Improving Classification of Multi-Lingual Web Documents using Domain Ontologies

Marina Litvak, Mark Last, and Slava Kisilevich

Department of Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel,
{litvakm, mlast, slaks}@bgu.ac.il

Abstract. In this paper, we deal with the problem of analyzing and classifying web documents to several major categories/classes in a given domain using domain ontology. We present the ontology-based web content mining methodology that contains such main stages as collecting a training set of labeled documents from a given domain, building a classification model above this domain given the domain ontology, and classification of new documents via the induced model. We tested the proposed methodology in a specific domain, namely web pages containing information about production of certain chemicals. Using our methodology, we are interested to identify all relevant web documents while ignoring the documents that do not contain any relevant information. Our system receives as input an OWL file built in Protege tool, which contains the domain-specific ontology, and a set of web documents classified by a human expert as "relevant" or "non-relevant". We use a language-independent key-phrase extractor with integrated ontology parser (defined in a given language) for creating the database from input documents and use it as a training set for the classification algorithm. The system classification accuracy using various levels of ontology is evaluated.The current version of our system supports web content mining in English, Arabic, Russian, and Hebrew languages.

1 Introduction

Over the last years, we have observed an explosive growth in the information available on the Web. To meet our information needs, we need more intelligent systems to gather the useful information from the huge amount of Web related data sources.

Web mining ([2]) is a new technology that has emerged as a popular area in the field of Web Intelligence ([4]). It is categorized into three areas of interest: web usage mining (finds access patterns from web logs), web structure mining (provides structural information about documents) and web content mining (finds useful information from the web content) [1]. It is obvious that data mining techniques (see [5], [6]) can be used for Web mining. One of the problems in this area is to represent the web documents as a meaningful, informative input for data mining algorithms, and then to "translate"/interpret the mining results.

In this paper, we introduce the ontology-based web content mining application for analyzing and classifying web documents in a given domain. We use *domain ontology*, which organizes concepts, relations and instances into a domain [11], for purpose of enriching the term vectors representing documents with concepts. This approach has two benefits: first, it resolves synonyms; and second, it introduces more general concepts. Our term vectors contain of terms and their importance weights, where term may be a phrase extracted from the

text of a document or related concept from the ontology (depending on the level of concept hierarchy or abstraction level induced by the user). For the purpose of classification, we can use any popular classification algorithm, like C4.5, Bayes Network and Naive Bayes.

The rest of the paper is organized in the following way. Section 2 summarizes the related work. Section 3 describes the methodology and the proposed system. Section 4 depicts the tested domain and the constructed ontology. In Section 5, we evaluate the results of initial experiments. Finally, in the last section we outline the conclusions and the future work.

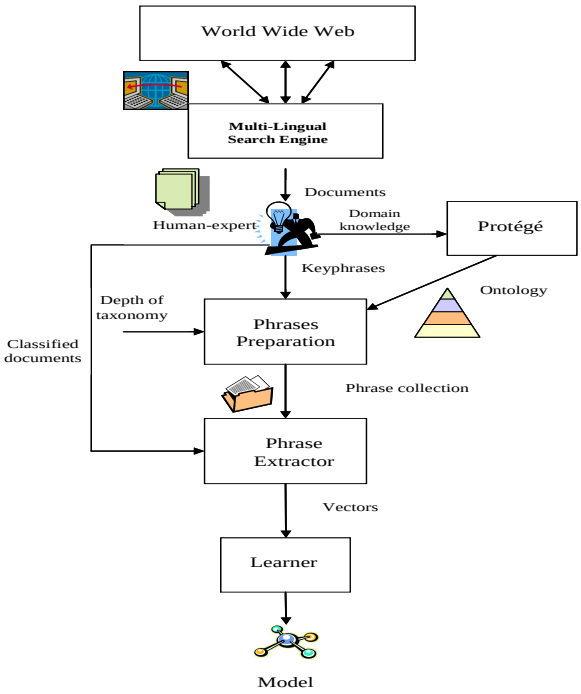


Fig. 1. Cross-Lingual Web Classification System

2 Related work

During the last decade, a huge amount of issues related to web content mining was investigated, like discovering of different patterns in the static content using conventional data mining [3], dynamic content mining (like mining news from online news sites) [7], predicting web information content [8], developing recommendation systems that can suggest the "information content" (IC) pages [9],

classifying web documents into Web hierarchy or topic ontology [20],[21], and many other.

