Knowledge Profiling for Social Networking News Releases

Lo Tse-Yi[a], Vincent Khoo Kay Teong[b]

[a,b]School of Computer Sciences,
Universiti Sains Malaysia, 11800 Minden, Pulau Pinang, Malaysia
Tel : 604-6533888, Fax : 604-6573335

[a]E-mail : tseyi.lo@live.com.my

[b]E-mail : vkhoo@cs.usm.my

ABSTRACT

This paper presents a case study of the news item categorization for a social networking mashup. Since it is time consuming to develop the hierarchical news categories on an iterative and incremental basis using a conventional knowledge mapping approach, an alternative approach is derived by using the Subject Reference System (SRS) guidelines as the reference for categorizing various types of news releases. With the ulterior aim of enabling automatic generation of knowledge maps, this research project adopts only a manual process for the creation of knowledge profile entries, which would then form the input references for future automatic generation of knowledge maps. The approach for transforming a news item to a knowledge profile entry involves some computer-assisted extraction of keywords, which is then linked to the Subject Reference System (SRS) guidelines repository for ease of knowledge profiling. It is argued that the alternative knowledge profiling approach is easier and more efficient than a manufacturing-oriented conventional approach, because the knowledge ontology and validation of news releases have already been proven and accepted by a large community of news agencies.

Keywords

Knowledge Mapping, Knowledge Profiling, News Metadata, Knowledge Ontology, Automatic Knowledge Map Generation

1.0 INTRODUCTION

Information presented in a non-meaningful context for an intended audience is one of the contributing factors for information overload. It has been shown that Office-based and PDF documents are the dominant file formats on the Intranet (Littlefield, 2002). With the revolution of the Web 2.0 technology, it is foreseeable that documents circulated through the Intranet would become a major part of the file distributions throughout the Internet. Similarly, large amount of rich text documents available as news releases posted online would become one of the contributing factors for information overload. Information can only be transformed to knowledge when a proper context is specified. Thus, 'what constitutes knowledge' is constantly pondered upon by computer scientists and experts (Alavi and Leidner, 1999).

Information can be ambiguous without a proper context. A typical knowledge management challenge faced by organizations is not having a standardized approach for sharing and leveraging knowledge internally and externally (Liebowitz, 2004). Since the same set of information can convey different meanings in different context and presentation, one of the challenges is finding the most suitable domain for structuring them (Lê and Lamontagne, 2002). In this regard, knowledge mapping has been described as the process, methods and tools for analyzing knowledge domains in order to discover their inherent features and visualize them in a comprehensive and transparent form (Speel et al., 1999).

Knowledge mapping is believed to be one possible way for resolving the problem of information overload. However, the requirement to redraw knowledge map is becoming increasingly demanding. This paper describes the initial attempt of a series of projects aimed at automatic generation of knowledge maps. In order to automate the process of knowledge mapping, some kind of a knowledge profile must first be established. In the next section, it reviews the related work on a six-step knowledge mapping approach used in a manufacturing industry. Section 3 explains an alternative five-step knowledge mapping approach for profiling social networking news releases. Section 4 briefly compares the two approaches, while section 5 concludes the paper by summarising the work done and proposing several tasks and milestones for future work.

2.0 RELATED WORK ON KNOWLEDGE MAPPING

2.1 Text Extraction and Knowledge Segmentation

Textual information can be obtained through text summarization. Summarization is carried out through techniques of text extraction and text abstraction. Text

extraction is about taking pieces of an original text based on either statistical or heuristic basis, and later joined together in a new shorter text. Text abstraction requires natural language processing of the domain knowledge. For the extraction techniques, they can be done at three different surface-meaning levels: word (Ercan and Cicekli, 2007), phrase (Turney, 2000) and sentence (Chan, 2006, Wesley and Jihoon, 2000). The importance of a sentence is indicated by its location (Baxendale, 1958) as the topic sentences tend to occur at the beginning or the end of documents or paragraphs (Edmundson, 1969).

On the other hand, metadata of document is useful in terms of documented information processing. Metadata extraction, as part of an information extraction activity, can be carried out based on formatted features (Giuffrida et al., 2000, Hu et al., 2006) and linguistic feature (Han et al., 2003). The findings of (Hu et al., 2006) show that the main title of a Microsoft-Word-based document can be successfully extracted using the formatted features.

