Institute of Internet and Intelligent Technologies
Vilnius Gediminas Technical University
Saulėtekio al. 11, 10223 Vilnius, Lithuania
http://www.isarc2008.vgtu.lt/

The 25th International Symposium
on Automation and Robotics in Construction

June 26–29, 2008

ISARC-2008

PRELIMINARY STUDY ON THE AUTOMATIC LESSONS-LEARNED FILE GENERATOR

Wen-Der Yu
Professor, Inst. of Const. Mgmt.,
CHU, Taiwan
wenderyu@chu.edu.tw

Yu-Teh Wei
Master Student,
Inst. of Cont.. Mgmt.,
CHU, Taiwan
m09516001@chu.edu.tw

Shen-Jung Liu
Assistant Vice President,
CECI Consult. Inc.,
Taiwan
sjliu@ceci.com.tw

Pei-Lun Chang
Engineer, Depart. of Business & Research,
CECI Consult. Inc.,
Taiwan
peilun@ceci.com.tw

ABSTRACT

Lessons-learned file (LLF) is commonly adopted to retain previous knowledge and experiences for future use in many construction organizations. Current practice in capturing LLF is mainly through the costly and time-consuming manual processes conducted by the construction engineers or managers. Moreover, many construction knowledge accumulated from previous projects is berried in the construction documents such as construction journals, proposals, as-built drawings, SPECs, plans, etc. It is impossible to develop LLFs from these documents manually. This paper presents the work of a preliminary attempt to develop an Automatic Lessons-Learned File Generator (ALLFG) based on text mining techniques. A prototype system is programmed. Case study is conducted to extract meaningful LLF from sample Chinese construction document automatically. Although the results are still experimental, promising potentials can be envisioned for practical applications.

KEYWORDS

Text mining, lessons-learned, corpus, knowledge management

1. INTRODUCTION

Construction engineering is an experience-based discipline. Experiences and knowledge accumulated from previous works play very important role in performing future projects. Unfortunately, the accumulated engineering knowledge and experiences usually disappear as the experienced staffs leave the organization. Many firms have established their knowledge management (KM) endeavours, e.g., Knowledge Management Systems (KMSs), to collect and store the knowledge generated through their daily business operations. Most of the organizations still choose to develop lessons-learned files (LLFs) for formal acquisition of previous construction knowledge and experiences. Most of the tasks are carried out manually by experienced staffs. There have been some efforts to build computer aided lessons-learned systems reported in literature. Some of them are reviewed in the following.

The Hypermedia Constructability System (HCS), Indiana Department of Transportation (INDOT), USA, was developed in collaboration between

INDOT and Purdue University [1]. The HCS stores historic lessons-learned in multi-media format so that construction engineers can learn from previous lessons more effectively. The Constructability Lessons Learned Database (CLLD) & Integrated Knowledge-Intensive System (IKIS) were developed by Kartam and Flood [2][3] to provide a repository for previously learned lessons. The major difference among CLLD, IKIS, and the other previous lessons-learned systems is that IKIS establish a formal process to verify historic lessons-learned by the domain experts before storing in the database. The Construction Industry Institute (CII) developed a Lessons-Learned Wizard (LLW) with the package of constructability program [4]. LLW provides a systematic approach to acquire and store historic lessons-learned. However, human involvement is still inevitable.

In spite of the different efforts mentioned above, none of them replaces the human's role in the lessons-learned process (LLP) addressed by Fisher et al. [5] as depicted in Fig. 1.

