulary, the conjunctive words such as: "namely", "hence", "also", "nevertheless", etc. can be obtained. By the logic sequence of the conjunctive words such as "since" and "furthermore" to compare the sequential relationship, the position of the "furthermore" should follow the position of "since" according to semantic logic. Hence, this paper further classifies the conjunctive vocabulary into the link vocabulary LinkC (CS) to smoothen abstract structure. Since the linking word is often the first word of the sentence, at this step, the first word of the sentence as a conjunctive vocabulary or the link vocabulary is judged to determine and acquire the conceptual sentences with conjunctive vocabulary. By Equation (37), the sentence has been segmented into word groups of two to six words. Then, this module judges whether the first meaning word of the sentence exists in the conjunctive vocabulary C(CS) or the link vocabulary LinkC(CS) as one of the conditions to build the structured summarization as shown in Step (B5).

Step (C4): Definition of Opening and Ending Conceptual Sentences

At this step, by Step (B2), the ending words $P^{To}(CW_k)$, the opening words $P^{From}(CW_k)$ and the ratio of the opening and ending $R[CW_k]$ are used to determine whether the word is the opening or ending word. The ending words of the structured summarization are mainly for the expression of goals such as: operation goal vocabulary, operation name vocabulary. Meanwhile, R $[CW_k]$ and the ratio of the word existence in the paragraph are calculated to judge whether the word is an opening one. If the ratio is the maximum value $MaxR[CW_k]$, it means that the word is the opening word of the sentence. In this way, the word is the ending word of the sentence can be determined. In addition, this module can also judges whether the ending word is followed by any punctuation mark (.,;, !, ?), if the sentence is followed by an ending punctuation mark, it means the word is the ending word of the sentence.

Step (C5): Calculation and Acquisition of Conceptual Sentences

The sentence has been segmented into word groups by Equation (37). At this step, the $Res(S_{i,j})$ is used to judge whether the resentence is in line with the conditions to build the structured summarization and represent it by the indicator value of the structured summarization of $Mk_{i,j}$. The principles are described as follows:

- If the candidate sentence $Res(S_i)$ contains the k'th word of the overall sequence of SCSW, then the structured summarization is in line with the indicator value at $Mk_{i,1}=1$.
- If the candidate sentence $Res(S_i)$ contains conjunctive vocabulary C(CS), then the structured summarization is in line with indicator value at $Mk_{i,2}=1$.
- If the candidate sentence $Res(S_i)$ contains conjunctive link vocabulary LinkC (CS), then the structured summarization is in line with the indicator value at $Mk_{i,3}=1$.
- If the candidate sentence $Res(S_i)$ contains the vocabulary review coefficient $R[CW_k]$ and the maximum value of the ratio of the total number of words, then the structured summarization is in line with the indicator value at $Mk_{i,4}=1$.

المنارة للاستشارات
www.manaraa.com

- If the candidate sentence $Res(S_i)$ contains the vocabulary review coefficient $R[CW_k]$ and the minimum value of the ratio of the total number of words, then the structured summarization is in line with the indicator value at $Mk_{i,5}=1$.
- If the candidate sentence $Res(S_i)$ contains the words belonging to the ending market set EndMark_Set, then the structured summarization is in line with the indicator value at $Mk_{i,6}=1$.

In summary of the above, the indicator values $Mk_{i,j}$ of the structured summarization corresponding to the candidate sentence $Res(S_i)$ can be summarized (Equation (44)). The structured summarization conformity indicator value $Mk_{i,j}$ can be used for consideration in the acquisition of candidate sentences as well as confirm the position of the candidate sentence. If the candidate sentence's $Mk_{i,4}$ is 1, the sentence is the beginning of the abstract. By Equations (45) and (46), the total sum of $Mk_{i,2}$ and $Mk_{i,3}$ at $Sum1(Res(S_i))$ are used as the criteria in the selection of the description stage. Finally, the total sum of $Mk_{i,5}$ and $Mk_{i,6}$ at $Sum2(Res(S_i))$ are used as the criteria for the acquisition at the assessment stage.

$$MK=\begin{bmatrix} \mathrm{MK}_{1,1} & \mathrm{MK}_{2,1} & \cdots & \mathrm{MK}_{i,1} \\ \mathrm{MK}_{1,2} & \mathrm{MK}_{2,2} & \cdots & \mathrm{MK}_{i,2} \\ \mathrm{MK}_{1,3} & \mathrm{MK}_{2,3} & \cdots & \mathrm{MK}_{i,3} \\ \mathrm{MK}_{1,4} & \mathrm{MK}_{2,4} & \cdots & \cdots \\ \mathrm{MK}_{1,5} & \mathrm{MK}_{2,5} & \cdots & \mathrm{MK}_{i,5} \\ \mathrm{MK}_{1,6} & \mathrm{MK}_{2,6} & \cdots & \mathrm{MK}_{i,6} \end{bmatrix} \tag{44}$$

$$Sum1(\mathrm{Re}\,s(S_i))=\sum_{j=2}^{3} MK_{i,j} \tag{45}$$

$$Sum2(\mathrm{Re}\,s(S_i))=\sum_{j=5}^{6} MK_{i,j} \tag{46}$$

According to the results of the calculation of vocabulary mutual dependence at Step (C2), the overall ranking sequence of the vocabulary, SCSW can be obtained. In this paper, the Equation (47) SCSW results are used as the main basis and other conditions (namely, conjunctive vocabulary or membership of ending or opening vocabularies) to determine sentence positions for building the structured summarization.

