A MODEL FOR KNOWLEDGE COMPONENT ANALYSIS

[1]JIANG-LIANG HOU, [2]JIANG-YU CHEN

Department of Industrial Engineering and Engineering Management, National Tsing Hua University
Email: adamhou@ie.nthu.edu.tw, g9534537@oz.nthu.edu.tw

Abstract- To efficiently search for or utilize the domain-specific knowledge from documents has become a key issue in regard to document and information management. If documents can be divided into several knowledge components, then the reader can efficiently search for specific domain knowledge. Critically implicit knowledge from documents is often given in the form of figures and tables, and domain specific knowledge is embedded within the paragraphs surrounding these figures and tables. This paper develops a model for knowledge component extraction in order to extract figure or table-related paragraphs from numerous documents based on knowledge components, and is designed to enable the efficient acquisition of domain-specific knowledge

Index Terms- Knowledge Component Analysis, Knowledge Management.

I. INTRODUCTION

As readers searched for specific domain knowledge through traditional knowledge extraction schemes, they had to digest a great deal of irrelevant content from entire documents in order to obtain the necessary domain-specific knowledge. In traditional knowledge extraction schemes, the figures and tables served as an auxiliary tool which provided the intended contents for the reader. As such, the time required for the readers to search the specific domain knowledge could not be reduced. The issues of this paper can be summarized as follows:

- Traditional knowledge extraction schemes for entire documents cannot assist a knowledge receiver/reader in the efficient extraction of domain specific knowledge within the document.
- Traditional knowledge extraction schemes cannot accurately provide the required information or knowledge for a knowledge receiver/reader.

This paper develops a knowledge component extraction model with which to extract figure or table-related paragraphs from numerous documents based on knowledge components, and is designed to enable the efficient acquisition of domain-specific knowledge. Using the strategy developed by this paper, a document can be divided into several figure or table-based knowledge components. Then, a reader can accurately extract domain-specific knowledge. This proposed model reduces the time and cost of reviewing an entire document, and the implicit knowledge within the document can be searched for and used by a reader with less effort than with previous strategies.

II. LITERATURE REVIEW

Figure or table knowledge component extraction is to systematically analyze document contents with respect to figures or tables in order to extract specific knowledge and concepts. The figures or tables were analyzed to obtain corresponding domain specific knowledge characteristics to enable the extraction of the textual domain knowledge in the figure or table via a textual knowledge analysis that includes sentence segmentation, grammar analysis and knowledge conceptualization. Since figure, image or table analysis technology was used to acquire figure or image characteristics, textual knowledge analysis technology was required for acquiring structured textual knowledge. The concept of figure or image-text knowledge extraction technology allows for the extraction of the figure or image-related text knowledge. Figure or image-based knowledge component extraction includes: textual knowledge analysis and figure or image content analysis and figure or image-text knowledge extraction.

2.1 Textual Knowledge Analysis

Textual knowledge analysis indicated that knowledge contents from a document were analyzed systematically. That is, textual knowledge analysis was used to extract domain specific knowledge and concepts from the documents. According to the structural analysis of domain specific knowledge, textual knowledge analysis can be categorized in natural language processing (NLP) and knowledge conceptualization technologies. Using NLP, technology, the vocabulary, grammar, semantics and context in the sentences can be analyzed in order to derive the segmented sentences [1] and grammar analysis [2] for use as a basis for knowledge conceptualization technology.

The aim of knowledge conceptualization technology is to help knowledge managers extract and classify structured textual knowledge in order to obtain integrity and validity. Textual knowledge conceptualization can be processed through grammar structure analysis and the reasoning rule. Grammar structure analysis was intended for analyzing

grammar structure and for structuring textual knowledge through vocabulary analysis, semantics analysis and grammar analysis [3][4]. The reasoning rule is applied to textual knowledge conceptualization where the knowledge extraction rules can be established manually or automatically [5].

2.2 Figure or image Content Analysis

Figure or image content analysis indicated that the characteristics contained in the figures or images were analyzed systematically. The analytical objects included graphic texts, images and visual features, such as color, texture and shape. With regard to figure or image content analysis, previous research can be divided into text extraction and characteristic analysis. Figure or image text extraction was used to identify and acquire the textual contents from complicated images [6]. Characteristic analysis was applied to analyze graphic patterns and visual features [7].

In addition, figure or image content analysis was used to accurately extract the text contained in the figures or images, identify graphic images and analyze graphic characteristics. Previous studies analyzed content for figure or image retrieval [8] and figure or image classification [9], using figure or image content analysis technology.

However, analysis based on visual features is difficult to use when exploring the figure or image semantics in the context of human interpretability. Some researchers extracted figure or image-text knowledge based on textual semantic information in order to derive a figure or image semantic description similar to that of human interpretation [10]. Other researchers attempted to extract the figure or image-text knowledge by integrating visual features and textual semantic concepts to improve the accuracy of figure or image textual knowledge extraction [10].

III. KNOWLEDGE COMPONENT EXTRACTION MODEL

The figures and tables in a document typically contain important implicit knowledge; the aim of this paper was to extract the knowledge components of figures and tables within the document. To extract these knowledge components, figure or table titles, contents and numbers need to be acquired in advance. Figure or table contents can be divided into textual contents and digital contents. The numbers and figure or table titles from the bottom or top correspond to the purpose or description.

