A General Framework for Text Document Classification Using SEMCON and ACVSR

Zenun Kastrati(✉), Ali Shariq Imran, and Sule Yildirim Yayilgan

Faculty of Computer Science and Media Technology,
Gjøvik University College, Gjøvik, Norway
{zenun.kastrati,ali.imran,sule.yayilgan}@hig.no

Abstract. The text document classification employs either text based approach or semantic based approach to index and retrieve text documents. The former uses keywords and therefore provides limited capabilities to capture and exploit the conceptualization involved in user information needs and content meanings. The latter aims to solve these limitations using content meanings, rather than keywords. More formally, the semantic based approach uses the domain ontology to exploit the content meanings of a particular domain. This approach however has some drawbacks. It lacks enrichment of ontology concepts with new lexical resources and evaluation of the importance indicated by weights of those concepts. Therefore to address these issues, this paper proposes a new ontology based text document classification framework. The proposed framework incorporates a newly developed objective metric called SEMCON to enrich the domain ontology with new concepts by combining contextual as well as semantic information of a term within a text document. The framework also introduces a new approach to automatically estimate the importance of ontology concepts which is indicated by the weights of these concepts, and to enhance the concept vector space model using automatically estimated weights.

Keywords: Ontology · Classification · Text document · SEMCON

1 Introduction

Web is the main source of information with large number of documents rapidly increasing every passing day. The information is usually kept in unstructured and semi-structured formats - be it text (e.g. word, pdf), images, video or audio. More than 80 % of the information produced by an organization is stored in unstructured textual format in the form of reports, email, views, news, etc [1]. Discovering and extracting useful information from these resources is therefore difficult without the organization and summarization of the document content. This is both an extremely vital and a tedious process in today's digital world [2]. Automatic classification in this respect plays a key role in organizing these massive sources of unstructured textual information into a structured format.Automatic text document classification (categorization) is the process of automatically assigning a

S. Yamamoto (Ed.): HIMI 2015, Part I, LNCS 9172, pp. 310–319, 2015.
DOI: 10.1007/978-3-319-20612-7_30

text document from a given domain to one or more class labels from a finite set of predefined categories.

Classification process has been tackled in two ways in literature: text based and semantic based. In the text based approach, the classification uses extraction of tokens and keywords thereby providing limited capabilities to capture and exploit the conceptualization involved in user information needs and content meanings. Aiming to solve these limitations, the semantic based approach follows the idea of using the content meaning rather than literal strings. It uses domain ontologies to exploit the content meanings of a particular domain.

Most of the indexing techniques used in the ontology based classification approach relies on the statistical vector space model, which represents text documents and categories as term vectors. The components of these term vectors are domain ontology concepts and their relevance represented by the frequency of concepts' occurrence.

Although the existing approaches use the domain ontology to index and retrieve text documents, they lack two important issues regarding the domain ontology (1) enrichment of ontology concept (2) importance of ontology concept.

Concept enrichment means linking new available lexical resources from a particular domain to the existing ontology concept and this is a crucial step for a domain ontology to be actually usable in real applications.

Importance of ontology concept defines the contribution of this concept in the classification process. This contribution depends on the position of concept where it is depicted in the ontology hierarchy, e.g. the higher the concept in the ontology hierarchy, the less the contribution in the classification and vice versa.

To address these issues, in this paper we propose a new ontology based text document classification framework. The proposed framework uses a new objective metric called SEMCON [9] developed at our laboratory to enrich the domain ontology with new concepts by combining contextual as well as semantic information of a term within a text document. In addition, the framework introduces a new approach to automatically estimate the importance of ontology concepts which is indicated by the weights of these concepts, and to enhance the concept vector space model using automatically estimated weights [10].

The reminder of this paper is organised as follows. Section 2 describes related work while Sect. 3 presents a detailed description of our proposed framework. In Sect. 4, we describe the process of labelling of unclassified document into appropriate category. Lastly, Sect. 5 makes some conclusions and gives some directions for future work.

2 Related Work

An increasing number of recent information retrieval systems make use of ontologies to help the users clarify and categorize their information needs and move towards semantic representations of documents. Many ontology based classification systems and models have been proposed in the last decade and all these follow the idea of considering the semantic relations between the terminology

information extracted from the text documents and the terminology information extracted from the domain ontology. An in depth review of ontology based classification approach is presented in [3,4]. The authors built an ontology in the economy domain for document classification. They indexed a corpus of economy related documents using the domain ontology by comparing the terminology extracted from the documents and the terminology extracted from the domain ontology.

