Pedagogical Document Classification and Organization Using Domain Ontology

Ali Shariq Imran(✉) and Zenun Kastrati

Faculty of Computer Science and Media Technology,
Norwegian University of Science and Technology (NTNU), Gjøvik, Norway
{ali.imran,zenun.kastrati}@ntnu.no

Abstract. One of the challenges faced by today's web is the abundance of unstructured and unorganized information available on the Internet in form of educational documents, lecture notes, presentation slides, and multimedia recordings. Accessing and retrieving the massive amount of such resources are not an easy task, especially educational resources of pedagogical nature. Much of the pedagogical content available on Internet comes from blogs, wikis, posts with little or no metadata, that suffer from the same dilemma. The content is out there but way out of the reach of the intended audience. For content to be readily available, it has to be properly organized into different categories and structured into an appropriate format using metadata. This paper addresses this issue by proposing an automated approach using ontology-based document classification. The paper presents a case study and describes how our proposed ontology model can be used to classify educational documents into predefined categories.

Keywords: Domain ontology · Document classification · eLearning · SEMCON

1 Introduction

People have been educating themselves since the era of dawn, shaping their minds to adapt to the changing needs. Not only education has helped mankind acquaint themselves with better tools and skills, and to find solutions to everyday problems, it has also helped progress in the field of technology. Over time, it has resulted in a technologically advanced society we now live in today.

Massive amount of digital content from all walks of life is produced on a daily basis with the advancement of technology, and education is no different. Hundreds and thousands of educational videos, audio recordings, presentation slides and lecture notes are uploaded to the Internet, creating a massive wealth of information and digital libraries of educational content. Most of which is, however, unstructured and unorganized, thereby making it difficult to find them amongst the wealth of information available on the Internet.

According to IBM, a computer giant, roughly 2.5 quintillion bytes of data is produced every day. This data is coming from various sources in forms of emails,

P. Zaphiris and A. Ioannou (Eds.): LCT 2016, LNCS 9753, pp. 499–509, 2016.
DOI: 10.1007/978-3-319-39483-1_45

chats, blogs, posts, social media, eLearning platforms, among others. This huge amount of data is expected to grow even at a faster pace in coming years. More than 80 % of the information coming from various sources is unstructured and unorganized [1]. This results in a loss of data and the information fail to reach to the users. This holds true for educational material roaming around on the Internet as most of it never reach to the audience.

To ensure easy retrieval and access to massive amount of digital data, we need to organize and structure it accordingly. Having said that, organizing and structuring massive wealth of information is a momentous task for humans. It is labour intensive, prone to errors and is time-consuming. Automatic approaches and methodologies such as the use of ontologies can help play a vital role in this regard.

The rest of the paper is as follow. In Sect. 2 we present the related work on ontology and eLearning. Section 3 describes the importance and the role of ontologies in organizing pedagogical content. Section 4 presents a case study where we show how our proposed ontology model can play an effective role in organizing educational material while Sect. 5 concludes the paper.

2 Related Work

Ontologies are being used in web portals and eLearning systems for nearly two decades to generate knowledge and aid processes of collaborative learning. 1999 was the start of era when the role of ontologies for intelligent educational systems was first recognized in a workshop [2]. With the boom of web 2.0 in 2004, ontologies gained popularity. By then numerous workshops, conferences, and journals were dedicated to ontologies for educational systems [2–8], which resulted in a significant amount of researches. This lead to the development of semantic tools (Jena, Sesame, KAON, JRDF, Protege (Pugin for Protege for OWL) and ontology based languages (DAML + OIL, RDF, RDF-S, OWL and its sublanguages: OWL Lite, OWL DL, and OWL Full). Thus giving rise to a semantic web, semantic databases and semantic searches in last decade. Since then ontologies have become an essential part of many eLearning systems.