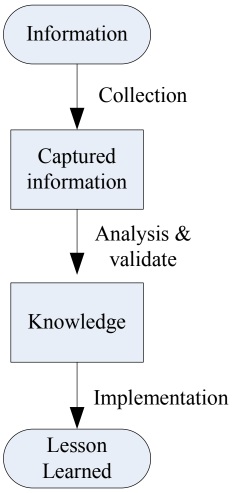


Figure 1. Concept of lesson-learned process [5]

Fisher et al. addressed that the core activity of LLP should be performed by the human knowledge holder in order to capture the information, analyze the validity of the information, and transform into a form of "conceived knowledge (or 'believed knowledge')" to be implemented as a lesson-learned. Such perspective is also supported by Nonaka's four-dimensional model for organizational knowledge creation (also known as "spiral of organizational knowledge creation") [6]. The requirement of human's involvement can result in bottleneck in generating valuable LLFs. Since the construction engineers and managers are already extremely busy in their daily works, it is almost impossible for them to spend extra time reviewing the construction documents and filling up the pre-designed form of LLF. Moreover, many historic lessons are berried in existing construction documents such as construction journals, proposals, as-built drawings, SPECs, plans, etc. It is much desired to develop an Automatic Lessons-Learned File Generator (ALLFG) that can automatically generate lessons-learned form existing construction documents.

The presented work is a preliminary attempt to develop an Automatic Lessons-Learned File Generator (ALLFG) for LLF generation from construction documents in Chinese language based on text mining techniques. A case study of the ALLFG is conducted to extract LLF from the document of a local A/E firm to evaluate the feasibility of the proposed method. The rest of this paper is presented in the following manner: the Chinese text mining techniques related to the proposed ALLFG are reviewed in the second section; the framework and theoretical backgrounds of ALLFG are described in the third section; a case study to test the preliminary ALLFG is demonstrated in Section Four to evaluate the feasibility of such method; discussions on the preliminary testing are provided in Section Five; finally, the findings and future works are concluded.

2. REVIEW OF TEXT MINING TECHNIQUES FOR CHINESE DOCUMENTS

2.1. Text Mining

Text mining (TM), also known as Knowledge Discovery from Text (KDT) or Document Information Mining, is a process to discover the implicit and useful information and knowledge stored in the documents [7][8]. The KDT process usually employs techniques such as Information Retrieval (IR), Information Extraction (IE), Computational Linguistics, Natural Language

Processing (NLP), DM, and knowledge representation. Each of the above techniques has formed a specific and quite matured domain of research. The main difference between traditional DM and KDT is that the former focuses on the structural data in the databases; while the latter tackles semi- or non-structural texts. Dörre [9] addressed two difficulties in KDT: (1) the manual approach for characteristic analysis of mass documents is inefficient; (2) the key attributes are uneasy to define as the dimensionality of text data is large. Therefore, KDT requires additional data selection process compared with the traditional DM.

2.2. Vector Space Model

A special technique of IE named Vector Space Model (VSM) is adopted to perform IE functions in abstracting lessons-learned files from Chinese construction documents. VSM is an algebraic model for representing text documents as vectors of identifiers. The original VSM was first proposed by Salton et al. and described as follows [10]:

For a document d_i and all words $w_l \in W$, frequencies $f(w_l \mid d_i)$ or probabilities are considered in following Eq. (1).

$$p(w_l \mid d_i) = \frac{f(w_l \mid d_i)}{\sum_m f(w_m \mid d_i)} \tag{1}$$

Using a vector space of $L = \|W\|$ dimensions, a document d_i is given as a vector $\vec{x}_i$ of word probabilities as the following Eq. (2):

$$\vec{x}_i = \left[p(w_1 \mid d_i), \ldots, p(w_L \mid d_i)\right]^T \tag{2}$$

Due to the large amount of distinct Chinese words in any non-trivial corpus ($L \approx 10^5 \sim 10^7$) the vector space is extremely high dimensional but sparsely occupied [11].

