Thai Metadata Extraction by Using Case-based Reasoning

Krisda Khankasikam

Faculty of Information and Communication Technology, University of Phayao, Thailand 56000, E-mail: KrisdaK@gmail.com

***Abstract* – This paper reports an experience of human-assisted process to extract metadata from Thai documents. Nowadays, a number of Thai archives are placed online for sharing increasingly because the Internet infrastructure is powerful preserving and sharing knowledge require appropriate processes. Metadata, data about data, is a very useful information technology today because it helps users to differentiate significant from non-significant documents. The manually harvesting of these metadata elements is highly labor-intensive, costly and time-consuming then automated is a key to successful preservation. The experiment, a prototype system by using Case-based Reasoning algorithm for metadata extraction is introduced. Cased-based Reasoning is an approach in artificial intelligence that differs from other approaches. The Thai metadata extraction were performed on some Thai articles which content related to sufficient economy and Thai folk wisdom and was evaluated the approach by using the standard precision, recall and f-measure indices. The study illustrated that this approach helps knowledge workers in a domain to come together, share educational material and greatly reduce the labor work of metadata creation process.**

***Keywords* – Case-based Reasoning, Metadata, Dublin Core, Information Extraction.**

1. INTRODUCTION

This section contains the background and problem description of the study. With the growth of the Internet and related tools, there has been a rapid growth of online resources. However, lack of metadata available for these resources stops their dissemination on the Internet [5]. Metadata can help resource discovery, according to Doane's estimation [17] a company's use of metadata in its intranet may save about $8,200 per employee by reducing employee time for searching, verifying, and organizing the files. According to Rosenfeld's presentation in the DCMI 2003 workshop [2], it would take about 60 employee-years to create metadata for 1 million documents.

Realizing the benefits of metadata, most modern digital libraries support processes for manually metadata extraction as part of the publication process [4, 13]. However, metadata does not exist for legacy documents that mostly have the form of scanned images either in Portable Document Format (PDF) format or some image formats. There are good commercial tools for scanning and applying Optical Character Recognition (OCR) to generate an electronic version of a document. Nevertheless, there is a lack of good tool that can take an electronic version of a scanned document and extract the metadata from the document [24]. The process of creating metadata manually is expensive, labor-extensive and time-consuming [3] for a large collection. The costs for manual metadata creation make a great inspiration for creating the automated metadata extraction tools.

Manually extracting metadata from Thai electronic documents has many problems. Three problems that are worthy to mention are variety of Thai electronic document formats, time-consuming and quality of extracted metadata. Regarding to a problem on variety of Thai electronic document form, the electronic documents gathering from various sources have various layout and forms. The information, we want to extract, are not in the same position. In relevant to a problem on quality of extracted metadata, the metadata produced by manually metadata extraction may contain errors both from original documents and error from human entering data. To obtain a high quality metadata, the extracted metadata should be reviewed carefully. However, manually reviewing all extracted metadata could be time-consuming and costly. The problem of time-consuming is described perfectly in Rosenfeld's paper in the DCMI 2003 [2], it would take about 60 employees per year to create metadata for one million documents. The costs for manual metadata creation make a great case for the automated metadata extraction tools. This paper proposed case-based reasoning and dublin core metadata to solve these problems.

Besides the three problems from a paragraph, the automatic Thai metadata extraction can be the cause of problem in the study. The reason of difficulty on automatic Thai metadata extraction is characteristic of Thai language as sentences are written as a long series of characters without word or sentence markers [15]. The Thai alphabet consists of 44 consonant 32 vowels 4 tone marks and there is no capital letter. There are no changes in word form or word inflection as an expression of tense, case or gender; word ordering plays an important part in determining the syntactic role of word. The same form of words in different positions contains different syntactic properties and therefore conveys different meanings [8]. To express tense and case, additional words often are inserted to clarify the meaning. Thai grammar does not follow the extended projection principle, as found in English, where a sentence must have an overt subject. The subject can be omitted even if it is pronominal; this characteristic is referred to as null subject parameter. Thai contains relatively few headwords. Many Thai words are

formed from a combination of different nouns, verbs and auxiliaries to form compound nouns [9]. This paper proposed information extraction that suitable for Thai language to solve this problem.