$$\begin{aligned} SCSW &= \mathrm{CW}'_{k\text{-}1}\rightarrow \mathrm{CW}'_{k}\rightarrow\cdots\rightarrow \mathrm{CW}'_{k+n}\rightarrow\cdots \\ &\quad Where\ \mathrm{R}[\mathrm{CW}'_{k\text{-}1}]\geq \mathrm{R}[\mathrm{CW}'_{k}]\geq\cdots \\ &\quad \geq \mathrm{R}[\mathrm{CW}'_{k+n}]\geq\cdots \end{aligned} \tag{47}$$

4. KNOWLEDGE DOCUMENT STRUCTURED SUMMARIZATION SYSTEM

This study uses the example of 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 "brief structured summarization" and "detailed structured summarization") to demonstrate feasibility of the proposed methodology and the developed system. The system operational structure is as shown in Figure 4, and the system application scenarios are described in detail as below.

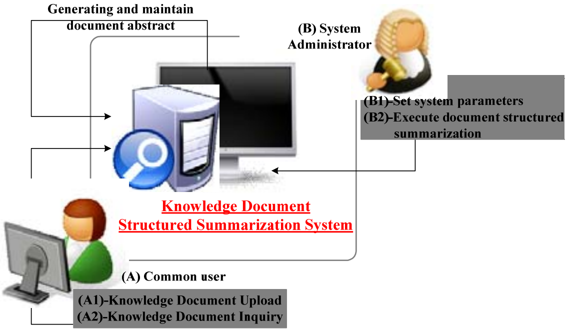
Generating and maintain document abstract
(B) System Administrator
(B1)-Set system parameters
(B2)-Execute document structured summarization
Knowledge Document Structured Summarization System
(A) Common user
(A1)-Knowledge Document Upload
(A2)-Knowledge Document Inquiry

Figure 4: System operational structure

- **Knowledge documents collected and uploaded common users**

Before this system generating the summary of knowledge documents, common users should collect knowledge reports/documents from Institute of Occupational Safety and Health website (see Figure 5) including document entitled "ergonomic hazards prevention technology" etc. After the system administrator has set the system parameters, common users can upload a document entitled "ergonomic hazards prevention technology" without structured abstract by knowledge document upload function of the knowledge document maintenance module to the system for the knowledge document expression item analysis and knowledge document structured summarization.

Figure 5: Ergonomic technology reports

- **Kernel module creates the structured summarization by system administrator**

After the uploading of ergonomic technology reports by the common users, the system administrator can execute the knowledge document structured summarization module including expression items analysis, brief structured summarization and detailed structured summarization functions to establish the structured abstract of the knowledge document. First, the expression items of

المنارة للاستشارات
www.manaraa.com

the knowledge document before generating an abstract should be analyzed via expression items analysis function by system administrator. The system then displays the conceptual statements of the expression items and the contents of the knowledge document (see Figure 6) for the users. After the completion of analysis, the centrality and structural scores of the conceptual statements by weights can be obtained via brief structured summarization function (see Figure 7). At meanwhile, the brief structured abstract of the knowledge document can be generated and maintained in this system (see Figure 8). Finally, system administrator also can executes the detailed structured summarization function to obtain the work frequency calculation results contained in the document (see Figure 9) to acquire the detailed structured summarization by forming the lexical chain of word combinations (see Figure 10). The system also maintains the detailed structured summarization in this system for the inquiry of users.