2.3. Corpus-Based VSM

Using Chinese word stems to represent documents usually results in the inappropriate fragmentation of multi-word concepts [12][11]. As a result, using pre-stored phrases instead of single words or word stems as the terms may produce a VSM that better represents the human recalling process, and result in more effective document retrieval. Kupiec et al. [13] proposed a Corpus-based approach based on Bayesian classifiers to enhance the document retrieval in VSM can be used for Chinese documents, too. The Corpus-based VSM is described as follows:

Assume that sentence *s* is any sentence considered to test document *S* and $F_1 \sim F_k$ are characteristics using to measure the importance of a sentence (e.g., the frequency of a set of keywords), then the probability of *s* belonging to the *summary* can be calculated by Eq. (3).

$$\mathrm{P}\left(s \in S \middle| F_1, F_2, \ldots, F_k\right) = \frac{\prod_{j=1}^{k} P\left(F_j \middle| s \in S\right) P\left(s \in S\right)}{\prod_{j=1}^{k} P\left(F_j\right)}, \tag{3}$$

In practical implementation, the following Eq. (4) is adopted to replace Eq. (3):

$$\mathrm{P}(s \in S) = \frac{\#(sentence \quad in \quad summary)}{\#(sentence \quad in \quad training \quad corpus)} \tag{4}$$

In Eq. (4), numbers of word frequency in the summary and that in the training corpus are calculated, respectively to determine the probability.

3. PROPOSED ALLFG

3.1. System Framework

The system framework of the proposed ALLFG is shown in Fig. 2.

The proposed ALLFG consists of six components: (1) the CKIP Phrase Database—a corpus base storing five million most commonly used Chinese words and phrases provided by the Institute of Information Science, Academia Sinica of Taiwan [14]; (2) Key Phrase Extraction Module—extracting key phrases from the CKIP Phrase Database; (3) Characteristic Extraction Module—Generating characteristic vectors from Key Phrase Database; (4) Key Phrase Database—storing the extracted key phrases by Key Phrase Extraction Module in the order of relative importance; (5) Segmentation Module—breaking down the target document into limited meaningful text segments; (6) Segment Rating Module—evaluating the text segments with respect to their characteristic values to determine the relative similarity; (7) Problem Domain Characteristics Database—storing the characteristic values of text segments in the document .

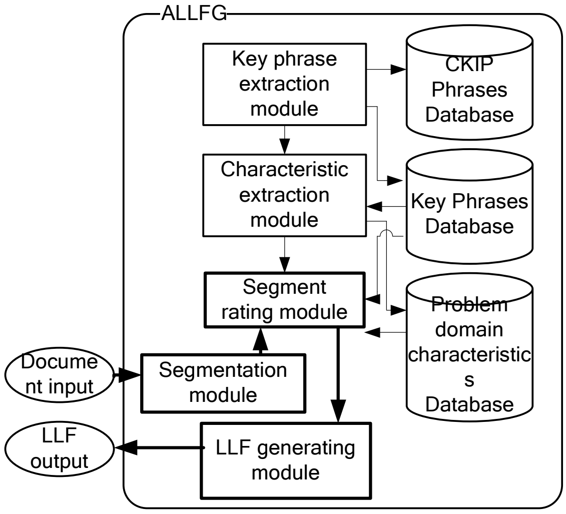


Figure 2. System framework of ALLFG

3.2. Computational algorithms

The computational algorithms of the proposed ALLFG for the KM documents of the CoP in a KMS are depicted in Fig. 3. The KM documents record the questions and associated responses from participants of CoP. Such Q&A type documents are accumulated in the KMS continuously. Before transforming into LLFs, the documents are difficult to reuse.

The computational algorithms of ALLFG consist of the three sub processes:

Questioning sub process:

A question is posed by the questioner in the CoP.

The question is separated into sentences according to interpunctions, e.g., “!, @, \t, $, %, ^, &, *, (,), \n, \r, -, _, +, =, {, }, [,], :, ;, <, >, ,, ., ?, /, ~, `, ，, 。, 「, 」”.

The sentences are segmented into keyword/keywords—the maximum matching algorithm (MM) is adopted to pick the longest possible phrases.