Many authors reduce building recommendation systems to the classification task. Billsus and Pazzani [12] trained a Naive Bayes classifier [13] to recommend news stories to a user, using a Boolean feature vector representation of the candidate articles, where each feature indicates the presence or absence of a word in the article. Jennings and Higuchi [14] trained one neural network for each user to represent a user's preferences for news articles. Anderson and Horvitz [15] built a Naive Bayes model to predict the candidates (pages or topics) that the user will view next in the session.

Document representations for test classification are typically based on the classical Bag-Of-Words paradigm. However, over last years, the authors tried to enhance the classical document representation through concept-based document retrieval ([26]). One of such enhanced approaches is ontology[1].

Currently, there are several existing approaches for classifying web pages into Web hierarchy. Koller and Sahami in [20] propose an approach that utilizes the hierarchical topic structure to decompose the classification task into a set of simpler problems, one at each node in the classification tree. Mladenic and Globelnik in [21] describe an approach to automatically mapping web pages onto ontology using the Yahoo! ontology of Web pages. The paper of McCallum *et. al* ([22]) shows that the accuracy of a Naive Bayes text classifier can be significantly improved by taking advantage of a hierarchy of topic categories of documents. Chakrabarti *et. al* in [23] explore how to organize large text databases hierarchically by topic to aid better searching, browsing and filtering.

Bloehdorn and Hotho in [25] propose document representation through concepts extracted from background knowledge. In another publication ([24]) Hotho *et. al* use ontologies to improve text document clustering. A paper by Cesarano *et. al* [16] presents a prototype of an ontology-based system for information retrieval on the web, where the global relevance grade is computed for each document.

3 Methodology

Figure 1 presents a high-level view of the proposed Cross-Lingual Web Classification System. In the absence of any detailed domain knowledge, a user can initiate the system operation by submitting a set of keyword queries in any language to a multi-lingual search engine (such as Google^{TM}). A human expert reads the documents and labels them as "relevant" or "irrelevant". Additional degrees of relevancy (e.g., "partially relevant") can be allowed. The task of the Learner module is to build a compact model (profile) of the pages collected

[1] According to the most cited definition in the literature [10], ontology is an explicit specification of a domain conceptualization. It accumulates and organizes knowledge about domain in a machine-processable and human-readable way providing a common understanding basis, facilitating information/knowledge dissemination and reuse. Therefore, ontology has the potential to improve information/knowledge capturing, organization, re-use and re-finding through meticulous domain organization principles and advanced reasoning tasks.

from the web so that new relevant pages can be reliably recognized by the system. We induce a classification model from a training collection that includes a mix of relevant and non-relevant pages. Each page is represented as a vector of $< term_i,\ weight_i >$ pairs received from Ontology-based Phrase Extractor module, described in the sub-section below. The phrases are extracted using a list of domain-specific terms and other ontology information. The term-frequency (tf) $weight_i$ indicates the frequency of a $term_i$ in the observed document.

3.1 Ontology Specification

An ontology defines explicitly the terms used to describe and represent an area of knowledge. Ontologies are used by people, databases, and applications that need to share domain information. Ontologies include computerusable definitions of basic concepts in the domain and the relationships among them. They encode knowledge in a domain and also knowledge that spans domains [19]. The term 'ontology' can be used for several ways. Ontologies can contain simple taxonomies and logical theories as well.

In this paper, an ontology represents the conceptual information of the domain of interest (see Section 4) and it is used for the purpose of conceptual document representation and improving the documents classification. In other words, our goal is extraction of more meaningful and relevant (even general) information from text of documents for the purpose of building more accurate classification models. Our ontology consists of individuals/instances, classes with their properties and hierarchical/taxonomic relationships between them. Each object/thing in the domain is associated with its unique class. Usually the names of classes are nouns. Each thing has a name (the name itself is not object of the domain but only symbolizes it) or several names that are synonyms. All names of an object are mapped to ontology as individuals of its class. The relationships among the things represent the existing taxonomy. The properties describe the things.