The rest of this paper is organized as follow. Section 2 provides a brief related works. Section 3 describes proposed framework, followed by a description of the experimental results in Section 4. Discussion is discussed and conclusion is drawn in Section 5 and 6 respectively.

2. RELATED WORK

This section aims to explore case-based reasoning and information extraction. Besides, this one describes details of metadata and dublin core metadata element set.

2.1 Case-based Reasoning

Case-based Reasoning (CBR) [1, 11] is an approach in artificial intelligence that differs from other artificial intelligence approach [23]. Instead of depending on general knowledge of a domain or depending on knowledge gained by deduction from rule of a problem domain, CBR depends on knowledge that is previous experience of problem solving [20]. In CBR, new problems are solved by remembering solution to problem which is similar to the current problems [10]. As the problem cases and the remembered cases are often not perfectly matched cases the remembered solutions are modified in a way that at least parts of the case can be used. This process is known as the case adaptation. The set of CBR principles are defined as a cycle composing four activities called the CBR cycle as shown in Figure 1.

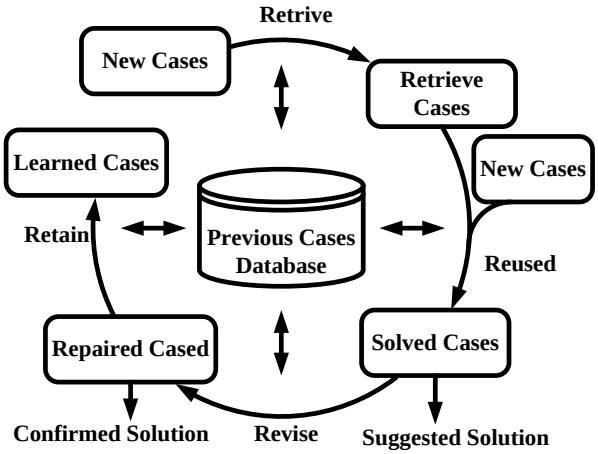

Figure 1 Case-based reasoning cycle

The CBR cycle was developed by Aamodt [1]. An initial description of a problem is a new case. This new case is used to RETRIEVE a case from the database. The retrieved case is combined with the new case through REUSE into a solved case. Through the REVISE process this solution is tested for success. During RETAIN, useful experience is retained for future reuse, and the database is updated by a new learned case.

Problem solving using CBR is more effective than other AI approaches [7] because of the following reason. Increasing problem solving efficiency, the problem solving process does not start from the blank. The appropriate solution is taken to derive the current problem. A similar new problem will recall the store solution of the similar problem. Better quality of solution, in Rule-Based reasoning, rules will be uncompleted when principle of domain is not well understood and the formulations of complex rules of some domains are difficult. But CBR allows problem solving informally, in such domain as cases capture associations between situations, solutions and outcomes. The solution suggested by cases may be more accurate than suggested by chains of rules [6]. User acceptance, because the solutions are on the basis of what really happened it is likely to be more confidently accepted [16].

2.2 Information Extraction

Information extraction is one of the Natural Language Processing (NLP) tasks. The purpose of information extraction, in terms of the NLP domain, is to create a system that process the digital documents written in any of the natural languages and identify essential information. The information extraction system can be known as a set of linguistic tools and resources that used to perform specific task related to the NLP. In general, the information extraction system can be presented as a black box (Figure 2) that uses linguistic resources to implement a given task. A user passes documents in a natural language to the information extraction system and receives the result of the processing. The results depend on the task characteristic.

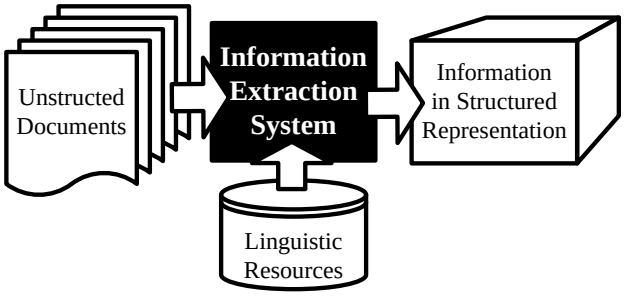

Figure 2 Information extraction system

The common architecture of information extraction system consists of 4 main modules (Figure 3). The general architecture of IE systems is based on a pipeline processing where set of modules is executed in given order. An output of one module is an input for another. The order in where subsequent modules are executed is essential because some modules provide or require information required or provided by other modules. Depending on the task specification and the language characteristic different types of modules are plugged into the pipeline.