2.2 Knowledge Mapping Approach

Kim et al. proposes a practical approach to capture, represent and visualize organizational knowledge of a manufacturing company (Kim et al., 2003). It emphasizes on the re-iterative building and validation of knowledge ontology manually, which is a tedious and time-consuming process. In this approach, the domain experts have to create and maintain the knowledge ontology. Knowledge map is the main output, and knowledge profiling is one of the steps in this approach for organizing the organizational knowledge.

The six steps are as follows:

- Defining Organizational Knowledge
- Process Map Analysis
- Knowledge Extraction
- Knowledge Profiling
- Knowledge Linking
- Knowledge Map Validation

Figure 1 shows the six-step approach, while Figure 2 shows a conceptual knowledge map.

Figure 1: Procedure for building knowledge map (Kim et al., 2003)

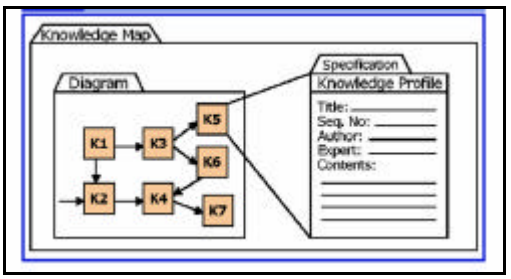

Figure 2: A conceptual model of knowledge map (Kim et al., 2003)

2.3 Subject Reference System (SRS) Guidelines

An example of context selection guidelines in the journalism domain is News Metadata Standards, such as Subject Reference System (SRS) or Publishing Requirements for Industry Standard Metadata (PRISM) (Fernández-García and Sánchez-Fernández, 2004). It is maintained by International Press Telecommunications Council (IPTC) as an international consortium of news agencies, editors and newspapers distributors. Its activities are mainly based on developing and publishing industry standards for the interchange of news data. SRS guidelines offer a hierarchy of subjects as metadata for categorising news. The components of SRS are objects, attributes, subject reference, synonyms and qualifiers. Three news representation standards exist in SRS namely, News Industry Text Format (NITF), News Markup Language (NewsML) and Information Interchange Model (IIM). These specifications are used to represent the knowledge of news items. NITF is a structural framework for news representation through standard XML text format (IPTC and NAA, 2003). It provides an universal language independent coding system for indicating the subject content in describing many features of news items.

3.0 ALTERNATIVE APPROACH FOR KNOWLEDGE PROFILING

In this paper, an attempt is made to describe an alternative knowledge profiling approach. This approach is easier and more efficient that the conventional approach involving incremental development of the knowledge ontology. It is assumed that if the steps for knowledge profiling are well conducted, the knowledge mapping task would not be posing any major problem. In this approach, the NITF specifications of SRS guidelines are adopted.

Based on the six-step knowledge mapping approach proposed by Kim's team (Kim et al., 2003), one obvious basic improvement is getting rid of the manual knowledge ontology building and validation process. The alternative approach consists of five steps as follows:

- Information Acquisition and Collection
- Identification of Information Context
- Knowledge Extraction and Organization
- Knowledge Profiling and Linking
- Knowledge Mapping

Figure 3 shows the alternative five-step knowledge mapping approach.

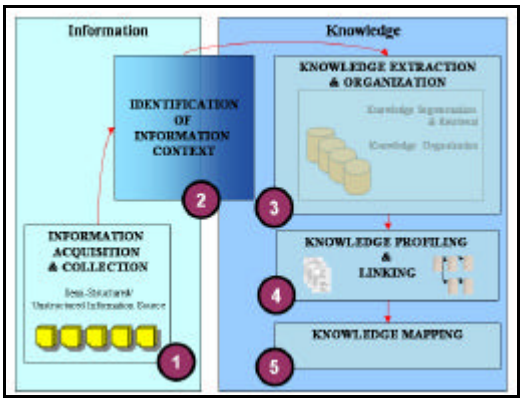

Figure 3: Alternative approach for knowledge profiling

3.1 Information Acquisition and Collection

Figure 4 shows one of the typical knowledge management frameworks (O'Dell and Grayson, 1997). Collection has been identified as one of the processes. In this scenario, the definition of information is one unit of document or chunks of unstructured data in rich-text files. In this research, news releases are selectively downloaded and converted from PDF to Word format or some Office-compatible formats.