The semantic indexing of documents using the domain ontology was also used later by researchers in [5]. They performed the semantic indexing and retrieval of biomedical documents through the process of identifying domain concepts extracted from the Medical Subject Headings (MeSH) thesaurus. The authors used a content-based cosine similarity measure. The semantic indexing of documents in domain of biomedicine was also subject of research in [6]. The authors established an ontology based information retrieval system (OBIRS) which uses a domain ontology and documents that are indexed using its ontology concepts e.g. genes annotated by concepts of the Gene Ontology or PubMed articles annotated using the Medical Subject Headings (MeSH). OBIRS system, through a user friendly interface, provides query formulation assistance through auto-completion and ontology browsing. The interface estimates the overall relevance of each document with respect to a given query. The relevance is obtained by aggregating the partial similarity measurements between each concept of the query and measurements indexing the documents. Finally the retrieved documents are ordered according to their overall scores, so that the most relevant documents are ranked higher than the least relevant ones. Furthermore, the scores are summarized in a small explanatory pictogram and an interactive semantic map is used to display top ranked documents.

In contrast to the above methods which use all concepts of a domain ontology to calculate similarity score between the ontology and documents, research in [7] presents a novel ontology-based automatic classification method which uses only a small number of ontology concepts. More precisely, the paper proposed using only the lowest level concepts (instances) of a ontology and this can be achieved thanks to the technique of ontology reasoning. In other words, the approach initially represents documents by a set of weighted terms and categories by ontologies. Then, ontology reasoning is performed to obtain the instances of an ontology. Finally, Google Distance is used to calculate similarity score between these instances and the set of weighted terms for each document. Documents are then classified into different categories according to the computed scores.

3 Proposed Model and Methodology

This section describes the proposed framework for automatic text document classification using the objective metric SEMCON and an adaptive concept vector space representation (ACVSR) model. The proposed framework, inspired by [12], is illustrated in Fig. 2. Our framework consists of four main components detailed below: (1) Domain Ontology, (2) Categories, (3) Building the semantics of the categories and (4) Unlabelled documents.

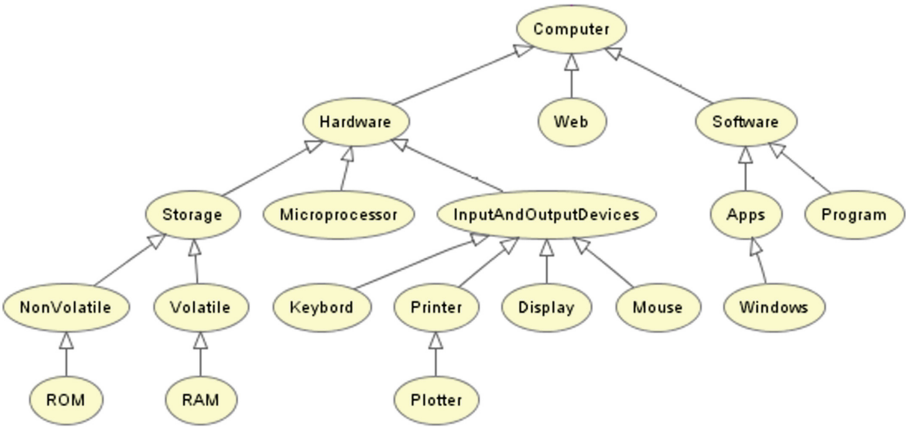

Fig. 1. Ontology sample of the computer domain

3.1 Domain Ontology

The work presented in this paper is in line with the ontology based approach and takes as a starting point the existence of a domain ontology. Domain ontologies are used to model in a formal way the basic vocabulary - concepts for describing a domain and interpreting a description of a problem in that domain.

A 5-tuple based structure [8] is a commonly used formal description to describe the concepts and their relationships in a domain. The 5-tuple core ontology structure is defined as:

$$O = (C,\ R,\ H,\ rel,\ A) \tag{1}$$

where:

- C is a non-empty set of concepts
- R is a set of relation types
- H is a set of taxonomy relation of C
- rel is a set relationship of C with relation type R, where $rel \subseteq C \times C$
- A is a set of description of logic sentences

Figure 1 shows an ontology tree specified for the concept computer, a part of computer domain.

The main purpose of introducing domain ontologies is to move from a document evaluation based on terms to an evaluation based on concepts, thus moving from lexical to semantic interpretation. The goal is to use the knowledge in domain ontologies to match categories and documents on a semantic level.