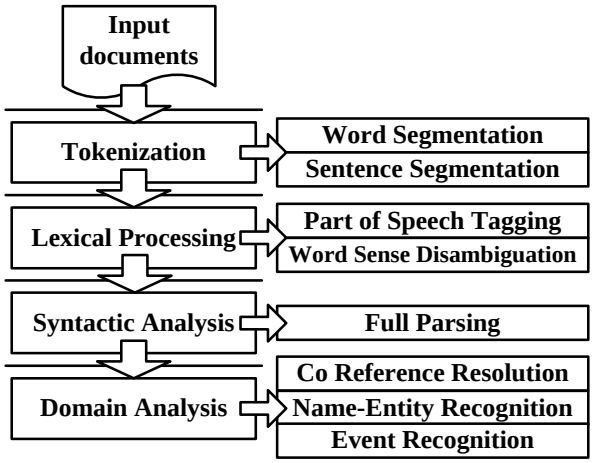

Figure 3 Common Architecture of information extraction

2.3 Metadata

This section presents the concept and significance of metadata usage. Nowadays, knowledge management has become a vital challenge. There are knowledge is placed in huge computer system in the form of digital documents. Digital documents can be based on individual stand-alone document files such as Adobe PDF, MS Word and MS PowerPoint documents or on internal document types in the computer system. Usage of digital documents has introduced many new sharing and efficiency dissemination knowledge. However, usage of computer systems can easily limit knowledge sharing while the correct documents are difficult to locate. With a fast increasing collection of documents, locating the correct document becomes more challenging.

The simple definition of metadata is ‘‘data about data’’. According to Metadata Basics [27], metadata is data that provides information about documentation. Metadata is also machine understandable information for the web. In the digital environment, the representative role of metadata is the key because many resources are not easily browse and others do not carry clear data about themselves. Metadata can be used to give descriptions of the document. These descriptions can be a part of the data used for document querying and retrieval. This is allowing new users to gain knowledge of the existing recourses and their most central characteristics.

2.4 Dublin Core

Dublin Core is a short form Dublin Core Metadata Initiative (DCMI) [25]. Dublin Core was developed in 1995 metadata workshop which is sponsored by the Online Computer Library Center and the National Center for Supercomputing Application. The objective of Dublin Core development is to advance the state of the art in the development of metadata records for networked information resources. The outcome of the first workshop is the set of 13 metadata element which is called Dublin Core Metadata Element Set (DCMES). Dublin Core has gained the special importance among the resource description communities [25].By the third workshop, Dublin Core is upgraded to 15 metadata element. The Dublin Core Metadata Element Set includes: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, and Rights. The details of each element are shown in Table 1.

Table 1 Detail of element set

Field	Description
Title	A name given to the resource.
Creator	An entity primarily responsible for making the content essence of the resource.
Subject	The topic of the content essence of the resource.
Description	An account of the content essence of the resource.
Publisher	An entity responsible for making the resource available.
Contributor	An entity responsible for making contributions to the content essence of the resource.
Date	A date associated with an event in the life cycle of the resource.
Type	The nature or genre of the content essence of the resource.
Format	The physical or digital manifestation of the resource.
Identifier	An unambiguous reference to the resource within a given context.
Source	A reference to a resource from which the present resource is derived.
Language	A language of the intellectual content essence of the resource.
Relation	A reference to a related resource.
Coverage	The extent or scope of the content essence of the resource.
Rights	Information about rights held in and over the resource.

Each Dublin Core element is optional and may be repeated. The DCMI has established standard ways to refine elements and encourage the use of encoding and vocabulary schemes. There is no prescribed order in Dublin Core for presenting or using the elements.

3. METHODOLOGY

This section describes the methodology of the automatic Thai metadata extraction. The methodology composed of three main components: case retrieval module for comparing problem case and stored case, metadata generating module for automatically extracting metadata from electronic Thai documents, and metadata verification module for identifying and correcting the errors in extracted metadata. The architecture of the system is shown in Figure 4.

