DOCINER: A Document Indexation Tool for Learning Objects

Suphakit Niwattanakul  
School of Information Technology,  
Saranaree 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 TEXTCOMON (TEXTi-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 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](image)

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

![Fig. 2 The framework of keyword and concept extraction method](image)

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 SEWSE [24].
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**

<table>
<thead>
<tr>
<th>Objects (Concepts)</th>
<th>Attributes (Keywords)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>discrete structure</td>
</tr>
<tr>
<td>civil engineering</td>
<td>X</td>
</tr>
<tr>
<td>computer science</td>
<td>X</td>
</tr>
<tr>
<td>electrical engineering</td>
<td>X</td>
</tr>
</tbody>
</table>

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.

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 indexing. Using the notation “list of keywords -> concept”, here are the rules that can be derived from or that case Fig. 4 for that case.

\[
\text{"mathematics", "discrete mathematics", "discrete structure"} \rightarrow \text{"civil engineering"}\\
\text{"mathematics", "discrete mathematics", "computer programming"} \rightarrow \text{"computer science"}\\
\text{"mathematics", "computer programming", "electrical communication"} \rightarrow \text{"electrical engineering"}
\]

- **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.

\[
\text{"mathematics", "discrete mathematics"} \rightarrow? \text{"civil engineering"}\\
\text{"mathematics", "discrete mathematics"} \rightarrow? \text{"computer science"}\\
\text{"mathematics", "computer programming"} \rightarrow? \text{"computer science"}\\
\text{"mathematics", "computer programming"} \rightarrow? \text{"electrical engineering"}
\]

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].

\[
\text{precision} = \frac{\text{ard}}{\text{ad}} \quad (1)
\]

\[
\text{recall} = \frac{\text{ard}}{\text{trd}} \quad (2)
\]

\[
F - \text{measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (3)
\]

Where \( \text{ard} = \) number of relevant documents in the result list \( \text{ad} = \) number of documents in the result list \( \text{trd} = \) total number of relevant documents in the document base

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>
<thead>
<tr>
<th>Steps of evaluation</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Finding keywords</td>
<td>0.9933</td>
<td>0.9900</td>
<td>0.9861</td>
</tr>
<tr>
<td>(with indexing rules)</td>
<td>1.00</td>
<td>0.9900</td>
<td>0.9945</td>
</tr>
</tbody>
</table>

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