3.2 Categories

The second module represents the predefined categories *cat* such as *software*, *hardware*, *web* etc., and the documents *doc* within these categories *cat* for a

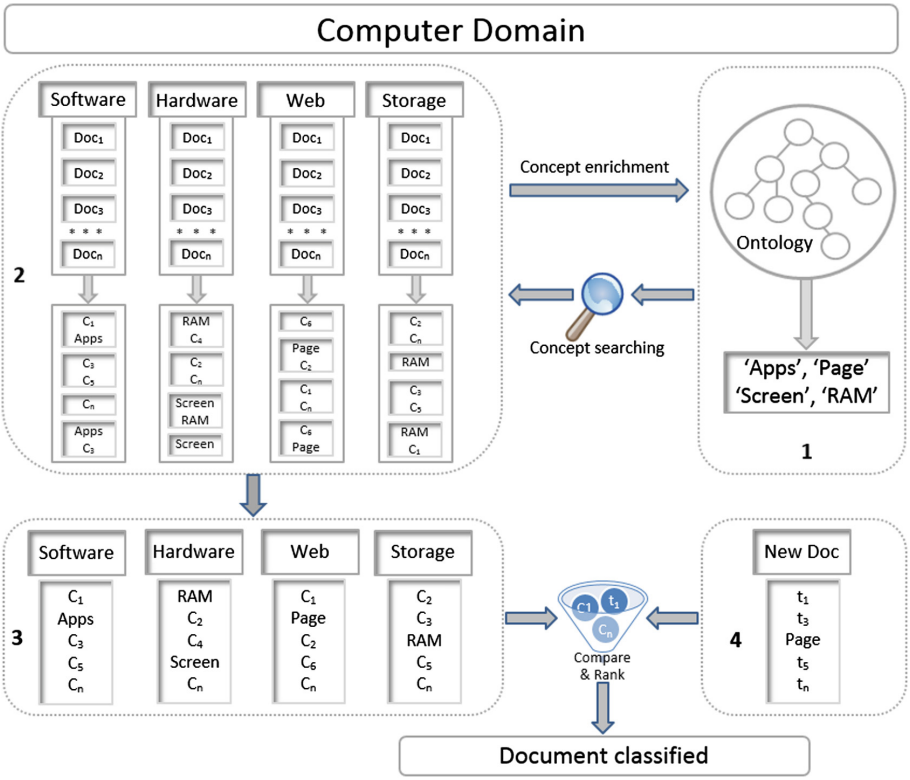

Fig. 2. The architecture of proposed model

certain domain (*Doc*). The documents are organized into appropriate categories manually by an expert from that domain. The documents are represented as plain texts without any semantics associated with them at this point. The semantic aspect is built in using the predefined domain ontology defined in 3.1. In other words, the semantic for each document (doc_1, doc_2,..., doc_n) is built by matching the terms t in the document doc with relevant concepts c in the domain ontology. This is achievable thanks to the presence/availability of at least one of concept labels within documents and/or through identification of associated terms.

The former is a straightforward process. There maybe single label concepts (*Windows*, *RAM*, etc.) in a domain ontology as well as compound label concepts (*InputAndOutputDevices*), as indicated in Fig. 1. For single label concepts, we use only those terms from the document for which an exact term exists in the domain ontology. For example, for concepts in the domain ontology such as *Windows*, *Printer*, *RAM*, etc., there exists exactly the same term extracted from the document. For compound label concepts, we use those terms from the document which are present as part of a concept in the domain ontology. For example,

consider *InputAndOutputDevices* as one of the compound ontology concept. In this case the ontology concept *InputAndOutputDevices* is matched if either term, *Input*, *Output* or *Device* is present in the document.

The latter is a more complex process. Rather than simply looking for an ontology concept, it looks for new terms within documents which are associated semantically with those concepts. To find these associated terms of a particular ontology concept within documents, the proposed framework employs the SEMCON model [9]. The SEMCON model is an objective metric which combines the contextual and semantic information of the given terms through its learning process. More formally, the SEMCON initially computes an observation matrix by exploiting the statistical features such as frequency of the occurrence of a term, term's font type and term's font size. The context 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 semantics by computing a semantic similarity score between two terms - term that is extracted from a document and term that already exists in the ontology as a concept.

3.3 Building the Semantics of the Categories

The third module deals with incorporating the semantics to the categories. By incorporating the semantics into categories, the overall classification system can 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, i.e. each category is represented as a vector whose components are concepts of a domain ontology. The relevance of a concept is defined by a weight computed based on the frequency of occurrence of the concept in a document.

The quantity of information given by the presence of concept c in a document depends on (1) the importance of concept c which is defined by the depth of c in the ontology graph and the number of concepts which subsume or are subsumed by c and (2) Relevance of concept c defined by the frequency of occurring of c in the document.