Ontologies are now being used in eLearning systems for domain knowledge, metadata, and entity representation. Use of ontologies in the context of eLearning can be categorized into [9]: (a) curriculum modeling and management, (b) describing learning domains, (c) describing learners data such as profile and personal data and, (d) describing eLearning services, all for the purpose of better content structuring and organization, and easy search and retrieval mechanisms. CURONTO is an ontology designed for entire curriculum management. Others include Gescur [10] and Crampon project [11] that aid curriculum designing. Learning ontologies consists of domain (subject) specific ontologies [12,13], and task-based ontologies involving pedagogy design [14,15], assessments [16,17], search and retrieval [18], and feedback [19,20]. Adaptive courseware Tutor [21], ONTODAPS [22], and work done by Panagiotopolos et al. [23] provides information about student's knowledge, progress, and personal information. They also

provide information describing learners data such as IMS learner information package (LIP), IEEE public and private information (PAPI), and friend of a friend (FOAF). While [24] and [25] are service related ontologies for creating learning object repositories (LOR) and mapping learning objects (LO) to a single common ontology. These facilitate the existing metadata standards such as Dublin Core DCMI, IMS learning resource metadata, IEEE LOM and SCORM via ontologies.

3 Importance of Ontologies in eLearning

This section briefly discusses eLearning platforms' content organization, defines ontology and establishes how ontology can play a role in the content organization in eLearning platforms.

3.1 eLearning Platforms

Over the years, numerous eLearning platforms and management systems have emerged. These platforms and websites such as Coursera, edX, Khan Academy, offers many online courses from all walks of life. These courses are usually divided into modules. Each module is further divided into different lessons and each lesson consists of a topic. The topics may be further split into smaller chunks of educational resources called LO. On a daily basis, hundreds and thousands of LOs are created and uploaded on various educational platforms. The benefit of these resources is certainly undeniable, however to benefit from the wealth of information available on the Internet, these resources have to be structured and grouped together into categories for easy search and retrieval. A domain ontology can play a vital role in this regard by incorporating semantics into eLearning platforms.

3.2 Ontology

Ontology is a fundamental element in semantic web and artificial intelligence (AI) based systems and is often defined as a '*specification of conceptualization*' [27]. It is a description of the real world concepts as entities and the relationship between them. Ontologies can be used in context of knowledge sharing and reuse. Given a domain ontology, queries, questions, and assertions can be made via AI agents/programs for content organization, structuring, and classification. Thus, It can also be described as a set of vocabulary of a particular domain.

A domain ontology consists of concepts and the relationships between these concepts for a particular domain (course) rather than specifying only generic concepts, as found in the upper ontologies such as SUMO, DOLCE, Cyc, among others. In other words, a domain ontology represents the vocabulary of a particular domain in a formal way and therefore it should closely match the level of information found in a text document in that domain.

3.3 Ontology Role in eLearning

Many learning management systems (LMS) and online learning platforms use open educational resources (OER) delivering high-quality pedagogical content in a form of LO. These OER and LO are often manually structured and organized into different categories, which demands a lot of manual work and is a time-consuming process. The OER and LO consist of different topics from various fields which can be depicted as concepts.

For instance, for a given chemistry domain, a list of commonly used terms can be prepared. These terms can be used as concepts to build an ontology for chemistry domain. The ontology is usually a hierarchical representation of these terms and the relationships between them. Thus, the terms for a chemistry domain can be represented as a hierarchical structuring consisting of classes and subclasses. To give an example, the term 'atom' can be a subclass of the term 'substance'. As both of them are concepts belonging to a chemistry domain, therefore, these terms can be used as labels in a domain ontology. Once the ontology is populated with a list of all the important concepts from the chemistry domain, it can be used to classify and organize different OER and LO using it.

In today's era, ontology plays a vital role in structuring and organizing pedagogical content on eLearning platforms. The next section presents a case study describing how ontologies can be used to structure and organize educational resources into different categories.

4 Case Study

In this section, we are introducing an example of classifying unlabelled text documents in the appropriate category within a pedagogical platform using the domain ontology. Employing a domain ontology enables to move from a document classification based on keywords to a classification based on content meanings (concepts), thus moving from lexical to semantic interpretation. The example presented in this section is composed of 4 components.

4.1 Domain Ontology

Text document classification presented in this case study is in line with ontology-based classification approach, therefore, it takes as a starting point the existence of a domain ontology. A domain ontology represents concepts for describing a domain and interpreting a description of a problem in that domain. A 5-tuple based structure [28] shown in Eq. 1 is commonly used to describe the concepts and their relationships of a particular domain.

$$D = (C,\ I,\ H,\ type,\ rel) \tag{1}$$

where:

- C is a finite set of concepts

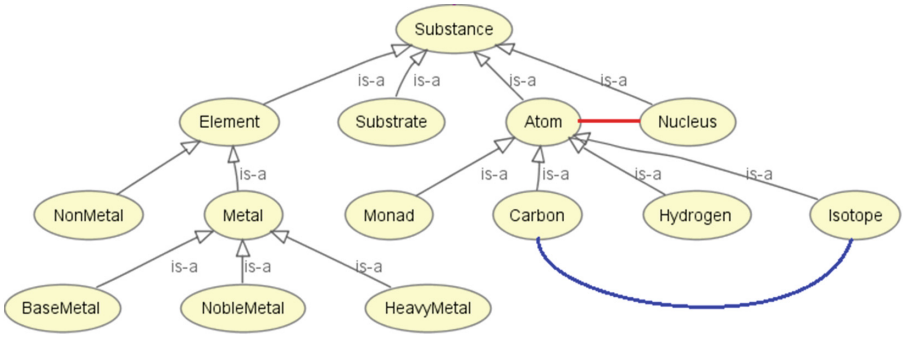

Fig. 1. A part of substance ontology from the domain of chemistry

- I is a finite set of lexical entries (Instances)
- H is a finite set of concept to concept relationships
- *type* is a finite set of instance to concept relationships
- *rel* is the finite set of instance to instance relationships.

Figure 1 shows a part of substance ontology built according to the Eq. 1. Additionally, it illustrates concepts (Element, Atom, Metal), instances of concepts (Carbon, Hydrogen) and the three types of relationships used to link these concepts. These relationships are (1) concept-to-concept (Metal is an element, Nucleus is part of an atom), (2) instance-to-concept (Monad is an atom with valence one), and (3) instance-to-instance relationship (Carbon has isotopes).

4.2 Predefined Categories and Semantics

The second component shows categories which have been predefined in a pedagogical platform. In this case study, the category shown in Fig. 2 is the subject of chemistry and the documents contained within this domain. The documents are organized into appropriate categories manually by an expert of that domain. At this point, these documents are represented as plain texts and there is no semantics associated with them.

The semantic information is added in using a domain ontology as defined in Subsect. 4.1. The semantic of documents, as shown in Fig. 3 (Doc_1, Doc_2,..., Doc_n), is incorporated by matching the terms t in a document Doc with the relevant concept c from the domain ontology. Adding semantics is possible thanks to (1) the presence of at least one of the concept labels within documents and/or (2) through identification of terms which are semantically close related to these concepts.

The former is a straightforward process. It simply employs the matching method [29] to find the concepts label within documents. A domain ontology consists of single label concepts such as Substance, Element, Atom and compound label concepts (BaseMetal, HeavyMetal) as well. For single label concepts, we use only those terms from the document for which an exact term exists in

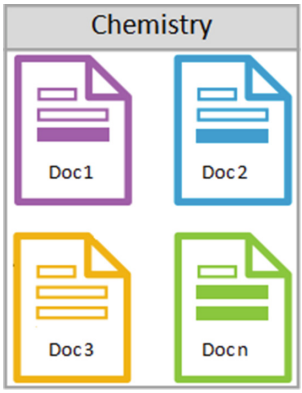

Fig. 2. Representation of a predefined category

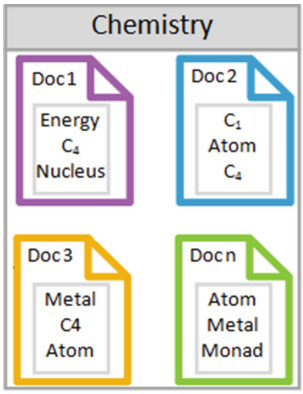

Fig. 3. Representation of documents after semantics have been incorporated

the domain ontology. For example, for concepts in the domain ontology such as Substance, Atom, and Element, there exists the same term extracted from the document. This process is known as exact term matching. For compound label concepts, we use those terms from the document which are present as part of a concept in the domain ontology. This type of concept matching is known as partial matching, and it represents cases when concept label contains terms extracted from the document in the corpus. The formal definition of exact and partial matches is given as follows.

If DO is the domain ontology, C the corpus composed of documents of this particular domain and $Doc \in C$ a document defined as a finite set of terms $Doc = \{t_1, t_2, ..., t_n\}$.