3.2 Ontology-based Phrase Extractor

This module includes Phrase Preparation and Phrase Extractor units (see Fig. 1). The module receives as input documents, ontology and abstraction level and creates term vectors.

The Phrase Preparation unit prepares phrase collection given ontology and abstraction level k — XML file including all general thing names as phrases with their associated classes of k^{th} level as related concepts (in case of abstraction level equal to 0 the collection does not include related concepts). Currently, we also add to this collection phrases that, by expert opinion, can characterize type of a document. In the future we are going to build a separate ontology containing these phrases or even embed them into an existing domain ontology.

The Phrase Extractor scans the phrases included in the collection, and every time it finds name of thing it references to the related concept. We used **Replace Terms by Concepts ("repl")** strategy (HYPINT) for replacing terms by

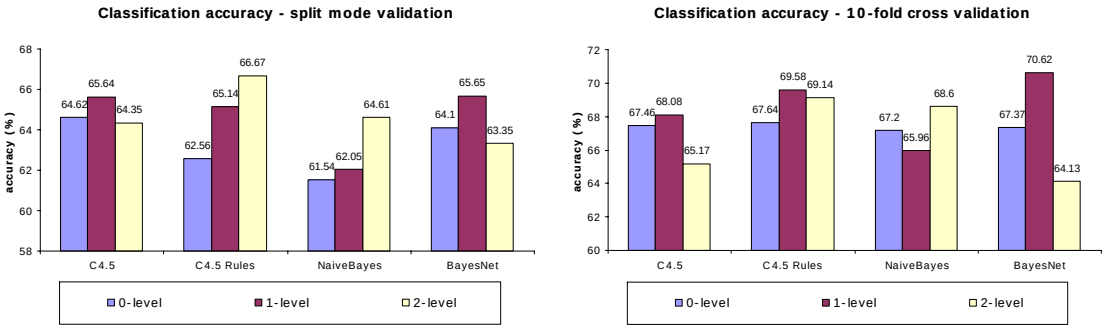
concepts and **All Concepts ("all")** strategy for disambiguation investigated in [24]. **Replace Terms by Concepts (repl)** strategy expels all terms from the vector representations for which at least one corresponding concept exists. Thus, terms that symbolize general things in domain ontology are only considered at the concept level, but terms that do not appear in ontology (provided directly by a human expert) are not discarded. The **All Concepts (all)** does not do anything about disambiguation and considers all concepts for augmenting the text document representation. The concept frequency is calculated as sum of the frequencies of all terms in document being related to that concept in the ontology.

The generic structure of this module enables to handle texts in virtually any language.

4 Experiments

The main goal of this research is increasing the classification accuracy through maintaining an ontology. We tested the proposed methodology in a specific domain, namely web pages containing information about production of certain chemicals. It is clear, that almost every chemical has many names (synonyms) - it may be a full name, an abbreviation, a formula or molecular structure. Our ontology stores class for each chemical in domain that contains all its known names as instances. Whenever the Phrase Extractor finds any name of chemical, it refers to the associated class. In addition, we define different properties for the chemicals and keep the hierarchical relationships between groups of them, like "available chemicals" (can be purchased or extracted from something), "rare chemicals" (complement to the first one), organic chemicals, salt, poisons and more. These groups may be joint as well as disjoint. The total time spent for ontology creation was about 20 hours including 2-3 meetings with a domain expert. Currently, our ontology includes 29 instances (things/names of chemicals) organized into 37 classes. We wish to extend it in the future experiments.