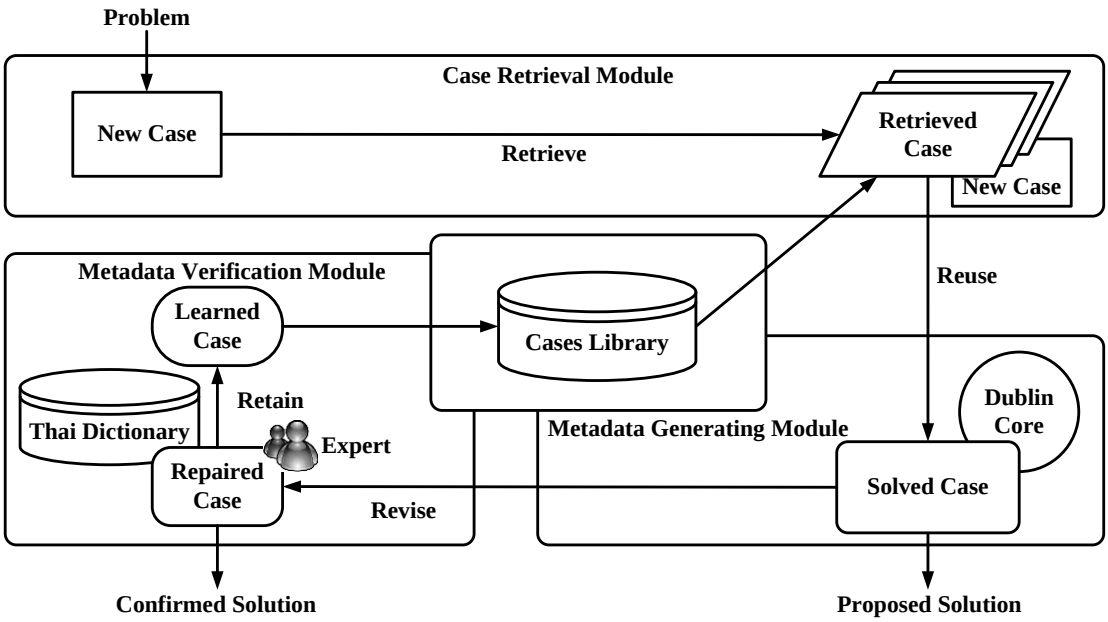

Figure 4 Architecture of system

3.1 Case Retrieval Module

A case retrieval module is used for a comparing problem case and a stored case using Nearest Neighbor Retrieval (NRR) technique [19]. NRR is a technique to measure how similar the target case comparing to a source case. It processes retrieval of cases by comparison of a collection of weighted attributes in the target case to source cases in the CBR library. If there is no matched case in the CBR library, CBR system will return the nearest matched source case. The return of the nearest case match can be represented by the following Equation.

$$\text{Similarity}(T,S) = \sum_{i=1}^{n} f(Ti,Si) * Wi \qquad (1)$$

Where T is the target case, S is the source case, n is the number of attributes in each case, i is an individual attribute from 1 to n, f is a similarity function for attribute i in cases T and S and W is the importance weighting of attribute i. The equation of the NNR represents the sum of similarity of the target case to the source case for all attributes multiplied by the importance weighting of individual attributes. The CBR system therefore retrieves a meaningful case that may provide a detailed solved problem description to a new problem.

3.2 Metadata Generating Module

A metadata generating module is responsible for automatically extracting the metadata from Thai electronic documents. In the research, Thai keywords that matched with Dublin core categories will be extracted into database. For example, the header of student thesis abstract, as shown in Figure 5, is roughly detected fifteen keywords which are Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage and Rights

