DOCINER: A Document Indexation Tool for Learning Objects

Suphakit Niwattanakul
School of Information Technology,
Suranaree University of Technology
Nakhon Ratchasima, Thailand
suphakit@sut.ac.th

Michel Eboueya
L3I
University of La Rochelle
La Rochelle, France
michel.eboueya@univ-lr.fr

Philippe Martin
Eurecom
Sophia-Antipolis, France
phmartin@phmartin.info

***Abstract*-In this paper, we present a method we implemented to help a user index documents (and, in particular, learning objects) according to a given set of concepts (terms referring to domains or topics). The user first associates keywords to the concepts. Our method uses such associations to suggest simple rules for indexing a document by concepts according to the keywords this document contains. Then, our system uses those rules to perform the indexation of documents.**

Keywords: document indexation, formal concept analysis

I INTRODUCTION

Nowadays, the use of online learning resources is increasingly common in education focusing on course development [1]. Many researchers pay attention to the issue of reusability of learning resources. Course developers aim to reuse these learning resources for developing a new course because the reuse of learning resources can save time and money for course development.

In terms of course development, a course generally consists of units of instruction called Learning Objects (LOs). A learning object is any digital resource that can be used or reused to support learning ([2], [3]). LOs can be texts, presentations, quizzes, video clips, tutorials, maps, animations, assessments, etc. LOs are accessible and searchable through Web-based repositories and mediators. In a repository, LOs reside within a database on the server hosting the Web-enabled gateway to the collection, whereas a mediator contains no LOs but links to objects residing on remote servers.

A Learning Object Repository (LOR) is a system that provides functions to collect LOs available on computer networks and/or Databases. LORs can play the role of a repository and/or a mediator. The metadata associated to documents in LORs facilitates the search and management of LOs. Many LORs are developed based on the IEEE LOM metadata standard [2] and its application profiles such as SCORM [4], CanCore [5], Normetic [6] and UK LOM Core [7].

The use of educational metadata standards allows LOs to index and classify by classification systems but these metadata standards lack a formal semantics and they introduce the problem of incompatibility between heterogeneous metadata descriptions or schemas across domains [8]. Ontologies can be used for indexing learning resources by using concepts (topics or domains).

Although the use of learning content management systems is becoming common in most educational organizations and the number of educational resources is huge, most of these resources are hidden in repositories and cannot be easily found. This can impede their potential use and reuse. Searching for LOs in LORs by using keywords leads to problems since different LOs may be about the same topic while containing different keywords.

Traditional information retrieval technology is based on the occurrence of words in documents. Semantic Web technologies ([9], [10]) may be used for information retrieval on the Web [11]. We use a lightweight semantic retrieval technique to ease the retrieval of LOs: 1) the user first associates keywords to the concepts (terms or lists of terms referring to domains or topics), 2) via a direct application of Formal Concept Analysis (FCA), our system uses such associations to suggest simple rules for indexing a document by concepts according to the keywords this document contains, 3) our system uses those rules to perform the indexation of documents. After presenting the framework of our technique, we present its second step.

II KNOWLEDGE ORGANIZATION SYSTEM BASED ON ONTOLOGIES

For knowledge sharing, “an ontology is a formal, explicit specification of a shared conceptualization” [12]. A specification of conceptualization consists in a list of objects and relations that hold among them. “Explicit” means that objects, concepts, and other entities are explicitly defined. “Formal” implies that the ontology should be machine-readable and logic-based. The main structure of an ontology model consists in concepts or classes, and relations.

Researchers are developing a method to automatically extracting structured information from documents by using information extraction technologies. Several tools or systems for building domain ontologies from text are TEXCOMON (TEXt-COncept Map-Ontology) ([13], [14]) and TEXT-TO-ONTO Ontology Learning Environment [15]. As described in [16], the process of concept indexing consists in (i) extracting entities from unstructured text-based content using lexical tags and rules, (ii) identifying concepts and adding ontology tags to them using semantic

rules, and (iii) merging entity and concept information into a concept index.

The term "Knowledge Organization System" (KOS) refers to all types of schemes for organizing information and promoting knowledge. KOSs include classification schemes that organize materials at a general level such as subject headings and authority files. Authority files are used to control variant versions of key information such as geographic names and personal names. KOSs also include highly structured vocabularies, such as thesauri, and less traditional schemes, such as semantic networks and ontologies [17].