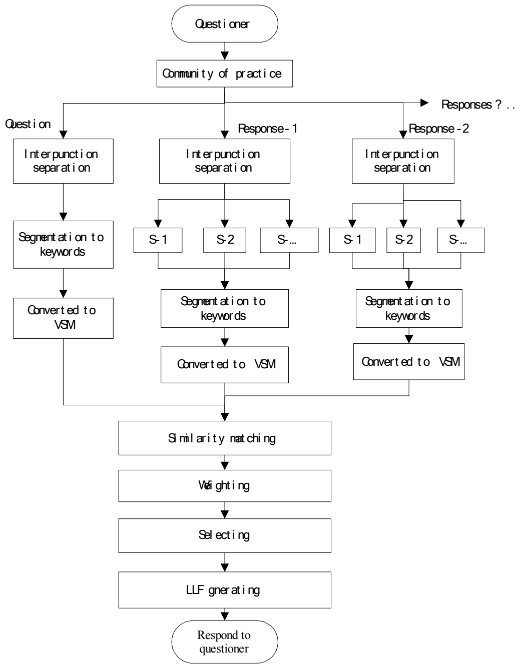


Figure 3. Computational algorithms of ALLFG

The question is converted into question VSM (Q-VSM) based on the importance factors (IMFs) calculated by Eq. (5) and represented as shown in Fig. 4, where $\overline{\mathrm{q}}$ is the VSM of question, $\overline{\mathrm{d}}_{\mathrm{j}}$ is the VSM of the i^{th} response, and w_i is the attributes adopted to represent a document ($\overline{\mathrm{q}}$ or $\overline{\mathrm{d}}_{\mathrm{j}}$).

$$IMF_{i,j} = \frac{L_j}{L_{i,\max}}(0.5 + 0.5\frac{tf_{i,j}}{tf_{i,\max}}) \times \log(\frac{N}{df_j}), \tag{5}$$

where, L_j is the length of j^{th} keywords; $L_{i,\max}$ is the longest keywords of the i^{th} document; $tf_{i,j}$ is the frequency of jth keywords in the ith document; $tf_{i,\max}$ is the keyword of the highest frequency in the i^{th} document; df_j is the number of documents with j^{th} keywords; and N is the number of total documents.

Responding sub process:

A response is provided by the responder in the CoP.

The response is separated into sentences according to interpunctions.

The sentences are segmented into keyword/ keywords.

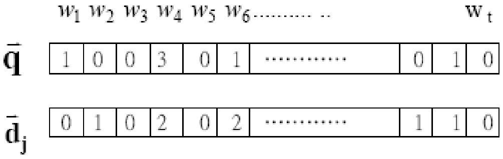
The response is converted into response VSM (R-VSM).

LLF generating sub process:

Similarity matching is performed between the Q-VSM ($\bar{q}$) and R-VSMs ($\bar{d}_j$) by Eq. (5) using inner product:

$$sim(\vec{d}_j,\vec{q}) = \frac{\vec{d}_j \bullet \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}, \tag{5}$$

where, $\bar{q}$ and $\bar{d}_j$ VSMs of the question and responses, respectively, and w_{ij} represents the elements Fig. 4.



	w_1	w_2	w_3	w_4	w_5	w_6………. ..		w_t	
$\bar{q}$	1	0	0	3	0	1 …………	0	1	0
$\bar{d}_j$	0	1	0	2	0	2 …………	1	1	0

Figure 4. VSM representation

Weighting the sentences with Eq. (6);

$$w_{i,j} = TF(i,j) \times IDF(j), \tag{6}$$

where *TF*(*i,j*) is the term frequency (characteristic) of w_j in the i^{th} document, and *IDF*(*j*) is a measure of Inter-document characterization. *TF*(*i,j*) and *IDF*(*j*) defined in Eq. (7) and (8) respectively.

$$\mathrm{TF}(i,j) = \frac{w_{i,j}}{\sum^{t} w_{i,k}} \tag{7}$$

$$IDF(j) = \log\frac{N}{n_j}, \tag{8}$$

where n_j is the number of documents related to characteristic w_j.

Selecting highly weighted sentences according to predefined criteria.