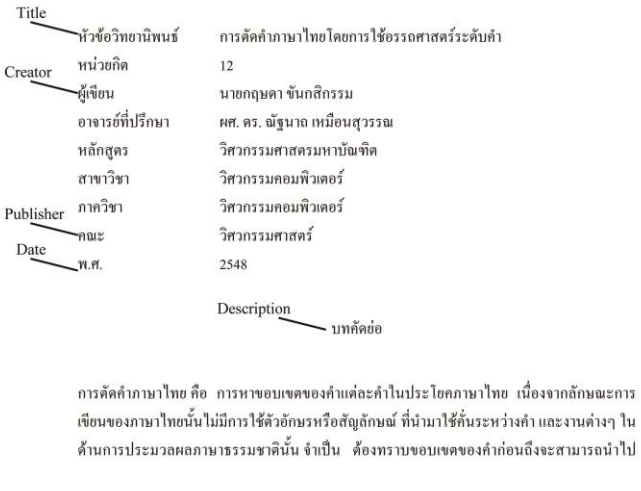

Figure 5 Thesis header

After analyzing the header of thesis abstract, the system can easily identify the boundary of each part by using special symbols (e.g. ‘:’) and keyword markers (e.g. |Title|, |Creator|, |Subject|, |Description|, |Publisher| and |Date|). Thus, the system can create an analyzed structure of thesis abstract by using those boundary markers as a part separation point. After extracted the metadata, the metadata verification module will help users identifying and correcting the error in extracted metadata in order to obtain a high quality metadata.

3.3 Metadata Verification Module

The extracted metadata may contain errors both from the metadata generating module and original documents. To gain a high precision metadata, it is necessary to identify and correct the errors before using the metadata [3]. The proposed framework is an integrated mechanism in order to help users to correct the errors. Regarding to errors in metadata creation module, the system may not be able to extract some documents due to incompleteness of source case or defect in the documents. In this process, the system will display error messages from metadata creating module to guide users for correcting the errors. The users make a decision to response with the errors. There are many sophisticated methods that can be employed in order to help the users detecting and correcting the errors. The good choice is to use a spelling correction technique [21] to detect errors and suggest the correction.

4. EXPERIMENTAL

The experiments with Thai metadata extraction are performed on the various sources which content related to sufficient economy and Thai folk wisdom. These documents are written in original Thai language without using words originate from English language or words standardize by The Royal Institute (TRI), an institution concerning academic matters as the compilation and publication of dictionaries, encyclopedias, terminologies and taxonomies. Because this system and its techniques can process data using original Thai words only.

4.1 Test Cases

The typical approach for testing a metadata extraction system is to create a perfect metadata by expert for comparing with the results. Unlike English, standard data set in Thai are not yet available for evaluating metadata extraction system. However, in order to observe characteristics of the proposed methodology, Thai documents where content related to sufficient economy and Thai folk wisdom, including Thai theses (D1.TT), news (D2.NW), columnist's article (D3.CA), academic papers (D4.AP), articles concerned with royal words (D5.RW) and text books (D6.TB) are collected to make the data sets. Each data set consists of 500 documents, and document sizes range from 1 to 10 pages. There are collaboration from a student in the Faculty of Humanities of Naresuan University for collecting the documents from various sources, such as from the Internet, conference proceeding and newspaper.

4.2 Evaluation

The results are evaluated by using three widely used methods [18] composed of Precision, Recall and F-measure indices. Let R*a* is the number of correctness extracted metadata, A is the number of ideal extracted metadata and R is the number of extracted metadata in actual answer. Precision of the algorithm can be calculated as the fraction between the numbers of correctness extracted metadata and the number of ideal extracted metadata. In this research work, the Precision index is defined as

$$\text{Precision} = \frac{\text{R}a}{\text{A}} \tag{2}$$

Then, the Recall index is the fraction between the numbers of correctness extracted metadata and the number of extracted metadata in actual answer; that is,

(3)

Finally, the average value of the Precision and Recall indices called "F-Measure index" can be calculated as follows:

$$\text{F}_{\beta} = \frac{(1+\beta^2) * \text{Precision} * \text{Recall}}{\beta^2 * \text{Precision} + \text{Recall}} \tag{4}$$

β is a factor which is used to adjust weight between Precision and Recall indices. Two other commonly used F-measures are the F_2 measure which weights Recall twice as much as Precision and the $F_{0.5}$ measure which weights Precision twice as much as Recall. In this research, β is defined as 1 because Precision and Recall are evenly weighted. This F_1 measure is also known as the Harmonic mean specified by

$$\text{F} - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

4.3 Results

In experimental result, the experiments are performed on a standard benchmark with the manually extracted metadata that generating by expert in metadata extraction of Naresuan University. In data analysis, there is collaboration from Thai metadata extraction experts of Naresuan University to check the correctness of results. The experimental evidence has provided that the proposed algorithm gives acceptable performance. Table 2 shows the summary of evaluation results from the system comparing with the existing metadata.