The research of knowledge representation is developing and testing the knowledge representation language [18]. Knowledge representation systems allow the concepts and inference rules to be used by machines. Nowadays, the SKOS (Simple Knowledge Organisation System) model is developing as a knowledge representation system and can be used for developing Web contents thanks to the Semantic Web [19].

The SKOS model is designed by the W3C Semantic Web Best Practices and Deployment Working Group. SKOS Core is a model designed for expressing the basic structure and content of concept schemes. A concept scheme is a set of concepts, optionally including statements about semantic relations between those concepts. Concept Schemes can be thesauri, classification schemes, subject heading lists, taxonomies, terminologies, glossaries and other types of controlled vocabulary.

III Framework of the Keyword and Concept Extraction Method

Our information indexation/extraction technique fits the definition of [20]: "*the identification, and consequent or concurrent classification and structuring into semantic classes, of specific information found in unstructured data sources, such as natural language text, making the information more suitable for information processing tasks.*" To achieve this, many information extraction methods have been proposed: name entity recognition, noun phrase coreference resolution, semantic role recognition, entity relation recognition, time line recognition, etc.

Our own named entity recognition technique starts by comparing words in texts with index words coming from a lexical database such as WordNet [21]. These words are then associated to keywords and these keywords are associated to concepts (also coming from WordNet and/or provided by the user).

As illustrated in Fig. 1, the keywords "Computer Programming" and "Mathematics" are identified according to words from the text. Then these two keywords are used for identifying the concept "Computer Science."

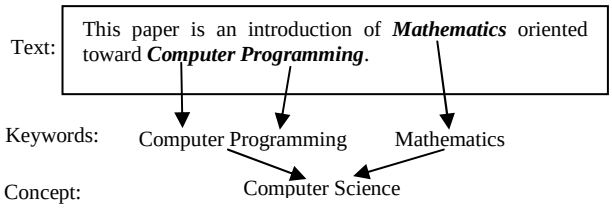

Fig. 1 Keyword and concept extraction method

The framework of our method is illustrated in Fig. 2.

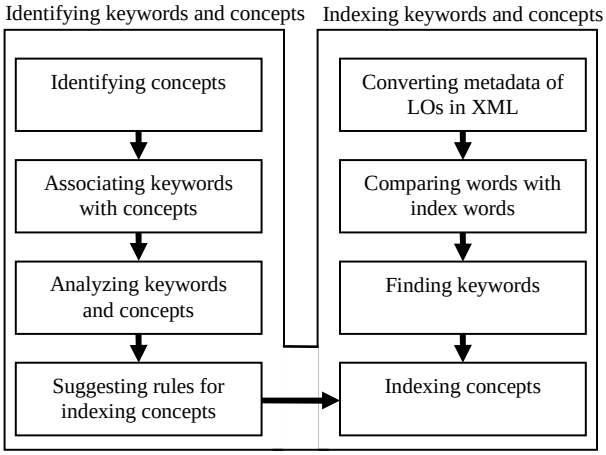

Fig. 2 The framework of keyword and concept extraction method

For identifying keywords and concepts, data from two sources are used for identifying keywords and concepts which are: (i) words and their information from the WordNet dictionary and (ii) words and keywords from experts. The two data sources are transformed into a database based on an ontology model. These concepts are classified via classification systems such as controlled vocabularies and taxonomies using the SKOS ontology. The concepts and their keywords are analyzed through an FCA (Formal Concept Analysis) system ([22], [23]) to suggest rules for indexing concepts.

To index documents, we propose a tool called DOCINER (DOCument INdexation for Educational Resources) that first converts the metadata of the source LORs in XML. Then, within that textual metadata, it isolates the keywords it knows. Finally, it uses the indexing rules to associate each LO with concepts (topics or domains).

IV Suggesting Rules for Relating Keywords to Concepts

In DOCINER, associations between keywords and concepts come from WordNet and/or the user, and are represented using the SKOS ontology.

Fig. 3 illustrates such associations. DOCINER is based on the knowledge annotation and retrieval server SEWESE [24].