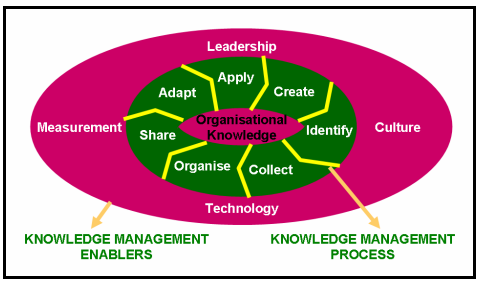

Figure 4: Knowledge management framework (O'Dell and Grayson, 1997)

3.2 Identification of Information Context

The Subject Reference component architecture within SRS identifies the general content of a News Object. It consists of three levels of hierarchy, which are Subject, Subject Matter and Subject Detail. Its reference mechanisms consist of a fixed eight-digit number. The first two digits indicate the top-level Subject. The next three digits show the Subject Matter, and the last three digits identify Subject Detail, as shown in Table 1.

Table 1: Example of SRS reference mechanism

Reference Number	Subject	Subject Matter	Subject Detail
05000000	05	000	000
	Education (EDU)	(none)	(none)

3.3 Knowledge Extraction and Organization

3.3.1 Knowledge Segmentation and Retrieval

In order to reduce the feature dimensionalities, the format-features extraction approach is adopted. The file title is first to be extracted from the first page of a document. The next task is to obtain the summary of the document, by segmenting the document into units of paragraph. The first sentence of each paragraph is then extracted to form part of the news release summary. Figure 5 shows two units obtained from a sample document.

Before Segmentation (Original Paragraph):
Nanyang Technological University (NTU) is a research-intensive university with globally acknowledged strengths in science and engineering. The university is located in a gardencampus in western Singapore, tracing its roots back to 1955.

After Segmentation:
Unit 1: [text="Nanyang Technological University (NTU) is a research-intensive university with globally acknowledged strengths in science and engineering."]
Unit 2: [text="The university is located in a gardencampus in western Singapore, tracing its roots back to 1955."]

Figure 5: Example for paragraph segmentation

3.3.2 Knowledge Organization

As shown in Table 2, four attributes values are stored. The file names is the last part of the file location path in the file system, the file location is the full file location

path in the file system, while the file title is the title of the document.

3.4 Knowledge Profiling and Linking

The primary role of a knowledge profile is to function as a template for machine-processing. Each knowledge profile entry is automatically given a unique identifier. The 'description' field contains the keywords of the summary. The linking of the knowledge profile entries depends on the taxonomy of the SRS guidelines. By doing so, the three sub-fields of 'relation' field - 'subject', 'subject matter' and 'subject detail' - of each knowledge profile entry could be populated based on the field value of 'description'.

3.4.1 Keyword Extraction from Summary

The keywords in the 'description' field are captured based on term weights. The field value should have a maximum of three words that have the highest term frequency with a minimum of two frequencies in the 'summary' field.

3.4.2 Knowledge Profile Creation and Knowledge Ontology Linkage

Based on the information captured in the 'description' field, the three sub-fields of 'relation' field -'subject', 'subject matter' and 'subject detail' - would be provided by the SRS, as shown in Figure 6. In the process of categorizing the news releases, knowledge profile entries are to be linked in according to the three levels of hierarchy of SRS. This process is shown in Figure 6. The knowledge profile entries are to be maintained based on the component architecture as mentioned in Section 3.2.

Table 2: Data definition of a knowledge profile entry

File Name	file01
File Location	.\University01\file01.pdf
File Title	NTU's CNI students to learn from top executives of Banyan Tree
Summary	Nanyang Technological University (NTU)'s Cornell-Nanyang Institute of Hospitality Management (CNI) has launched its inaugural Master Class in collaboration with Banyan Tree Hotels and Resorts. Nanyang Technological University (NTU) is a research-intensive university with globally acknowledged strengths in science and engineering.

Description			
Relation	Subject		
	Subject Matter		
	Subject Detail		

4.0 COMPARISON OF KNOWLEDGE MAPPING APPROACHES

The advantages and disadvantages between the two approaches are compared and shown in Table 3. Since the primary objective of this work is to research into possible software tools for automating the generation of knowledge profile entries, it is acceptable to sacrifice the possible slight inaccuracy in the SRS specifications. Table 4 shows the features comparison of both approaches. It is felt that the alternative approach is more user-friendly and efficient for profiling news releases.