The mapping of term $t_i \in Doc$ into concept $c_j \in DO$ is defined as exact match $EM(t_i, c_j)$, where

$$EM(t_i, c_j) = \begin{cases} 1, & \textit{if label}(c_j) = t_i \\ 0, & \textit{if label}(c_j) \neq t_i \end{cases} \tag{2}$$

The mapping of term $t_i \in Doc$ into concept $c_j \in DO$ is defined as partial match $PM(t_i, c_j)$, where

$$PM(t_i, c_j) = \begin{cases} 1, & \textit{if label}(c_j) \textit{ contains } t_i \\ 0, & \textit{if label}(c_j) \textit{ not contain } t_i \end{cases} \tag{3}$$

If $EM(t_i, c_j) = 1$, term t_i and concept label c_j are identical and term t_i is then replaced with concept c_j.

If $PM(t_i, c_j) = 1$, term t_i is part of concept label c_j and term t_i is then replaced with concept c_j. For example, the *BaseMetal* compound ontology concept shown in Fig. 1 contains terms extracted from the document such as *Base* and/or *Metal*.

The latter is a more complex process. It searches for new terms within documents which are associated semantically with ontology concepts. To find these terms, we employ the SEMCON model [26] which uses an aggregated contextual and semantic information of the particular term. SEMCON exploits the statistical features such as frequency of occurrences of a term, term font type, and term font size to build the observation matrix. Contextual information is then defined by using the cosine measure where the dot product between two vectors of the observation matrix reflects the extent to which two terms have a similar occurrence pattern in the vector space. In addition to the context information, the SEMCON incorporates the semantic information to a term, by computing a semantic similarity score between two terms - a term that is extracted from a document and term that already exists in the domain ontology as a concept.

The next step is incorporating semantics of the categories. This is a process where the overall classification system tries to replicate the way an expert organizes/categorizes the documents into each category. The category semantics is built by aggregating the semantics of all documents which belong to the same category, as shown in Fig. 4.

Figure 5 shows all the steps taken through the process of incorporating the semantics of the chemistry category.

Each category is represented by a vector with two members: (1) concepts of a domain ontology, and (2) weight of these concepts. Category vector representation is given in Eq. 4.

$$Cat_j = \{(c_1, w_1), (c_2, w_2), (c_3, w_3), ..., (c_j, w_j)\} \tag{4}$$

Where, c_j is a concept appearing in the domain ontology and w_j is the weight of this particular concept.

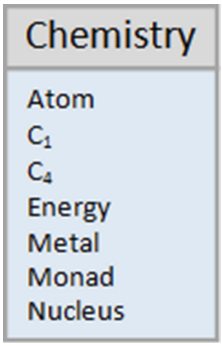

Fig. 4. Incorporating category semantics

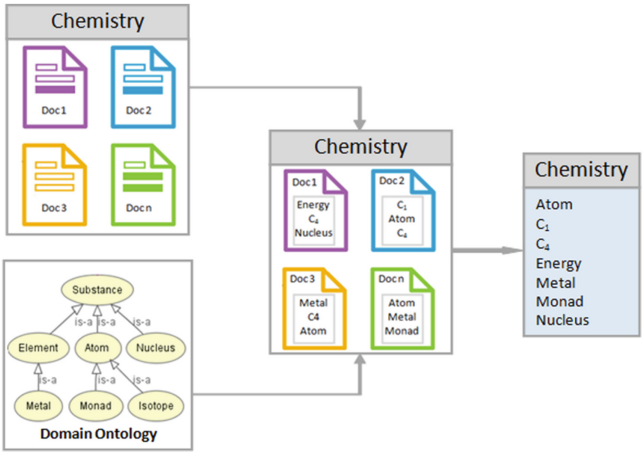

Fig. 5. The overall process of incorporating category semantics

Weight of concepts of the category vector is computed by aggregating the concept importance and concept relevance. The value of a concept weight given in Eq. 5 is in the range of [0,1].

$$w(c_j) = Imp(c_j) \times Rel(c_j) \tag{5}$$

Concept importance shows how important a concept is in the domain ontology and this is reflected in the number of relations this particular concept has to other concepts. The concept importance is computed automatically using the approach described in [30]. More concretely, the approach takes the ontology and map that into a graph and then implements one of the Markov-based algorithms (PageRank) to compute the concept importance.

Concept Relevance reflects the contribution of a particular concept to a category vector by the frequency of the occurrences of this concept in that category

alone and it is computed using the Eq. 6.

$$Rel(c_j) = \sum_{i=1}^{m} Freq(c_j) \tag{6}$$

Where, $Freq(c_j)$ is the frequency of occurrences of a concept c_j in the corpus.