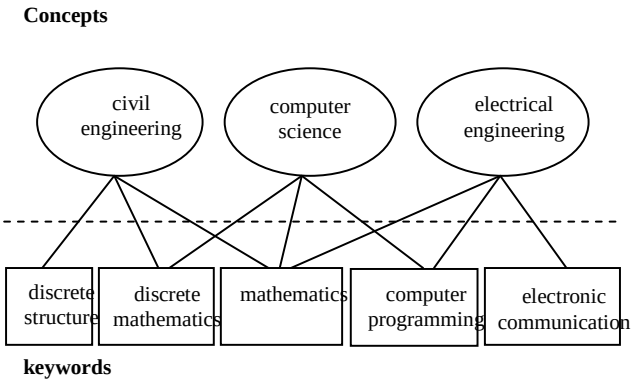

Fig. 3 Relating keywords and concepts

These associations can also be represented as in TABLE 1. This format permits you to apply basic techniques of FCA where two types of items (objects and attributes) relate to each other. In FCA, each relationship between an object and its related attributes is called a "formal concept". In TABLE 1 the formal concepts are shown via three rectangles.

TABLE 1
A FORMAL CONTEXT OF KEYWORDS AND CONCEPTS

Objects (Concepts)	**Attributes (Keywords)**				
	discrete structure	discrete mathematics	Mathematics	computer programming	electronic communication
civil engineering	X	X	X		
computer science		X	X	X	
electrical engineering			X	X	X

By using the above mentioned basic techniques of FCA, ToscanaJ [25] which is an open source is used as a tool for analyzing data and presenting these data with concept lattices in an image. The notation graph referred to as a "concept lattice" or a "Galois lattice" is used for representing formal concepts. A central notation of a concept lattice is a duality namely a "Galois connection" used for representing between two types of related items. The "concept lattice" shown in Fig. 4, can be derived from the previous table for representing formal concepts.

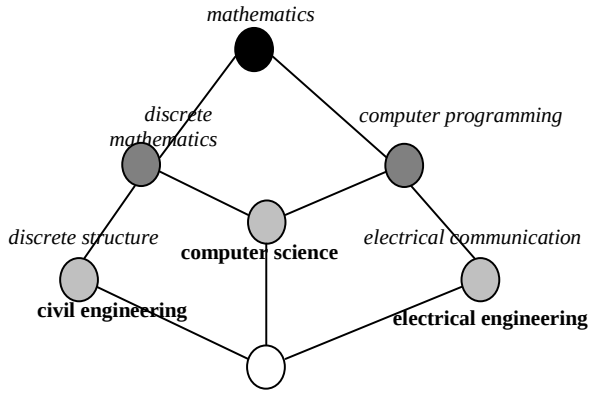

Fig. 4 A concept lattice for TABLE 1

From such a lattice, our method draws simple rules for indexing documents by concepts based on the keywords in these documents.

- **Rules that do not need to be approved by the user.**

For some keywords, there is only one related concept. In such a case, there is no ambiguity for document indexation. Using the notation "*list of keywords -> concept*", here are the rules that can be derived from or that case Fig. 4 for that case.

{"mathematics", "discrete mathematics", "discrete structure"} -> "civil engineering"
{"mathematics", "discrete mathematics", "computer programming"} -> "computer science"
{"mathematics", "computer programming", "electrical communication"} -> "electrical engineering"

- **Rules that need to be approved by the user.**

When a keyword is related to several concepts (i.e., domains or topics), the user might want to make a selection. Using the notation "*list of keywords ->? concept*", here are the rules that can be derived from or that case Fig. 4 for that case.

{"mathematics", "discrete mathematics"} ->? "civil engineering"
{"mathematics", "discrete mathematics"} ->? "computer science"
{"mathematics", "computer programming"} ->? "computer science"
{"mathematics", "computer programming"} ->? "electrical engineering"

The rules are represented in tuProlog ([26], [27]) and searched via the query mechanisms of tuProlog.

Our document indexation approach is close to the ones adopted in the TEXCOMON system ([13], [14]) and PALOMA [28] developed in the framework of LORNET (Learning Object Repositories Network) [29], both of which perform knowledge management from educational resources. However, these systems do not suggest indexation rules to the user. Indexation rules can be used for indexing documents by concepts to help retrieve these

documents. The DOCINER approach suggests such indexation rules by using an FCA system.