We learned and tested four classification models on the following document representations: vectors of original phrases (without any knowledge about concepts and relationships between them kept in the domain ontology), the same documents after phrase extraction with synonyms handling (1-level conceptualization), and after 2-level conceptualization (referring extractor every time it finds name of chemical to the parent classes of its thing class), and then compared between the accuracy rates of the resulting models. We were given 114 HTML pages classified as relevant and non-relevant by a domain expert (41 pages or 36% are relevant). Charts in Fig.2 demonstrate the classification accuracy of different models depending on level of ontology conceptualization. We applied C4.5, C4.5 Rules, Bayes Network and Naive Bayes algorithms using Weka Data Mining Software [18] and two testing modes: 10-fold cross-validation and test split (66% training set and remainder the testing set). The values in Fig.2 are averages from 10 runs of each mode. We used t-test (two-tailed paired, $\alpha = 0.05$) for each algorithm to assess whether the accuracy values of different abstraction



(a) (b)

Fig. 2. Classification Accuracy depending on Abstraction Level

levels	C4.5	C4.5 Rules	Naive Bayes	Bayes Network
0 – 1	↑ 0.037*	↑ 0.032*	↑ 0.738	↑ 0.336
1 – 2	↓ 0.340	↑ 0.455	↑ 0.053	↓ 0.124
0 – 2	↓ 0.832	↑ 0.104	↑ 0.164	↓ 0.480

levels	C4.5	C4.5 Rules	Naive Bayes	Bayes Network
0 – 1	↑ 0.411	↑ 0.039*	↓ 0.044*	↑ 0.001*
1 – 2	↓ 0.007*	↓ 0.762	↑ 0.001*	↓ 0.000*
0 – 2	↓ 0.079	↑ 0.313	↑ 0.076	↓ 0.006*

Table 1. Results of t-test — split mode (left table) and 10-fold cross validation (right table)

levels are statistically different from each other. The results of the t-test are presented in Table 1.

As it can be seen from the results of the experiments, the C4.5 and C4.5 Rules based on the split mode and C4.5 Rules and Bayes Network based on the 10-fold cross validation are significantly improved in 1-level abstraction with respect to the 0-level. On the other hand, the accuracy value of the Naive Bayes is decreased.

When we classified the 2-level abstraction represented documents, the Naive Bayes model based on the 10-fold cross validation has improved, while accuracy of the C4.5 and Bayes Network models have decreased.

Algorithms performances at 0-level abstraction with respect to 2-level are not significantly different, except decrease of accuracy of the Bayes Network model.

We explain the accuracy decreasing of most models in case of 2-level abstraction by losing some specific information in more general representation of documents. The "strange" behaviour of Naive Bayes model the in 10-fold cross validation mode, by our opinion, is justified by its specific constraints: first, it confirms independence of variables, and, second, it builds model based on all available features, while decision tree is using a feature selection procedure.

As we all know, the size of the training set affects the classification model accuracy. We believe that given a larger training set (currently in preparation) we can get more accurate results.

5 Conclusions and Future Work

In this paper we presented a new ontology-based methodology for classification of web documents to main categories according to the user "Information Needs". The main contribution of this work is using domain-based Multi-Lingual Ontology in the conceptual representation of documents. We tested our method on the specific chemicals domain, where the synonyms and the taxonomic relationships were handled. Despite the small training set, quite good results were obtained. We intend to improve current results by increasing the training set and the set of keyphrases as well as by enhancing our methodology in the following ways:

- Learning a multi-lingual domain ontology exploiting machine learning techniques.
- Elaborating (or use some existing tools like GATE [17]) for automatic construction of ontologies on specific domain. Such update will enable us to make an ontology-based classification system completely domain-independent.
- Using several ontologies for the same set of documents (or one ontology including several hierarchies).
- Mapping web documents into Web hierarchy (it may be topic ontology) to improve the classification accuracy.

Acknowledgement. We wish to thank D. Berenstein, the domain expert, for helping us in the ontology construction and collection of the training set for the learning algorithms.