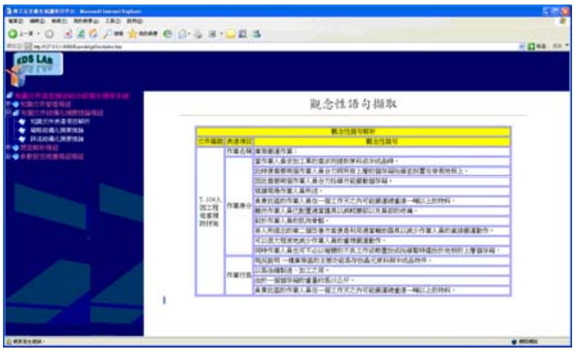

Figure 6: Document expression items analysis result

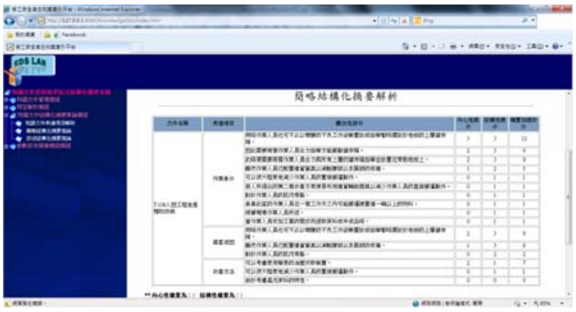

Figure 7: Brief structured summarization result (1)

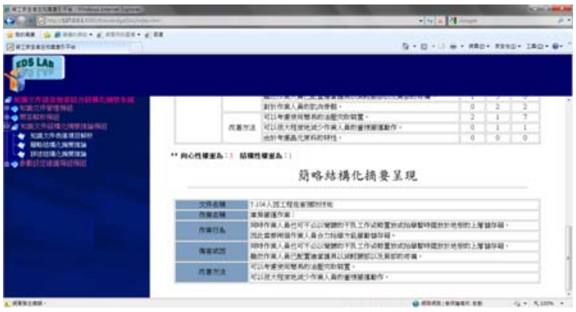

Figure 8: Brief structured summarization result (2)

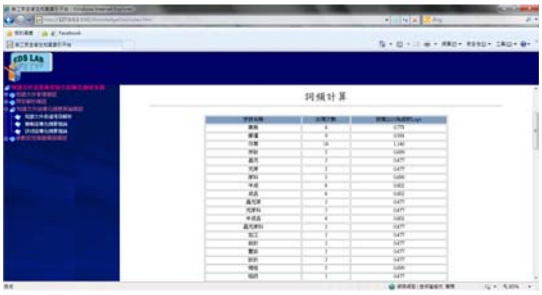

Figure 9: Detailed structured summarization result (1)

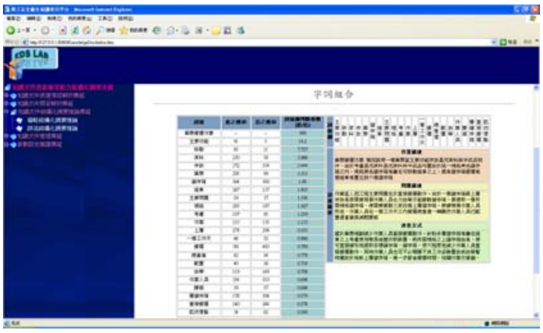

Figure 10: Detailed structured summarization result (2)

5. CONCLUSION

Most of the knowledge document searching methods use keyword inquiring and document classification methods to enhance information acquisition efficiency. However, as faced with specific knowledge fields, the specialty of knowledge would impose obstacle and lengthen the search time. Hence, this paper establishes a knowledge document structured summarization model for knowledge demanders to efficiently obtain needed knowledge documents. The proposed model analyzes the representative vocabularies of the document and builds the domain vocabulary database and vocabulary determination rules accordingly for the sentence acquisition. Then, this model calculates the sequential relationship of words through vocabulary mutual dependence as the conditions for structured summarization. After that, this model acquires the most representative words from the document as the judgment standards and builds the structured summary to enhance the searching by knowledge demanders. In addition, a web-based system is developed based on the proposed model. Finally, a real case namely ergonomic technology reports collected from Institute of Occupational Safety and Health is used to demonstrate feasibility of the proposed methodology and the developed system.

REFERENCES

1. Bouras, C., Poulopoulos, V. and Tsogkas, V., 2008, "PeRSSonal's core functionality evaluation: Enhancing text labeling through

المنارة للاستشارات
www.manaraa.com

personalized summaries," *Data&Knowledge Engineering*, Vol. 64, No. 1, pp. 330-345.