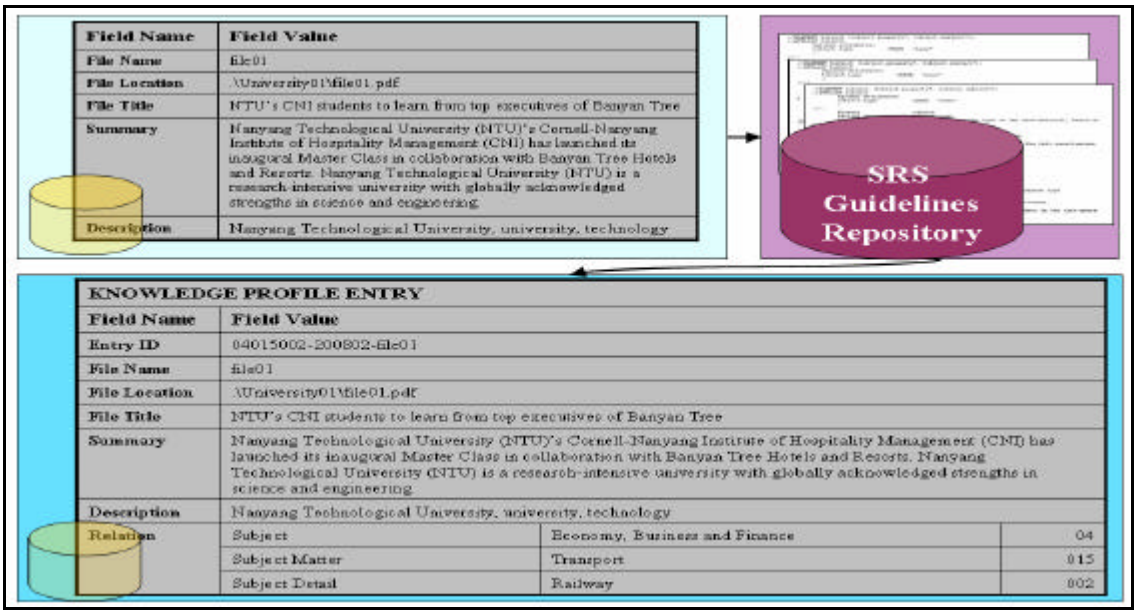

Figure 6: Linking to knowledge ontology

Table 3: Merits comparison of the two approaches

Attributes	Kim et al. Approach (Kim et al., 2003)	Alternative Approach using SRS
Advantages	It is required to implement knowledge map validation checking, which counter-check the accuracy of the knowledge profiling and linking.	It is not required to develop knowledge ontology. SRS is used.
Disadvantages	It is time-consuming to build and validate an organization's knowledge ontology manually, through the techniques of questionnaires and interview in deciding the scope and level of details of the organizational knowledge.	It may be inaccurate if the taxonomies of the SRS specifications are not done correctly.

Table 4: Features comparison of the two approaches

Features	Kim et al. Approach (Kim et al., 2003)	Alternative Approach using SRS
Definition of Knowledge	Information captured through interviews, brainstorming sessions, focus groups, qualitative text analyses and task environment analyses.	A form of knowledge in a specific context based on SRS.
Knowledge Base Construction	In accordance with a manufacturing company environment.	In accordance with SRS guidelines upon news releases.
Output	Expert knowledge about the hot rolling process, a core part of steel production process.	News releases for social networking community.
Step1 - Knowledge Definition	This is done by deciding on the scope and level of details for the intended information by using questionnaires and interview techniques.	This step is not required.
Step2 - Identification of Information Context	Not Applicable.	Extracted document information has to be specified in a certain context.
Step3 - Process Map Analysis	Define knowledge flow of the processes of a hot rolling mill through human interactions.	Not Applicable.
Step4 - Knowledge Extraction	People are involved to identify the sources of knowledge such as operation manual and domain knowledge of onsite technician.	Programs are used for the extraction.
Step5 - Knowledge Profiling & Linking	Knowledge profile consists of attributes such as title, creation date, author, expert, location, description. Knowledge linking is built for identifying the knowledge flow.	Knowledge profile entries population and linkages among the entries could be achieved by connecting to the SRS knowledge ontology.
Step6 - Knowledge Map Validation	Manual validation is required.	Validation is done automatically by SRS.