The titles, contents, and numbers can be compared with the contents of target documents in order to obtain the specific paragraphs related to the figure or table knowledge. In addition, the locations of figure or table knowledge paragraphs in the target document can be identified from the description characteristics of the first and last sentences from arranged free-form documents. This paper develops a knowledge component extraction model including a keyword mapping algorithm (KMA) and a sentence mapping algorithm (SMA) used to extract the description paragraphs in figures and tables. The detailed concepts of the algorithms are described below.

3.1 Document Content Analysis

The figure or table titles, figure or table contents, and figure or table numbers description paragraphs of free-form documents are typically used to describe specific figures and tables. To extract the correct locations of the description paragraphs, the keywords related to figures or tables can be extracted for comparison with the document contents. In addition, the target document could be divided into several complete sentences in advance for extraction of description paragraphs related to figures and tables.

- Step (A1): Extraction of figure or table keywords
- Step (A2): Segmentation of sentences

3.2 Keyword Mapping Algorithm

For most domain specific documents, the figure or table titles, figure or table contents, and figure or table numbers are used to describe critical information. To extract the correct parts of description paragraphs related to figures or tables, the normalized keyword frequencies of figure or table titles, figure or table contents, and figure or table numbers from each sentence could be calculated for extraction of the figure or table description sentences. The procedures of the keyword mapping algorithm consists of three major steps.

- Step (B1): Calculation for the normalized frequency of figure or table keywords
- Step (B2): Calculation of the normalized frequency of figure or table keyword set
- Step (B3): Determination of the figure or table description sentences

3.3 Sentence Mapping Algorithm (SMA)

In the first or end description sentences of figure or table in free-form documents, the specified semantic structures are typically used to represent the first or end description sentences of figures or tables. The specified semantic structure is the combination of specified words in the sentence with established usages, figure or table title keywords, figure or table content keywords, and figure or table numbers. In addition, the specified words in the sentence with established usages can be divided into first sentence's first word, first sentence's last word , end sentence's first word, and end sentence's last word. This paper attempts to figure out these characteristics of first and end sentences in figure or table description sentences of free-form documents and then determine the figure or table description sentences accordingly.

- Step (C1): Acquisition of the rank values of words

in first sentences of target documents
- Step (C2): Acquisition of the rank values of words in end sentences of target documents
- Step (C3): Comparison of the representative semantic structure of the first sentence characteristics

CONCLUSION

Overall, owing to the growing complexity of document content as well as a significant increase in domain specific knowledge, it is difficult for knowledge receivers to digest domain specific knowledge. Notably, traditional knowledge extraction schemes typically provide complete documents to readers, a process which requires significant amounts of time to digest this knowledge. The aim of this paper was to develop a knowledge component extraction model in order to improve the efficiency of extracting domain specific knowledge from the figure or table-based documents. The proposed knowledge component extraction technology can support the efficient management and utilization of numerous documents with figures and tables. Readers can efficiently and accurately acquire knowledge implicitly contained in documents.

REFERENCES

[1] Xue, N., 2003, "Chinese word segmentation as character tagging," Computational Linguistics and Chinese Language Processing, Vol. 8, No. 1, pp. 29-48.
[2] Chandrasekar, R. and Srinivas, B., 1997, "Automatic induction of rules for text simplification," Knowledge-Based Systems, Vol. 10, No. 3, pp. 183-190.
[3] Bichindaritz, I. and Akkineni, S., 2006, "Concept mining for indexing medical literature," Engineering Applications of Artificial Intelligence, Vol. 19, No. 4, pp. 411-417.
[4] Kim, J. T. and Moldovan, D. I., 1995, "Acquisition of linguistic patterns for knowledge-based information extraction," IEEE Transactions on Knowledge and Data Engineering, Vol. 7, No. 5, pp. 713-724.
[5] Akgobek, O., Aydin, Y. S., Oztemel, E. and Aksoy, M. S., 2006, "A new algorithm for automatic knowledge acquisition in inductive learning," Knowledge-Based Systems, Vol. 19, No. 6, pp. 388-395.
[6] Chen, D., Odobez, J. and Bourlard, H., 2004, "Text detection and recognition in images and video frames," The Journal of the Pattern Recognition Society, Vol. 37, No. 3, pp. 595-608.
[7] Ruaro, M. E., Bonifazi, P. and Torre, V., 2005, "Toward the neurocomputer: Image processing and pattern recognition with neuronal cultures," IEEE Transactions on Biomedical Engineering, Vol. 52, No. 3, pp. 371-383.
[8] Favela, J. and Meza, V., 1999, "Image-retrieval agent: Integrating image content and text," IEEE Intelligent Systems, Vol. 14, No. 5, pp. 36-39.
[9] Zenou, E. and Samuelides, M., 2005, "Characterizing image sets using formal concept analysis," EURASIP Journal on Applied Signal Processing, Vol. 2005, No. 13, pp. 1931-1938.
[10] Barnard, K., Duygulu, P. and Forsyth, D. A., 2003, "Recognition as translating images into text," Proceedings of the SPIE, Vol. 5018, pp. 168-178.

★★★