4.3 Representation of Unclassified Documents

The last component deals with the unlabelled documents which have to be classified into the appropriate category. The following preprocessing steps have to be undertaken to bring these new and unclassified documents into the appropriate form for further processing: text is cleaned by removing all punctuation and capitalization, and a tokenizer is used to separate the text into individual terms (words); all terms resulted by tokenization process are passed through the term stemmer to convert them into their base or root form to develop a list of potential terms which are a noun, a verb, an adverb or an adjective; the stop words are removed; and finally the weight of terms is computed using one of the techniques from the Information Retrieval such as Term frequency *tf* or Term Frequency Inverse Document Frequency *tf*idf*.

After the preprocessing step, the incoming unlabelled document is finally represented as a document vector composed by a finite set of weighted terms and it is described by the tuple given in Eq. 7.

$$Doc_i = \{(t_1, w_1), (t_2, w_2), (t_3, w_3), ..., (t_i, w_i)\} \tag{7}$$

Where, t_i is the i^{th} term appearing in this particular document and w_i is the weight of this particular term.

The final task, after the unlabelled document is brought in the document vector form according to Eq. 7, is then to classify it into its appropriate category automatically. This ultimate goal is achieved using the similarity measure. It finds the similarity between category vector and document vector. The higher the similarity score, the closer the relationship between the document and the category. In other words, the higher the similarity score between a document and a category, the document more likely belongs to this category. The mathematical definition of similarity measure is given in Eq. 8.

$$Similarity(Doc_i, Cat_j) = \frac{\overrightarrow{Doc_i} \times \overrightarrow{Cat_j}}{\| \overrightarrow{Doc_i} \| \cdot \| \overrightarrow{Cat_j} \|} \tag{8}$$

Where, Doc_i and Cat_j represent the document vector and category vector, respectively.

5 Conclusion and Future Work

In this paper, we presented a case study depicting how the proposed ontology-based model can be used to classify educational documents into predefined

categories in a pedagogical platform. The model classifies documents based on the content meanings thereby trying to replicate the way an expert organizes/categorizes the documents into each category. To achieve this, the model initially build the semantics of the documents using the domain ontology. Aggregating the semantics of all these documents belonging to a particular category builds the semantics of category. Finally, an unlabelled document is classified into a category which has the highest similarity score with this particular document.

The proposed approach can be an ideal choice for educational platforms such as massive open online courses (MOOC) and LMS where content organization and structuring is inevitable for easy search and retrieval. In the future we are planning to implement the proposed model to classify documents for a particular course within a pedagogical platform.

References

1. Raghavan, P.: Extracting and exploiting structure in text search. In: SIGMOD Conference, p. 635 (2003)
2. Workshop on "Ontologies for Intelligent Educational Systems", in conjunction with AI-ED 1999, Le Mans, France, 18–19 July 1999
3. Workshop on "Concepts and Ontologies in Web-based Educational Systems", in conjunction with ICCE 2002, Auckland, New Zealand, 3–6 December 2002
4. Workshop on "Semantic Web for Web-based Learning", in conjunction with CAISE 2003, Klagenfurt/Velden, Austria, June 2003
5. Workshop on "Applications of Semantic Web Technologies for Web-based ITS", in conjunction with ITS 2004, Macei-Alagoas, Brazil, 30 August–03 September 2004
6. Workshop on "Learning Design and Topic Maps", Oslo, Norway, 26–27 January 2005
7. Anderson, T., Whitelock, D.M.: The educational semantic web: visioning and practicing the future of education. J. Interact. Media Educ. **2004**(1), p.Art. 1 (2004). http://doi.org/10.5334/2004-1
8. Sampson, D.G., Lytras, M.D., Wagner, G., Diaz, P.: Special issue on Ontologies and the Semantic Web for E-learning. J. Educ. Technol. Soc. **7**(4) (2004)
9. Al-Yahya, M., George, M., Alfaries, A.: Ontologies in e-learning: review of the literature. Int. J. Softw. Eng. Appl. **9**(2), 67–84 (2015)
10. Dexter, H., Davies, I.: An ontology-based curriculum knowledge-base for managing complexity and change. In: 9th IEEE International Conference on Advanced Learning Technologies, pp. 136–140 (2009)
11. Fernandez-Breis, J.T., Castellanos-Nieves, D., Hernandez-Franco, J., Soler-Segovia, C., Robles-Redondo, M., Gonzalez-Martinez, R., Prendes-Espinosa, M.P.: A semantic platform for the management of the educative curriculum. Expert Syst. Appl. **39**(5), 6011–6019 (2012)
12. Lee, M.-C., Ye, D., Wang, T.: Java learning object ontology. In: 5th IEEE International Conferences on Advanced Learning Technologies, pp. 538–542 (2005)
13. Sameh, A.: Ontology-based feedback e-Learning system for mobile computing. In: Mastorakis, N., Mladenov, V., Kontargyri, V.T. (eds.) Proceedings of the European Computing Conference. LNEE, vol. 27, pp. 479–488. Springer, New York (2009)
14. Isotani, S., Mizoguchi, R., Capeli, O., Isotani, N., Jaques, P.: A semantic web-based authoring tool to facilitate the planning of collaborative learning scenarios compliant with learning theories. Comput. Educ. **63**, 267–284 (2013)