Generating LLF with highest weighted sentences.

3.3. Prototype ALLFG System

The prototype ALLFG system has been programmed with Microsoft VB.net language on a PC platform. All three corpus bases are developed with Microsoft SQL Server 2005 and integrated with ALLFG. In addition to the CKIP phrases, more than 10,000 domain specific phrases are added to the Key Phrases Database. The prototype ALLFG system is tested with pre-evaluated sample KM documents of the CoP provided by the industrial partner to assure the performance.

4. DEMONSTRATED CASE STUDY

In order to test the feasibility of the proposed method, a KM document (the record of knowledge management activities in CoP) is selected for case study. The KM document is in form of Q & A containing a question posed by a staff of the firm (namely the Questioner) and a set of responses provided by other staffs with relevant knowledge or experiences (namely the Responder).

4.1. Selected Documents

The selected KM document is shown below (in Chinese):

Table 1. Selected KM document (partial)

Question	請教各位先進，停工期間為配合業主相關單位需求而辦理趕工，是否有違約情形?
Response 1	個人認為除非配合趕工部分之工作為違法的問題，否則應算為部分停工，趕工部分應照算工期。
…	…
Response *n*	(Description of response *n*)

4.2. Converted VSM

Before converting the document in to VSM, the document should be segmented first. Take the question document as an example, the segmentation result of the sentences for the question is shown in table 2. The converted Q-VSM is as follows:

$\bar{q}$ ={(W1,0.7),(W2,0.22),(W3,1.0),(W4,0.7),(W5,1.0),(W6,0.7),(W7,0.7),(W8,1.0),(W9,1.0),(W10,1.0),(W11,1.0),(W12,0.4),(W13,1.0),(W14,0.52) }

Table 2. Example of segmentation (the question)

Sentence	Description	Segmentation
S_1	請教各位先進	請教 先進
S_2	停工期間為配合業主相關單位需求而辦理趕工	停工 期間 配合 業主 相關 單位 需求 辦理 趕工
S_3	是否有違約情形	是否 違約 情形

The response documents are converted into R-VSMs through the similar process.

4.3. Similarity Matching

The similarity matching is performed, and the similarity index is calculated according to Eq. (5). The results are shown in Table 3.

Table 3. Similarity matching

Response	Segment	Similarity	Order
1	1	0.98	6
2	1	1.14	4
3	1	0.57	10
	2	1.05	5
4	1	0.7	8
	2	0.57	10
5	1	1.39	1
	2	0.4	11
	3	0	12
	4	0.65	9
6	1	0.81	7
	2	1.26	2
7	1	1.23	3

4.4. LLF Generating

Base on the similarity indexes, the LLF is generated by selecting the highest three weighted segments from the responses. The generated LLF is shown in the below.

4.5. Qualitative Evaluation

The generated LLF is presented to the domain experts for qualitative evaluation in Table 4. A preliminary evaluation result shows a roughly 7 out of 10-point-scale can be achieved. The result is promising although there is still room for improvement.

Table 4. Generated LLF

Question		請教各位先進，停工期間為配合業主相關單位需求而辦理趕工，是否有違約情形?
Solution	5(1)*	停工已經業主同意，如需配合相關單位需求辦理分項計畫趕工，且是業主同意。
	6(2)	停工期間若業主不察而配合辦理趕工或擅自復工，將造成違法而被移送法辦。
	7(1)	若屬全面停工，即無所謂趕工，此期間之營造綜合保險一般為中斷狀態，於趕工前應先辦理復工，否則應屬違法及違約情形。

*: Response (segment)

5. DISCUSSIONS

5.1. Major improvements

It is found from the case study that the proposed text mining based ALLFG can automatically generate the Chinese LLF for a specific problem encountered before based on the provided documents. The automation of LLF generation is feasible. Such improvement provides profound potentials to the construction organization in converting their digitalized construction documents into intellectualized assets (e.g., lessons-learned files). Such improvement has not yet found in the literature. It shows a great improvement for construction knowledge preservation.