It is a well established fact that some concepts are better at discriminating between documents than others, in the classification process. In other words, concepts of a domain ontology do not contribute all equally. The contribution depends on the position of concepts where they are depicted in the ontology hierarchy and a concept's contribution is indicated by its corresponding weight. For example, if a concept is positioned in a lower level in the hierarchy tree this means that it is more abstract and it hardly belongs to other domain ontology. On the other side a more general concept is positioned in a higher level in hierarchy tree and it may exists in much ontology.

The ontology hierarchy consists of classes, subclasses and instances that may have different weights to reflect the concepts' importance. The concept's importance is calculated automatically using the model described in [10]. To show

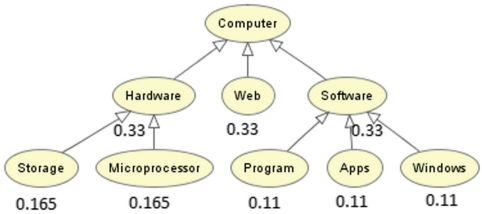

Fig. 3. Ontology representation for concept "Computer"

how the model computes the concept's importance, we have illustrated a simple example which can be explained starting from the lightweight ontology indicated in Fig. 3.

Let us consider concept *Program* which is a descendant of another concept *Software* which has m children including *Program*. Concept *Software* is a descendant of a concept *Computer* which has k children including *Software*. Concept *Program* is a leaf of the graph representing the computer domain ontology. For instance, considering a document containing only *Program* and *Software* concepts, the importance of concept *Program* in the document is $1 + \frac{1}{m}$. In the document containing *Program*, *Software*, *Computer* concepts, the importance of concept *Program* is $1+\frac{1}{m(1+\frac{1}{k})}$.

Aggregating the relevance and the importance of concept c, we obtain the overall weight given by the presence of concept c in a document and mathematically, it is formulated in Eq. 2.

$$w(c) = Freq(c) + \sum_{Path(c,\ldots,H)\in t_n} \sum_{m=2}^{depth(c)} \left(\frac{Freq(c_m)}{\prod_{k=2}^{m} |children(c_k)|}\right) \tag{2}$$

where *w(c)* is the overall weight of concept c and *Freq(c)* is the frequency of occurring of concept c in the document. The first part of sum given in Eq. 2 represents the relevance of concept c while the right part indicates the importance of concept c.

Finally, categories of a given domain are represented by the following tuple:

$$cat_i = \{(c_1, w_1), (c_2, w_2), (c_3, w_3), \ldots, (c_n, w_n)\} \tag{3}$$

where c_i is concept in domain ontology and w_i is its weight which is calculated using Eq. 2.

3.4 Unlabelled Documents

The last module encompasses a corpus of new documents which have to be classified. This module performs following preprocessing steps to represent and bring the new and unclassified documents into an appropriate form for further processing:

1. The text is cleaned by removing all punctuation and capitalization,
2. A tokenizer is used to separate the text into individual terms (words),
3. Passing all terms through the term stemmer to convert them in their base or root form to develop a list of potential terms which are a noun, a verb, an adverb or an adjective,
4. Remove all the stop words, words which do not contain important significance to be used,
5. Normalizing the texts using one of the techniques from the Information Retrieval such as Term frequency *tf* or Term Frequency Inverse Document Frequency $tf \times idf$.

From the list of potential terms, we extract only the nouns as part-of-speech (POS) of a language, since they represent the most meaningful terms in a document [11].

Finally, an unlabelled document to be classified is represented by a finite set of weighted terms as described by the following tuple:

$$doc_j = \{(t_1, w_1), (t_2, w_2), (t_3, w_3), \ldots, (t_n, w_n)\} \tag{4}$$

where t_i is term occurring in document and w_i is its weight which is calculated using the Term Frequency Inverse Document Frequency $tf \times idf$ model.

4 Assigning Documents to Categories

Consequently, the main goal of the above proposed model is to classify every new unclassified text document to its appropriate category automatically. Automatic text classification is the process of automatically assigning a text document from a given domain to one or more class labels from a finite set of predefined categories. For instance, for a binary text classification, it is the task of assigning a single binary value to each pair $(doc_j, cat_i) \in Doc \times Cat$, where Doc is a domain of documents and $Cat = (cat_1, cat_2, cat_3, cat_i)$ is a set of predefined categories. A threshold value T is assigned to (doc_j, cat_i) that denotes a decision to file doc_j under cat_i, on the contrary a value of F denotes a decision not to file doc_j under cat_i. The task is to approximate the unknown target function $\phi : Doc \times Cat \prec (T, F)$ by means of a function $\phi' : Doc \times Cat \prec (T, F)$ called the classifier, rule or model such that ϕ and ϕ' coincide as much as possible [13].