References

1. R. Kosala and H. Blockeel. Web mining research: a survey. SIG KDD Explorations, Vol. 2, pp. 1-15, July 2000.
2. O. Etzioni. The World Wide Web: Quagmire or Gold Mine? Communications of the ACM, Vol. 39, No. 11, pp. 65-68. Nov. 1996.
3. M. Montes-y-Gomez, A. Gelbukh and A. Lopez-Lopez. Mining the News: Trends, Associations, and Deviations. Computacion y Sistemas, Vol. 5 No. 1, pp. 14-24, Julio-Septiembre 2001.
4. Y.Y. Yao, N. Zhong, J. Liu and S. Ohsuga, Web Intelligence (WI): research challenges and trends in the new information age, in Zhong et al., eds., Web Intelligence: research and development, LNAI 2198, Springer-Verlag, pp. 1-17, 2001.
5. R. Agrawal, T. Imielinski and A. Swami, Database mining: a performance perspective, IEEE Transactions on Knowledge and Data Engineering, Vol. 5, no. 6, pp. 914-925, 1993.
6. M. S. Chen, J. Han, and P. S. Yu, Data mining: an overview from a database perspective, IEEE Transactions on Knowledge and Data Engineering, Vol. 8(, no. 6, pp. 866-883, 1996.
7. A. Mendez-Torreblanca, M. Montes-y-Gomez and A. Lopez-Lopez. A Trend Discovery System for Dynamic Web Content Mining, [citeseer.ist.psu.edu/695212.html], 2002.
8. Tingshao Zhu, Russ Greiner and Gerald Houbl. Predicting Web Information Content. Workshop on Intelligent Techniques for Web Personalization (ITWP '03), 2003.

9. Tingshao Zhu, Russ Greiner, and Gerald Haubl. An effective complete-web recommender system. In The Twelfth International World Wide Web Conference(WWW2003), Budapest, Hungary, May 2003.
10. T. R. Gruber. A translation approach to portable ontologies. Knowledge Acquisition, Vol. 5, no. 2, pp. 199-220, 1993.
11. G. van Heijst, A.Th. Schreiber, and B.J. Wielinga. Using explicit ontologies in KBs development. IJHCS, pp. 183-291, 1997.
12. D. Billsus and M. Pazzani. A hybrid user model for news story classification. Proc. of the Seventh International Conference on User Modeling (UM '99), Banff, Canada, 1999.
13. Richard Duda and Peter Hart. Pattern Classification and Scene Analysis. Wiley, New York, 1973.
14. Andrew Jennings and Hideyuki Higuchi. A user model neural network for a personal news service. User Modeling and User-Adapted Interaction, Vol. 3, no. 1, pp. 1-25, 1993.
15. Corin R. Anderson and Eric Horvitz. Web montage: A dynamic personalized start page. Proc. of the 11th World Wide Web Conference, 2002.
16. C. Cesarano, A. d'Acerno, A. Picariello. An Intelligent Search Agent System for Semantic Information Retrieval on the Internet. Proc. of the Fifth ACM International Workshop on the Web Information and Data Management, November 7-8, 2003, New Orleans, Louisiana, USA, pp. 111-117, 2003.
17. GATE - General Architecture for Text Engineering, The Natural Language Processing Research Group, Department of Computer Science, University of Sheffield [http://gate.ac.uk/].
18. Weka - Data Mining Software in Java [http://www.cs.waikato.ac.nz/ml/weka/].
19. Li, Yuefeng and Zhong, Ning. Web Mining Model and Its Applications for Information Gathering. Knowledge-Based Systems, Vol. 17, no. 5-6, pp. 207-217, 2004.
20. Koller, D., Sahami, M. Hierarhically classifying documents using very few words. Proc. of ICML 1997.
21. Mladenic, D., Grobelnik, M. Mapping documents onto web page ontology. Web mining: from web to semantic web: EWMF 2003, Springer Lecture Notes 2004.
22. McCallum, A. et al. Improving text classification by shrinkage in a hierarchy of classes. Proc. of ICML 1998.
23. Chakrabarti, S., Dom, B., Agrawal, R., Raghavan, P. Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. The VLDB Journal (1998), Spinger-Verlag 1998.
24. Andreas Hotho, Steffen Staab, Gerd Stumme: Ontologies Improve Text Document Clustering. ICDM 2003.
25. Stephan Bloehdorn, Andreas Hotho. Text classification by boosting weak learners based on terms and concepts. Proc. of the Fourth IEEE International Conference on Data Mining, 331-334. IEEE Computer Society Press, NOV 2004.
26. Information Mapping Project. [http://infomap.stanford.edu/index.html#papers]