Table 2 Experimental result

Data set	Result of the proposed system				Result of manually extracted		
	Precision	Recall	F-measure		Precision	Recall	F-measure
D1.TT	90.78	91.52	91.15		83.64	84.32	83.98
D2.NW	75.77	86.24	80.67		70.39	80.12	74.94
D3.CA	65.32	82.44	72.89		72.26	91.20	80.63
D4.AP	89.88	90.98	90.43		80.67	81.66	81.16
D5.RW	62.31	89.99	73.63		67.58	97.60	79.86
D6.TB	85.36	87.16	86.25		78.12	79.77	78.94
Average	78.24	88.06	82.50		75.44	85.78	79.92

From Table 2, the average results of precision, recall and f-measure from the proposed system are better than from manually extracted metadata 2.80% (78.24% - 75.44%), 2.28% (88.06% - 85.78%) and 2.58% (82.50% - 79.92%), respectively. The experimental results show that using the proposed methodology can reduce the labor work of metadata extraction process. Further, experiments on larger number of documents are needed to determine the performance of the system.

5. DISCUSSION

According to the definition of the previous three measures, these measure indices are direct performance measures how well metadata extraction does [18]. In general, the Precision index is used to indicate direct correctness of metadata extraction. For both Recall and F-measure indices, they do not directly indicate metadata extraction performance. However, they are useful for indirect correctness measure of metadata extraction. Either Recall index or F-measure index is one of the most important performance measures in information retrieval system. Actually, the F-measure index represents the compromise between the Precision and Recall indices.

Moreover, the framework and architecture of the research can be applied with some Asian language that characteristic similar to Thai language such as Chinese Japanese and Korean. But the techniques used in algorithm design must be adjusted to be suitable for those languages [12, 22].

6. CONCLUSION

In this paper, the framework for automatic metadata extraction from electronic heterogeneous Thai documents is presented. There are many researches on metadata extraction problems. However, in Thai, we are in the initial stage of developing mechanism for automatic metadata extraction. It is a challenge to extract metadata in Thai documents, because they are extremely different from documents written in English. The structure of written Thai is highly ambiguous, which requires more complex techniques than are necessary to perform comparable metadata extraction tasks in most European languages. The experimental result suggested that using the proposed framework is efficient and contains high-precision metadata extraction. The results with 3,000 documents show that using this system can reduce the labor work of metadata creation process. The system performs the level of precision at 62.31% - 90.78% depending on the characteristic of the input.

7. ACKNOWLEDGMENT

This work was supported by the Department of Applied Science, Faculty of Science and Technology, Nakhon Sawan Rajabhat University. The authors would like to thank the support from College of Arts, Media and Technology, Chiang Mai University.

8. REFERENCES

[1] A. Aamodt, and E. Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches," *Artificial Intelligence Communications*, vol. 7, pp. 39-59, 1994.

[2] A. Crystal, and P. Land, "Metadata and Search," *in Proc. the Global Corporate Circle Dublin Core Metadata Initiative 2003 Workshop*, Seattle, Washington, USA, 2003.

[3] A. Kawtrakul, and C. Yingsaeree, "A Unified Framework for Automatic Metadata Extraction from Electronic Document," in *Proc. the International Advanced Digital Library Conference*, Nagoya, Japan, 2005.

[4] D. B. Ganesh, "Organizing knowledge in the knowledge development cycle," *Journal of Knowledge Management*, vol. 4, pp. 15-26, 2000.

[5] E. Motta, S. B. Shum, and J. Domingue, "Ontology-driven document enrichment: principles, tool and applications," *International Journal of Human-Computer Studies*, vol. 52, pp. 1071-1109, 2000.

[6] E. Tshi, J. B. Garner, and S. Staab, "The role of Artificial Intelligence in Knowledge Management," *Knowledge-Based Systems*, vol. 13. pp. 235-239, 2000.