V Evaluation

Our evaluation relies on classic precision and recall measures (possibly combined in a F-measure) to assess the performance of the retrieval. Equations (1), (2), and (3) are used to calculate the values of precision, recall and F-measure [20].

$$precision = \frac{ard}{ad} \quad (1)$$

$$recall = \frac{ard}{trd} \quad (2)$$

$$F-measure = \frac{2 * precision * recall}{precision + recall} \quad (3)$$

Where ard = number of relevant documents in the result list
trd = total number of relevant documents in the document base
ad = number of documents in the result list

Values of precision, recall and F-measure are calculated by comparing the keywords in the result lists with keywords identified by an expert. **TABLE 2** shows the average results for 30 example documents. The method is tested in two steps, finding keywords and indexing concepts.

TABLE 2
Evaluation of the Keyword and Concept Extraction Method

Steps of evaluation	Precision	Recall	F-measure
Finding keywords	0.9933	0.9900	0.9861
Indexing concepts (with indexing rules)	1.0	0.9900	0.9945

As regards the precision values of finding concepts and indexing concepts, the precision value is increased in the process of indexing concepts. After identifying the concepts by using indexing rules, non-relevant keywords to such concepts are removed. However, the proposed method is only a prototype. It needs to be developed for an application in the future.

VI Conclusion

We have presented a document indexation approach. This approach can help users to associate documents or educational resources to concepts (terms referring to domains or topics) by using the occurrence of keywords in such documents in order that those documents can be retrieved by using the concepts. The advantage of this method is the suggestion of indexation rules to the user by implementing them in a way of knowledge management systems. The use of indexation rules help to remove non relevant keywords. The limit of this method is that concepts cannot be identified if there are no relevant words related to such concepts.

We shall evaluate our method by comparing our results with other concept/rule identification tools. To that end we shall re-use similarity measures between concepts and between keywords by using the well-known formula of similarity measures which is Jaccard's coefficient as described in [30].

References

[1] Caws, C., Friesen, N. & Beaudoin, M. A New Learning Object Repository for Language Learning: Methods and Possible Outcomes. Editor: Alex Koohang. In the Interdisciplinary Journal of Knowledge and Learning Objects. Volume 2, 2006.

[2] IEEE LTSC (Learning Technology Standards Committee). IEEE 1484.12.1-2002. Draft Standard for Learning Object Metadata, Institute of Electrical and Electronics Engineers. 2002. Retrieved from http://ltsc.ieee.org/ wg12/files/LOM_1484 _12_1_v1_Final_Draft.pdf

[3] Wiley, D.A. Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. 2002. Retrieved from http://reusability.org/ read/chapters/wiley.doc

[4] Betsy, S. Introduction to the SCORM for Instructional Designers. ADL (Advanced Distributed Learning). 2004. Retrieved from http://www. adlnet.gov/scorm/articles/article.aspx?id=4

[5] Friesen, N., Fisher, S. & Roberts, A. CanCore Guidelines Version 2.0: Introduction. 2003. Retrieved from http://www.cancore.ca/guidelines/ drd/

[6] Normetic. Profile d'application Normetic, version 1.1. 2006. Retrieved from http://www.normetic.org/

[7] UK LOM Core. UK Learning Object Metadata Core Draft 0.2. 2004. Retrieved from http://zope.cetis.ac.uk/profiles/uklomcore/uklomcore_ v0p2_may04.doc

[8] Stojanovic, L., Staab, S. & Studer, R. (2001). eLearning based on the Semantic Web. In WebNet2001 - World Conference on the WWW and Internet, Orlando, Florida, USA. Retrieved from http://citeseer. ist.psu.edu/ 501440.html

[9] Berners-Lee, T. Semantic Web on XML. XML 2000, Washington DC. Retrieved from http://www.w3.org/2000/Talks/1206-xml2k-tbl/

[10] Matthews, B. Semantic Web Technologies. JISC Technology and Standards Watch. 2005. Retrieved from http://www.jisc.ac.uk/ uploaded_documents/jisctsw_05_02bpdf.pdf

[11] Guha, R., McCool, R. & Miller, E. Semantic Search. WWW2003, May 20-24, 2003, Hungary. Retrieved from http://www2003. org/cdrom/papers/refereed/p779/ess.html

[12] Gruber, T. R. Toward principles for the design of ontologies used for knowledge sharing. International Journal of Human-Computer Studies, Vol. 43, Issues 4-5, November 1995, pp. 907-928.