2. 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.
3. Chan, S. W. K., 2006, "Beyond keyword and cue-phrase matching: A sentence-based abstraction technique for information extraction," *Decision Support Systems*, Vol. 42, No. 2, pp. 759-777.
4. Dahab, M. Y., Hassan, H. A. and Rafes, A., 2008, "TextOntoEx: Automatic ontology construction from natural English text," *Expert Systems With Applications*, Vol. 34, No. 2, pp. 1474-1480.
5. Elhadad, N., Kan, M. Y., Klavans, J. L. and McKeown, K. R., 2010, "Customization in a unified framework for summarizing medical literature," *Artificial Intelligence in Medicine*, Vol. 33, No. 2, pp. 179-198.
6. Legara, E. F., Monterola, C. and Abundo, C., 2011, "Ranking of predictor variables based on effect size criterion provides an accurate means of automatically classifying opinion column articles," *Physica A: Statistical Mechanics and its Applications*, Vol. 390, No. 1, pp. 110-119.
7. Li, Q. and Chen, Y. P., 2010, "Personalized text snippet extraction using statistical language models," *Pattern Recognition*, Vol. 43, No. 1, pp. 378-386.
8. Lorch, R. F. Jr., Lorch, E. P., Ritchey, K., McGovern, L. and Coleman, D., 2001, "Effects of Headings on Text Summarization," *Contemporary Educational Psychology*, Vol. 26, No. 2, pp. 171-191.
9. Moens, M. F., Uyttendaele, C. and Dumortier, J., 1999, "Information extraction from legal texts: the potential of discourse analysis," *International Journal of Human-Computer Studies*, Vol. 51, pp. 1155-1171.
10. Nomoto, T. and Matsumoto, Y., 2001, "A New Approach to Unsupervised Text Summarization," *Proceedings of the 24th International Conference on Research in Information Retrieval*, pp. 26-34.
11. Ouyang, Y., Li, W., Li, S. and Lu, Q., 2011, "Applying regression models to query-focused multi-document summarization," *International Journal of Medical Informatics*. Vol. 47, No. 2, pp. 227-237.
12. Pattern, R. L., 2008, "An effective sentence-extraction technique using contextual information and statistical approaches for text summarization," *Pattern Recognition Letters*, Vol. 29, No. 9, pp. 1366-1371.
13. Ruiz-Casado, M., Alfonseca, E. and Castells, P., 2007, "Automatising the learning of lexical patterns: An application to the enrichment of WordNet by extracting semantic relationships from Wikipedia," *Data & Knowledge Engineering*, Vol. 61, No. 3, pp. 484-499.
14. Sweeney, S., Crestani, F. and Losada, D. E., 2008, "'Show me more': Incremental length summarisation using novelty detection," *Information Processing and Management*, Vol. 44, No. 2, pp. 663-686.
15. Tao, Y. H., Liu, S. C. and Lin, C. L., 2011, "Summary of FAQs from a topical forum based on the native composition structure," *Expert Systems With Applications*, Vol. 38, No. 1, pp. 527-535.
16. Uzuner, O., Mailoa, J., Ryan, R. and Sibanda, T., 2010, "Semantic relations for problem-oriented medical records," *Artificial Intelligence in Medicine*, Vol. 50, No. 2, pp. 63-73.
17. Xie, S., Liu, Y., 2010, "Improving supervised learning for meeting summarization using sampling and regression," *Computer Speech & Language*, Vol. 24, No. 3, pp. 495-514.
18. Yang, C. C. and Wang, F. L., 2007, "An information delivery system with automatic summarization for mobile commerce," *Decision Support Systems*, Vol. 43, No. 1, pp. 46-61.
19. Zajic, D., Door, B. J., Lin, J. and Schwartz, R., 2007, "Multi-candidate reduction: Sentence compression as a tool for document summarization tasks," *Information Processing and Management*. Vol. 43, No. 6, pp. 1549-1570.
20. Zhou, L., Tao, Y., Cimino, J. J., Chen, E. S., Liu, H., Lussier, Y. A., Hripcsak, G. and Friedman, C., 2006, "Terminology model discovery using natural language processing and visualization techniques," *Journal of Biomedical Informatics*, Vol. 139, No. 6, pp. 626-636.

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 currently graduate student in the Department of Information Management at Nanhua University. Her research interests are knowledge management and knowledge retrieval.

(Received February 2012, revised August 2012, accepted August 2012)

www.manaraa.com

知識文件結構化摘要模式

楊士霆[*]、龔鈺婷
南華大學資訊管理學系
嘉義縣大林鎮中坑里南華路一段 55 號

摘要

人們已習慣透過網路搜尋方式取得資訊與知識，並延伸出關鍵字搜尋、文件分類等方式相關研究與技術以協助搜尋。雖有資訊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。因此，本研究針對特定領域知識文件發展一套「知識文件結構化摘要」模式，先行對「勞工安全知識網」之人因工程工作場所改善報告進行目標文件解析，擷取此類型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；其次，透過「觀念性語彙模組」係藉由各項語彙法則取得文件最具觀念性或是代表性語句，以協助知識文件關鍵字語意比對亦可作為結構化摘要之候選語句。最後，「結構化摘要模組」乃計算並擷取具有文件之代表性語句，並整合成制定摘要呈現方式，進而形成一套知識文件結構化摘要模式，期望知識搜尋者即可針對問題需求直接閱讀所需部分，以確保能於短時間內決定與篩選該文件是否為知識搜尋者所需資料。

關鍵詞：勞工安全知識網、知識管理、資料探勘、文件摘要技術
（*聯絡人：stingyang@mail.nhu.edu.tw）

www.manaraa.com

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.

www.manaraa.com