A PILOT STUDY ON ENHANCING THE APPLICATION OF KNOWLEDGE MANAGEMENT SYSTEMS USING SEMANTIC SEGMENTATION

Ji-Wei Wu
National Chiao Tung University, Taiwan, R.O.C.
jwwu.cs97g@nctu.edu.tw

Judy C.R. Tseng
Chung Hua University, Taiwan, R.O.C.
judycrt@chu.edu.tw

Wen-der Yu
Chung Hua University, Taiwan, R.O.C.
wenderyu@chu.edu.tw

Jyh-bin Yang
Chung Hua University, Taiwan, R.O.C.
jyhbin@chu.edu.tw

Wen-Nung Tsai
National Chiao Tung University, Taiwan, R.O.C.
tsaiwn@cs.nctu.edu.tw

Shun-Min Lee
CECI Engineering Consultants, Inc., Taiwan, R.O.C.
smlee@ceci.com.tw

Abstract

Owing to extensive knowledge management activities are promoted in various enterprises. How to enhance the application of knowledge management systems (KMS) becomes a critical issue. One of the enhanced applications of KMS is the SOS system, which aims at sharing knowledge for solving emerging problems. While it takes time for experienced employees to share their knowledge via the SOS system, it is advantageous to introduce an Automatic Problem Answering (APA) mechanism. When an emerging problem is issued, the APA mechanism will find actively suitable knowledge from the Intellectual Asset Repository (IAR), which consists of various sources of knowledge. In this paper, a semantic segmentation method is proposed and is employed in building a semantic segmentation module (SSM) in APA. SSM will automatically extract the knowledge corpuses embedded in documents. The extracted knowledge corpuses will be reused in APA to solve emergent problems and thus the application of KMS is enhanced.

KEYWORDS: Knowledge management, Text mining, Semantic segmentation
INTRODUCTION

More and more construction organizations have adopted commercial Knowledge Management Systems (KMS) for developing their own Knowledge Management (KM) functionalities. The existing KMS are mostly developed based on Communities of Practice (COP) for knowledge sharing and exchange (Yu, Yang, Tseng, & Yu, 2007). The SOS system is a specialized COP that provides real-time aids for engineers/managers who are encountered with emergent problems. Once a problem is posed by a questioner in SOS, it will prompt automatically on the entry page of the KMS of every member. The SOS system has been proved to be very beneficial to the firm. Both tangible and intangible benefits were resulted significantly (Yu, Chang, & Liu, 2006).

The essential problem of existing COP in solving emergent problems is that the problem posed in the COP should “wait” the domain experts to provide solutions. Such “passive” mode of problem solving assumes that the domain experts can “see” the problem and respond with their solution timely. However, previous research found that such reactive problem solving (RPS) approach has caused inefficiency of timeliness and cost effectiveness of the KMS (Yu et al., 2007). To tackle with these problems, the Model of Proactive Problem Solving (MPPS) is proposed (Yu et al., 2007). Unlike RPS, the PPS proactively solves the problem through Automatic Problem Answering (APA) module. The Automatic Problem Answering module (APA) is an automatic problem solving system (APS) that searches the solution database called Intellectual Asset Repository (IAR) to provide the most appropriate answer to the problem.

One major limitation of the proposed PPS is the requirement of the historic lesson learned files (LLF). The LLF are compiled manually by the questioner who obtained solution from the domain experts. Compilation of previous experience of problem solving is expensive. Moreover, the project final reports, plans, proposals, and other knowledge documents contain tremendous engineering experiences that are valuable sources for solutions of solving problems. An automatic functionality should be developed to compile the explicit knowledge stored in those the abovementioned documents. In this pilot study, a semantic segmentation method from text mining is proposed and is employed in building a semantic segmentation module (SSM). Final reports are selected from a construction organization to serve as the research subject. The knowledge corpuses embedded in these reports were extracted by SSM automatically. The extracted knowledge corpuses will nourish IAR and be used in APA for solving future emergent problems.

The rest of the paper starts with reviews of related works to provide required backgrounds for SSM; the Model of Proactive Problem Solving (MPPS), which SSM is built on, is then briefly reviewed; then the implementation of SSM is presented to demonstrate the applicability of SSM; finally, conclusions and future works are drawn based on our research findings.

RELATED WORKS

Text mining, like data mining or knowledge discovery (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996), is often regarded as a process to find implicit, previously unknown, and potentially useful patterns from a large text repository (Tseng, Lin, & Lin, 2007). Semantic
segmentation is one of the techniques in text mining. The aim of semantic segmentation is to segment a document into several thematic segments. Most of semantic segmentation researches (Beeferman, Berger, & Lafferty, 1999; Hearst, 1994) are based on the well known vector space model (VSM) (Salton & McGill, 1983; Salton, Wong, & Yang, 1975). Due to its popularity, the VSM based semantic segmentation techniques were employed in this study. VSM and related techniques of semantic segmentation will be described in the following subsections.