5.0 CONCLUSION

It is felt that an existing knowledge mapping approach (Kim et al., 2003), which requires a continuous development and validation of the organizational knowledge ontology manually, is both tedious and time-consuming. An alternative knowledge mapping approach based on the SRS guidelines as the knowledge ontology is then presented. It does not need any manual knowledge ontology development and validation. And the knowledge profile entries could be populated by linking to the SRS guidelines.

As part of the future work, a prototype will first be developed to use the SRS guidelines for maintaining a knowledge profile for news releases to be published in a social networking mashup for students intending to study abroad. The news releases will be manually copied, and the SRS-based knowledge profile will also be manually maintained. The next milestone is the research of a knowledge mapping tool for representing the news releases in an easily navigational manner. Once the knowledge mapping tool has been user-accepted, software agents will be designed and developed to automate the moving of news releases to the mashup. Subsequently, the team will pursue its ulterior aim of dynamically generating knowledge profiles and thus knowledge maps, as and when required.

REFERENCES

Alavi, M. & Leidner, D.E. (1999). Knowledge management systems: issues, challenges, and benefits. *Communications of the AIS,* Vol. 1, No. 2es.

Baxendale, P.B. (1958). Machine-Made Index for Technical Literature—An Experiment. *IBM Journal of Research and Development,* Vol. 2, No. 4, pp. 354-361.

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.

Edmundson, H.P. (1969). New Methods in Automatic Extracting. *J. ACM,* Vol. 16, No. 2, pp. 264-285.

Ercan, G. & Cicekli, I. (2007). Using lexical chains for keyword extraction. *Information Processing & Management,* Vol. 43, No. 6, pp. 1705-1714.

Fernández-García, N. & Sánchez-Fernández, L. (2004). Building an Ontology for NEWS Applications. In *International Semantic Web Conference.* Hiroshima, Japan.

Giuffrida, G., Shek, E.C. & Yang, J. (2000). Knowledge-based metadata extraction from PostScript files. *Proceedings of the fifth ACM conference on Digital libraries*, pp. 77-84.

Han, H., Giles, C.L., Manavoglu, E., Hongyuan Zha, A.H.Z., Zhenyue Zhang, A.Z.Z. & Fox, E.A.A.F.E.A. (2003). Automatic document metadata extraction using support vector machines. In *Digital Libraries, 2003. Proceedings. 2003 Joint Conference on.* Giles, C.L. (Ed.). pp. 37-48.

Hu, Y., Li, H., Cao, Y., Teng, L., Meyerzon, D. & Zheng, Q. (2006). Automatic Extraction of Titles From General Documents Using Machine Learning. *Information Processing & Management*, Vol. 42, pp. 1276-1293.

IPTC & NAA (2003). IPTC - NAA Subject Reference System Guidelines version 3 / November 2003.

Kim, S., Suh, E. & Hwang, H. (2003). Building the knowledge map:an industrial case study. *Knowledge Management*, Vol. 7, No. 2, pp. 34-45.

Lê, T.H. & Lamontagne, L. (2002). Structuring Organizational Memories using Multi-Dimensional Knowledge Networks. *Workshop on Knowledge Management & Organizational Memories*.

Liebowitz, J. (2004). Will knowledge management work in the government? *Electronic Government, an International Journal*, Vol. 1, No. 1, pp. 1-7.

Littlefield, A. (2002). Effective enterprise information retrieval across new content formats. In *Proceedings of the seventh search engine conference.*

O'Dell, C. & Grayson, C.J. (1997). If We Only Knew What We Know: Identification and Transfer of Internal Best Practices. In *Best Practices White Paper* American Productivity & Quality Center.

Speel, P.-H., Shadbolt, N., Vries, d., Dam, W.V., P.H. & O'Hara, K. (1999). Knowledge mapping for industrial purposes. In *12th Workshop on Knowledge Acquisition Modeling and Management.* Alberta, Canada.

Turney, P.D. (2000). Learning Algorithms for Keyphrase Extraction. *Information Retrieval*, Vol. 2, No. 4 pp. 303-336.

Wesley, T.C. & Jihoon, Y. (2000). Extracting sentence segments for text summarization: a machine learning approach. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval.* Athens, Greece: ACM.