[7] G. Walsham, "Knowledge Management: The Benefits and Limitations of Computer Systems," *European Management Journal*, vol. 19, pp, 559-608, 2001.

[8] G. W. Furnas, T. K. Laudauer, L. M. Gomez, and S. T. Dumais, "The Vocabulary Problem in Human-System Communication," *Communications of the ACM*, vol. 30, pp. 964-971, 1987.

[9] H. T. Koanantakool, T. Karoonboonyanan, and C. Wutiwiwatchai, "Computers and the Thai Language," *IEEE Annals of the History of Computing*, Vol. 31, pp. 46-61. 2009.

[10] I. D. Watson, *Applying Case-Based Reasoning: Techniques for Enterprise Systems*. California: Morgan Kaufmann, 1997.

[11] I. D. Watson, "Knowledge Management and Case-Based Reasoning: a Perfect Match?," in *Proc. the 14th International Florida Artificial Intelligence Research Society Conference*, Key West, Florida, USA, 2001.

[12] J. Liebowitz, "Knowledge management and its link to artificial intelligence," *Expert Systems with Applications*, vol. 20. pp. 1-6. 2001.
[13] J. O. Everett, D. G. Bobrow, R. Stolle, R. Crouch, V. D. Paiva, C. Condoravdi, M. V. D. Berg, L. Polanyi, "Making ontologies work for resolving redundancies across documents," *Communication of the ACM*, vol. 45, pp. 55-60, 2002.
[14] K. Khankasikam, "Knowledge Capture for Thai Word Segmentation by Using CommonKADS," in *Proc. The 2nd international conference on Computer and Automation Engineering*, pp. 307-311, Singapore, 2010.
[15] K. Khankasikam, and N. Chakpitak, "A Unified Framework for Thai Metadata Extraction Using Case-based Reasoning," in *Proc. the international conference on Advanced Computer Theory and Engineering*, pp. 210-214, Phuket, Thailand, 2008.
[16] K. M. Wiig, "Knowledge management: an introduction and perspective," *Journal of Knowledge Management*, vol. 1, pp. 6-14, 1997.
[17] M. Doane, "Metadata: Search and Meaningful ROI," *in. Proc. the Global Corporate Circle Dublin Core Metadata Initiative 2003 Workshop*, Seattle, Washington, USA, 2003.
[18] R. Baeza-Yates, and B. Ribeiro-Neto, *Modern Information Retrieval*, New York: Addison Wesley, 2002.
[19] S. H. Kang, and S. K. Lau, "Intelligent Knowledge Acquisition with Case-Based Reasoning Techniques," in *Proc. the 13th Australasian conference on Information Systems*, pp.1-8, Melbourne, Australia, 2002.
[20] S. Russell, and P. Norvig, *Artificial Intelligence: A Modern Approach*. 2nd ed, New York: Prentice-Hall, 2002.
[21] W. Aroonmanakun, "Collocation and Thai Word Segmentation," in *Proc. the 5th Symposium on Natural Language Processing & the 5th Oriental COCOSDA Workshop*, pp. 68-75, Hua Hin, Thailand, 2002.
[22] W. C. Holsapple, and D. K. Joshi, "A Collaborative Approach to Ontology Design," *Communications of the ACM*, vol. 45, pp. 42-47, 2002.
[23] W. Dubitzky, G. Buchner, and J. Azuaje, "Viewing Knowledge Management as a Case-Based Reasoning Application," *in Proc. the 16th National Conference on Artificial Intelligence and the 11th Annual Conference on Innovative Applications of Artificial Intelligence*, pp. 23-27, Orlando, Florida, USA, 1999.
[24] Y. Malhotra, "Knowledge Management: Life Like Autonomous Agents," *Communications of the ACM*, 38, pp. 108-114, 1998.
[25] (2010, Jan.). DCMI Specifications, [Online]. Available: http://dublincore.org/specifications/
[26] (2010, Jan.). Dublin Core Metadata Element Set, Version 1.1., [Online]. Available: http://dublincore.org/documents/dces/
[27] (2010, Mar.). *Metadata Basics*, [Online]. Available: http://dublincore.org/metadata-basics/