**Vector Space Model**

Vector Space Model (VSM) is the most popular model in information retrieval (IR) systems (Wu, Wei, & Tseng, 2007). In VSM, documents are represented as a Characteristic Vector (CV) which consists of n weights. Each weight represents the importance of a unique keyword in the document. The CV of document $D_i$ can be represented with the Equation (1):

$$CV(D_i) = \{(k_1, w_{k1}), (k_2, w_{k2}), \ldots, (k_j, w_{kj}), \ldots, (k_n, w_{kn})\} \quad (1)$$

where $CV(D_i)$ is the characteristic vector of the $i$th document; $K_j$ represents $j$th keyword; and $W_{ij}$ is the weighting value of $K_j$ in document $D_i$.

**TF×IDF**

A common approach used to determine the weight of a unique keyword is the well known TF×IDF (Term Frequency × Inverse Document Frequency) (Salton & McGill, 1983; Salton et al., 1975) method. It is a statistical measure for evaluating the importance of a term to the document it is contained. The importance increases proportionally to the frequency of the term (Term Frequency, TF) in the document, but is offset by the frequency of the term in the whole document collection (Inverse document frequency, IDF). TF×IDF is given as Equation (2):

$$W_{ij} = TF_{ij} \times IDF_j, \quad IDF_j = \log \left( \frac{N}{DF_j} \right) \quad (2)$$

where $TF_{ij}$ is the frequency of the $j$th term in the $i$th document ($D_i$); $DF_j$ is the number of documents in which the $j$th term occurs at least once. The inverse document frequency ($IDF_j$) is calculated from the number of documents ($N$) divided by document frequency ($DF_j$).

**IMpOrtance Factor (IMF)**

While TF×IDF is well known, the length of a term, which implies the importance of a term, is not considered. Based on TF×IDF, Wu et al. propose a novel modified TF×IDF scheme, called Importance Factor (IMF) (Wu et al., 2007). IMF is given as Equation (3):

$$IMF_{ij} = \frac{L_j}{L_{j,\text{max}}} \cdot (0.5 + 0.5 \frac{TF_{ij}}{TF_{i,\text{max}}}) \cdot IDF_j, \quad IDF_j = \log \left( \frac{N}{\sum_{i=1}^{N} C_{ij}} \right) \quad (3)$$

where $IMF_{ij}$ is the weighting of the $j$th term in the $i$th document ($D_i$); $L_j$ is the length of the $j$th term; $L_{j,\text{max}}$ is the maximum length of terms in $D_i$; $TF_{ij}$ is the number of occurrences for the $j$th term in $D_i$; $TF_{i,\text{max}}$ is the maximum number of occurrences for the terms in $D_i$; $N$ is the number of LLFs in the repository; and $C_{ij} = 1$ if $D_i$ contains term $j$, $C_{ij} = 0$, otherwise.
BRIEF REVIEW OF THE MODEL OF PROACTIVE PROBLEM SOLVING (MPPS)

The Model of Proactive Problem Solving (MPPS) is depicted in Figure 1. In the integrated framework, MPPS solves construction problems in two modes: (1) Automatic problem answering mode (APA mode)—the problem solving process is shown in Figure 1 as bold solid arrows, where the solution is searched automatically from historic lessons learned based on problem characteristics; (2) Automatic problem dispatching (APD) mode—the problem solving process is shown in Figure 1 as dashed arrows, where the unsolved problems (by APA mode) is automatically dispatched to the most related domain experts according to the problem characteristics and the Knowledge Capacity Matrix (KCM). The functions of problem solving in the traditional KMS is preserved and exercised in MPPS as shown in Figure 1 where the unsolved problem is posted in the COP of the KMS before entering the APD mode.

Figure 1: The Model of Proactive Problem Solving (MPPS)

Both the problems solved in APA and APD models are new lessons learned for future problems. This process is actually a verification of the knowledge to generate a higher level of intellectual asset called “wisdom”. This process is performed by Lessons Learned Wizard (LLW) as shown in Figure 1.

Versatility of the Intellectual Asset Repository (IAR) is the key to the success of APA. One major limitation of the MPPS is the nourishment of IAR depends only on historic LLFs. Since the manual compilation of previous experience of problem solving into LLF is costly and time consuming, IAR grows slowly. However, existing documents, such as proposals and final reports, of an organization are so versatile that knowledgeable works has been adequately kept in them. If the knowledge corpus can be extracted automatically from
existing documents, IAR will be quickly nourished and the MPPS will become much more useful. In this study, a semantic segmentation method is proposed and is employed in building a semantic segmentation module (SSM) to tackle with this problem.