15. Cobos, C., Rodriguez, O., Rivera, J., Betancourt, J., Mendoza, M., Leon, E., Viedma, E.: A hybrid system of pedagogical pattern recommendations based on singular value decomposition and variable data attributes. Inf. Process. Manag. **49**(3), 607–625 (2013)
16. Castellanos-Nieves, D., Fernandez, J., Garcia, R., Martinez, R., Moreno, M.: Semantic web technologies for supporting learning assessment. Inf. Sci. **181**(9), 1517–1537 (2011)
17. Litherland, K., Carmichael, P., Garcia, A.: Ontology-based assessment for accounting: outcomes of a pilot study and future prospects. J. Account. Educ. **31**(2), 162–176 (2013)
18. Lee, M., Tsai, K., Wang, T.: A practical ontology query expansion algorithm for semantic aware learning objects retrieval. Comput. Educ. **50**(4), 1240–1257 (2008)
19. Del, M., Breis, J., Castellanos, D., Morales, F., Espinosa, M.: Semantic web technologies for generating feedback in online assessment environments. Knowl. Based Syst. **33**, 152–165 (2012)
20. Kazi, H., Haddawy, P., Suebnukarn, S.: Leveraging a domain ontology to increase the quality of feedback in an intelligent tutoring system. In: Aleven, V., Kay, J., Mostow, J. (eds.) ITS 2010, Part I. LNCS, vol. 6094, pp. 75–84. Springer, Heidelberg (2010)
21. Grubii, A., Stankov, S., Žitko, B.: Adaptive courseware model for intelligent e-learning systems. In: International Conference on Computing, e-Learning and Emerging Technologies, vol. 16, no. 1 (2014)
22. Nganji, J., Brayshaw, M., Tompsett, B.: Ontology driven disability-aware e-learning personalisation with ONTODAPS. Campus-Wide Inf. Syst. **30**(1), 17–34 (2012)
23. Panagiotopoulos, I., Kalou, A., Pierrakeas, C., Kameas, A.: An ontology-based model for student representation in intelligent tutoring systems for distance learning. In: Iliadis, L., Maglogiannis, I., Papadopoulos, H. (eds.) Artificial Intelligence Applications and Innovations. IFIP AICT, vol. 381, pp. 296–305. Springer, Heidelberg (2012)
24. Raju, P., Ahmed, V.: Enabling technologies for developing next generation learning object repositories for construction. Autom. Constr. **22**, 247–257 (2012)
25. Arch-Int, N., Arch-Int, S.: Semantic ontology mapping for interoperability of learning resource systems using a rule-based reasoning approach. Expert Syst. Appl. **40**(18), 7428–7443 (2013)
26. Kastrati, Z., Imran, A.S., Yayilgan, S.Y.: SEMCON: semantic and contextual objective metric. In: 9th IEEE International Conference on Semantic Computing, pp. 65–68 (2015)
27. Gruber, T.R.: A translation approach to portable ontologies. Knowl. Acquisition **5**(2), 199–220 (1993)
28. Maedche, A.: Ontology Learning for the Semantic Web. Kluwer Academic Publishers, Norwell (2002)
29. Deng, S., Peng, H.: Document classification based on support vector machine using a concept vector model. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp. 473–476 (2006)
30. Kastrati, Z., Imran, A.S., Yayilgan, S.Y.: An improved concept vector space model for ontology based classification. In: 11th International Conference on Signal Image Technology and Internet Based Systems, Bangkok, Thailand (2015)