5.2. Remained problems

Although the case study shows feasibility and potentials of the proposed ALLFG for automatic LLF generation, there are still problems retained to be solved: (1) the accuracy of segmentation is still corpus-sensitive, construction specific corpuses are needed to be established; (2) computation speed for text mining is slow, more efficient algorithms are expected; (3) the generated LLF is sentence (or segment)-based rather paragraph-based, natural language solution is desired.

6. CONCLUSION AND FUTURE WORK

6.1. Conclusion

In this paper, an Automatic Lessons-Learned File Generator (ALLFG) is proposed to automatically generate the LLF from historic documents. The prototype system along with the computational algorithms is proposed. An application of the proposed ALLFG is demonstrated for the KM document of the CoP in an A/E firm. The case study result shows that the proposed method has great potential to convert digitalized construction documents into intellectualized assets (e.g., LLFs). Problem remained to solve are also addressed. It is concluded that the proposed ALLFG is very feasible for automatic LLF generation.

6.2. Future Work

Future research can be pursued on the following directions: (1) establishing construction specific corpuses to improve the segmentation accuracy; (2) developing new algorithms to improve the computation speed; (3) incorporating natural language schemes to improve the readability of the generated LLFs.

7. ACKNOWLEDGEMENT

The founding of this research project was partially supported by the National Science Council, Taiwan, under project No. 96-2221-E-216-054. Sincere appreciations are given to the sponsor by the authors. The valuable case study information presented in this paper was provided by CECI, Taipei. The authors would like to express sincere appreciations to the contributor, too.

REFERENCES

[1] McCullouch, B., and Patty, B., (1994) Hypermedia constructability system, Proceedings Computing in Civil Engineering (New York), ASCE, New York, NY, USA., n 2, pp. 1397-1404.

[2] Kartam, N., and Flood, I. (1997) Constructability feedback systems: Issues and illustrative prototype, Journal of Performance of Constructed Facilities, ASCE, Vol. 11, No. 4, pp. 178-183.

[3] Kartam, N. (1994) Knowledge-intensive database system for making effective use of construction lesson learned, Proceedings Computing in Civil Engineering (New York), ASCE, New York, NY, USA., n 2, pp. 1139-1145.

[4] Construction Industry Institute (1993) Constructability implementation guide, Publication 34-1, Austin, TX, U.S.A.

[5] Fisher, D., Deshpande, S., and Livingston, J. (1998) Modelling the Lessons Learned Process, *Research Report 123-11*, Construction Industry Institute, Austin, TX, USA.

[6] Nonaka, I. (1994) A dynamic theory of organizational knowledge creation, *Organization Science*, 5(1), 14-37.

[7] Feldman, R. and Dagan, I. (1995) Knowledge Discovery in Textual Databases (KDT), *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD 1995)*, 112-117, 1995.

[8] Sullivan, D. (2001) *Document Warehousing and Text Mining*, Wiley Computer Publishing, Yew York, USA.

[9] Dörre, J., Gerstl, P., and Seiffert, R. (1999) Text Mining: Finding Nuggets in Mountains of Textual Data, *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 1999)*, 1999.

[10] Salton. G., Wang, A., and Yang, C. S. (1975) A Vector Space Model for Automatic Indexing, *Communications of the ACM*, Vol. 11, 613-620.

[11] Pullwitt, D. (2002) Integrating contextual information to enhance SOM-based text document clustering, *Neural Networks*, 15, 1099-1106.

[12] Mao, W. and Chu, W. W. (2006) The phrase-based vector space model for automatic retrieval of free-text medical documents, *Data & Knowledge Engineering*, 61, 76-92.

[13] Kupiec, J., Pedersen, J., and Chen, F. (1995) A Trainable Document Summarizer", *Proceedings of the 18th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 68-73.

[14] Institute of Information Science Web Site: http://www.iis.sinica.edu.tw/, Academia Sinica, visited 2008/3.