THE PROPOSED SEMANTIC SEGMENTATION METHOD

Project final reports, plans, proposals, and other knowledge documents are served as the intelligence collection of construction organizations. These knowledge documents contain tremendous engineering experiences that are valuable source for solving problems. Nevertheless, the employees are not always available to find the related documents when they are in trouble. Moreover, it would be too time consuming for employees to filter out the solutions they need, especially in long text documents which contain dozens of pages. To achieve fully automation of knowledge extractions for further enhancing the application of knowledge management systems, the semantic segmentation module (SSM) is developed.

Semantic Segmentation Module (SSM)

The enhancement of MPPS with SSM is shown in Figure 2. First, a knowledge document is loaded from the document repository. The characteristics of the text paragraphs included in the document are analyzed by SSM. Then, the similarities between each two consecutive text paragraphs are measured. The document is segmented into several semantic paragraphs by identifying the semantic segment boundaries according to a threshold of the similarity measurement. Then the semantic paragraphs extracted are saved in IAR. Once a question is posed by the questioner, the characteristics of the posed problem are analyzed by APA. Then, APA searches the IAR to find the most relevant solutions. Finally, the solution is retrieved and returned to questioner. As long as IAR is quickly nourished by the semantic segments (representing knowledge corpuses), it is more likely that the question can be solved.

![Figure 2: the enhancement of MPPS with SSM](image-url)
The Process of SSM

The process of SSM is shown as Figure 3; it can be split into three main stages:

Text paragraph characteristic analysis
In this step, each of the text paragraphs contained in a document is transformed into a characteristic vector (CV) using VSM. At first, domain keywords contained in each paragraph are extracted. Second, the importance weightings of the keywords are calculated to form the CV of the paragraph using the Importance Factor (IMF) method as described in Section 2.3.

Similarity measurement
In this step, similarities between each two consecutive text paragraphs are measured using the inner product similarity measurement described in Equation (3):

\[ S_{ik} = \sum_{j=1}^{n} (W_{ij} \times W_{kj}) \]  

(3)

where \( S_{ik} \) is the similarity between text paragraph \( i \) and \( k \); \( W_{ij} \) is the weighting of the \( j \)th element of the \( CV(P_i) \); \( W_{kj} \) is the weighting of the \( j \)th element of the \( CV(P_k) \). The pairs of text paragraphs with lower similarities are more likely to be segmented into different semantic paragraphs.

Semantic paragraph segmentation
The aim of semantic paragraph segmentation is to segment the document into several thematic segments, by determining boundaries between text paragraphs. In order to determine the boundaries, a threshold is empirically determined in advance. The pairs of text paragraphs will be segmented into different semantic paragraphs if the similarity measured is lower than the threshold. Finally, all the semantic paragraphs are extracted from the document and saved into IAR.

Figure 3: the process of SSM
SYSTEM IMPLEMENTATION

The proposed SSM has been implemented at current stage. Figure 4 shows an example of knowledge corpuses of a final report. In this example, the 2-page text is segmented into four semantic paragraphs by SSM.

Figure 4: An example of segmented knowledge corpuses

Figure 5 shows the interface of the APA module for posing the question. Once a question is posed by the questioner, APA searches the intelligent assets repository (IAR) to find the most relevant intellectual assets (either LLFs or semantic paragraphs). Then, the relevant intellectual assets can be explored by the questioner.

Question description

Question classification

Figure 5: APA module—questioning
Figure 6 shows the interface of presenting a semantic paragraph to the questioner. The Questioner can not only read the relevant semantic paragraph but also explore the previous/next semantic segments or even the whole document. Figure 7 shows the interface that used for further exploring the whole document.
CONCLUSIONS

A semantic segmentation module (SSM) is proposed and developed in this paper. SSM automatically extract the knowledge corpuses embedded in reports and documents which are collected over years in an enterprise. The SSM not only improve the previously developed PPS system, but also begin to intellectualize the existing knowledge assets. The extracted knowledge corpuses will be reused in APA to solve emergent problems and thus the application of KMS is enhanced.

Even though preliminary results show promising benefits of the proposed SSM, future verification are required to validate the implementation of SSM, including: (1) quantitative evaluation of time and cost effectiveness; (2) verification of correctness of semantic segments extracted by SSM. Moreover, the proposed SSM can only extract text type knowledge assets. In the future, figure extraction, table extraction and semantic relation construction techniques will be added to MPPS to deal with more kinds of knowledge assets.

REFERENCES