In our ontology-based classification framework, the assigning of an unlabelled document to a given category depends on the similarity score between them. The higher the score, the closer the relations between the document and the category. In other words, the document is more likely belongs to this category.

Once the category and document vectors given in Eq. 3 and in Eq. 4 are constructed, the similarity measure between a document doc_j and the category cat_i is computed as:

$$Similarity(doc_j, cat_i) = \frac{\overrightarrow{doc_j} \times \overrightarrow{cat_i}}{\| \overrightarrow{doc_j} \| \cdot \| \overrightarrow{cat_i} \|} \quad (5)$$

Finally, we will use a threshold λ to determine which category to assign the document. After doing the same treatment for all documents and ontologies by using above method, documents are assigned to their respective categories.

5 Conclusion and Future Work

In this paper we proposed a new framework for automatic text document classification. The proposed framework is ontology based utilizing concepts of a domain ontology rather than keywords, to capture and exploit the conceptualization involved in user information needs and content meanings.

The indexing technique used in this framework relies on the statistical vector space model, which represents text documents and categories as term vectors. The components of the term vectors are domain ontology concepts and their weights. The ontology concepts have to be supplemented with new lexical resources of this particular domain in order to capture better the content meanings. To achieve this, the paper introduced the SEMCON model which combines the semantic and contextual information.

The paper also proposed to use the importance of the concept in addition to the concept's relevance defined by the frequency of occurring. The importance is defined by the depth of concept in the ontology graph and the number of concepts which subsume or are subsumed by this concept. Thus, the overall weight of a concept in the term vectors is defined by aggregating the importance and the relevance of the concept.

In the future work, we plan to implement this framework in a real application domain and to evaluate and compare the performance of this approach with existing ontology based classification approaches.

References

1. Raghavan, P.: Extracting and exploiting structure in text search. In: SIGMOD Conference, pp. 635 (2003)
2. Al-Azmi, A.-A. R.: Data, Text, and web mining for business intelligence: a survey. Int. J. Data Min. Knowl. Manag. Process (IJDKP), 3(2) (2013)
3. Song, M.-H., Lim, S.Y., Kang, D.-J., Lee, S.-J.: Ontology-based automatic classification of web documents. In: Huang, D.-S., Li, K., Irwin, G.W. (eds.) ICIC 2006. LNCS (LNAI), vol. 4114, pp. 690–700. Springer, Heidelberg (2006)
4. Song, M., Lim, S., Kang, D., Lee, S.: Automatic classification of web pages based on the concept of domain ontology. In: Proceedings of the 12th Asia-Pacific Software Engineering Conference (2005)
5. Dinh, D., Tamine, L.: Biomedical concept extraction based on combining the content-based and word order similarities. In: Proceedings of the ACM Symposium on Applied Computing, SAC Q1, pp. 1159–1163, NY, USA (2011)

6. Sy, M.-F., Ranwez, S., Montmain, J., Regnault, A., Crampes, M., Ranwez, V.: User centered and ontology based information retrieval system for life sciences. BMC Bioinform. **13**(Suppl 1), S4 (2012). doi:10.1186/1471-2105-13-S1-S4
7. Fang, J., Guo, L., Niu, Y.: Documents classification by using ontology reasoning and similarity measure. In: Proceedings of the 7th International Conference on Fuzzy Systems and Knowledge Discovery (2010)
8. Maedche, A.: Ontology Learning for the Semantic Web. Kluwer Academic Publishers, Norwell (2002)
9. Kastrati, Z., Imran, A.S., Yayilgan, S.Y.: SEMCON: semantic and contextual objective metric. In: Proceedings of the 9th IEEE International Conference on Semantic Computing, Anaheim, California, USA (2015)
10. Kastrati, Z., Imran, A.S.: Adaptive concept vector space representation using Markov chain model. In: Janowicz, K., Schlobach, S., Lambrix, P., Hyvönen, E. (eds.) EKAW 2014. LNCS, vol. 8876, pp. 203–208. Springer, Heidelberg (2014)
11. Liu, J.N.K., He, Y.-L., Lim, E.H.Y., Wang, X.-Z.: A new method for knowledge and information management domain ontology graph model. IEEE Trans. Syst. Man Cybern. Syst. **43**, 115–127 (2013)
12. Calvier, F.-E., Plantié, M., Dray, G., Ranwez, S.: Ontology Based Machine Learning for Semantic Multiclass Classification. In: TOTH- Terminologie & Ontologie: Théories et Applications, Chambéry, France (2013)
13. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. (CSUR) **34**(1), 1–47 (2002)