[13] Zouaq, A. & Nkambou, R. Building Domain Ontologies from Text for Educational Purposes. To be published in the IEEE Transaction on Learning Technologies, 2008.

[14] Zouaq, A., Nkambou, R. & Frasson, C. Enhancing Learning Objects with an Ontology-based Memory. To be published in the IEEE Transactions on Knowledge and Data Engineering, 2008.

[15] Maedche, A. & Staab, S. The TEXT-TO-ONTO Ontology Learning Environment. 2000. Retrieved from http://citeseer.ist.psu.edu /275146.html

[16] Setchi, R.M. & Tang, Q. Concept Indexing using Ontology and Supervised Machine Learning. In Proc. of World Academy of Science, Engineering and Technology. Vol. 21 January 2007. ISSN 1307-6884.

[17] Hodge, G. Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files. The Council on Library and Information Resources. 2000. Retrieved from http://www.clir.org/ pubs/reports/ pub91/contents.html

[18] Kiryakov, A., Popov, B., Terziev, I., Manov, D. & Ognyanoff, D. (2005). Semantic Annotation, Indexing, and Retrieval. Elsevier's Journal of Web Semantics, Vol. 2, Issue (1), 2005. Retrieved from http://www.websemantics journal.org/ps/pub/2005-10

[19] Miles, A. & Brickley, D. SKOS Core Guide. W3C Recommendation. 2005. Retrieved from http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/

[20] Moens, M.F.. Information Extraction: Algorithms and Prospects in a Retrieval Context. Springer : Netherlands. 246 p. 2006.
[21] Fellbaum, C. WordNet An Electronic Lexical Database. The MIT Press. May 1998. Retrieved from http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=8106
[22] Priss, U. Formal Concept Analysis in Computer Science. In B. Cronin (Ed.). Annual Review of Information Science and Technology, ASIST, Vol. 40. 2006.
[23] Wolff, K. E. A First course in Formal Concept Analysis. In Proc. of the SoftStat'93. Gustav Fischer Verlag. 1994. Retrieved from http://www. fcahome.org.uk/fca.html
[24] Gandon, F & Durville, P. SeWeSe: Semantic Web Server. INRIA, France. Retrieved on November 15th, 2007 from http://www-sop.inria.fr/acacia/soft/sewese/
[25] Becker, P., Hereth, J. & Stumme, G. ToscanaJ An Open Source Tool for Qualitative Data Analysis. Advances in Formal Concept Analysis for Knowledge Discovery in Databases. In V. Duquenne and B. Ganter and M. Liquiere and E. M. Nguifo and G. Stumme (Eds.). 2002. Retrieved from http://citeseer.ist.psu.edu/587107.html
[26] Piancastelli, G. & Omicine, A. tuProlog 2.0: One Step Beyond. ALP Newsletter Digest 20(1). Association for Logic Programming, February-March 2007. Retrieved from http://alice.unibo.it/xwiki/bin/view/Tuprolog/Documents
[27] Piancastelli, G., Benini, A., Omicini, A. & Ricci, A. The Architecture and Design of a Malleable Objet-Oriented Prolog Engine. (Slide) 23th ACM Symposium on Applied Computing (SAC 2008), 16-20 March, 2008, Fortaleza, Ceará, Brazil. Retrieved from http://alice.unibo.it/xwiki/bin/view/Tuprolog/Documents
[28] Paquette, G. Apprentissage sur l'Internet: des plateformes aux portails à base d'objets de connaissance. In S. Pierre (Ed), Innovations et tendances en technologies de formation et d'apprentissage. Presses de l'école polytechnique de Montréal, pp. 1-30. 2005. Retrieved from http://www.licef.teluq.uquebec.ca/gp/eng/publications/campus_virtuel.htm
[29] Paquette, G. De la recherche à la pratique – La planète universitaire se met en réseau, 2006. Retrieved from http://www.ledevoir.com/2006/05/20/109509.html?282
[30] Doan, A., Madhavan, J., Domingos, P. & Halevy, A. Learning to map between ontologies on the semantic web. In Proc of the 11th International WWW Conference. 2002. Retrieved from http://www.cs.washington.edu/homes/alon/site/files/glue.pdf