

Clustering for Ontology Evolution

George Tsatsaronis, Reetta Pitkänen, and Michalis Vazirgiannis

Department of Informatics,
Athens University of Economics and Business,
76, Patission street, Athens 104-34, Greece

Abstract. The Semantic Web initiative aims at automating semantics' embedding in Web pages so that richer information retrieval, data integration and improved navigation can be supported. Domain ontologies are used in this direction, providing a way to semantically characterizing Web documents if a mapping of the documents to the ontology concepts can be managed. Given such an ontology, the main problem that arises due to the rapid changes of the Web resources is monitoring the domain changes and update the given ontology respectively. In this paper we propose a semi-automated method for ontology evolution using documents' clustering, from the results of which ontology enrichments and updates are extracted.

1 Introduction

Over the years a large number of research issues have been raised concerning the existence and the use of the ontologies in the Semantic Web. Primarily, a vast amount of research has been committed on how to represent knowledge using ontologies, and on the languages that can be used for defining and handling them. Furthermore, research issues are brought forward in Maedche et. al. (2003) concerning the use of ontologies for Enterprise Knowledge Management with emphasis on the e-commerce. Attempts have been made so that an integrated enterprise-knowledge management architecture for implementing an ontology-based knowledge management system is created.

In this paper, we are tackling the issue of semi-automated ontology evolution and propose a method that utilizes document clustering for this task. Our notion of an ontology complies with the ontology model introduced in Motik et. al. (2002). The method capitalizes on the THESUS system (Nguyen et. al. (2003)) that enables organization of Web documents into semantic clusters. Our goal is to automatically suggest for a given domain ontology four types of changes to the user: new concept insertion, new instance insertion, concept merging and concepts declining. Concisely, the contribution of the work presented in this paper is two-fold: 1) A new semi-automated method for ontology evolution is proposed based on clustering a well defined subset of the WWW. The innovation of the method lies in mining the constantly changing Web resources with a clustering algorithm to update the conceptualization of a domain. 2) The underlying method mechanisms utilize a hierarchical thesaurus to map Web documents to ontology concepts and constitute an innovative approach for semantically characterizing Web documents with a given ontology's concepts.

The remaining of this paper is organized as follows: Section 2 introduces the notion of some preliminary concepts, like ontology evolution, THESUS system and thesauri. Section 3 introduces our evolution approach and formally defines the four types of the method's proposed ontology changes. Section 4 describes an experimental set up created to evaluate our method's performance. Section 5 musters the related work in the field of ontology evolution. Section 6 concludes and sets a potential basis for the continuation of this work.

2 Preliminary concepts

Since the proposed semi-automated ontology evolution method capitalizes on the THESUS system, this section gives a brief introduction on THESUS, as well as to the process of ontology evolution. Furthermore, the role and use of a thesaurus in our method is elucidated and its relation with the ontology evolution process is explained.

2.1 The problem of ontology evolution

In this paper we shall adopt the definition of an ontology evolution process given by Maedche et.al (2003), according to which ontology evolution is "the timely adaptation of an ontology and consistent propagation of changes to the dependent artefacts". The problem of ontology evolution is partitioned into single ontology evolution and multiple ontology evolution (Maedche et. al. (2003)). In this work we tackle with the problem of single ontology evolution, and consider this ontology residing in a single node.

Changes in an ontology evolution process are separated into *elementary* and *composite* (Stojanovic et. al. (2002)). The afflicted parts of ontology changes are the ontology concepts, the ontology instances and the IsA relations. Some of the elementary changes, as described in Stojanovic et. al. (2002), that can be applied in an ontology are the addition of a new concept/instance/IsA relation, the deletion of an existing concept/instance/IsA relation and the modification of an IsA relation. In this work we deal with the addition of new concepts and instances and the deletion of existing ones. A proposed placement of the new concepts/instances in the ontology is suggested to the user.

Combinations of the aforementioned changes can lead to composite changes, as described in Stojanovic et. al. (2002), some of which are the merge of existing concepts, the split of existing concepts into several concepts and the extraction of a superconcept from existing concepts. The proposed method deals with the merging of existing concepts.

2.2 THESUS system

The Thesus system (Nguyen et. al. (2003)) organizes Web documents into semantic clusters, exploiting the semantics of both document content and links and utilizing two different clustering algorithms, DB-Scan and COBWEB (see Nguyen et.

al. (2003) and references therein). THESUS comprises of five basic modules: 1) the *information acquisition* module, that gathers a set of URLs which appear relevant to the topic under consideration, 2) the *information extraction* module, that extracts keywords from Web documents, 3) the *information enhancement* module that enhances extracted hyperlink information with semantic information by mapping extracted keyword sets to sets of concepts in an ontology, 4) the *clustering* module, that partitions the set of URLs into semantically coherent subsets based either on the extracted keywords or on the respective ontology concepts describing the topic under consideration, and 5) the *query engine* that enables searching in the collection. Our work capitalizes on the THESUS system, and thus the existence of an initial domain ontology and a thesaurus, the later of which will be used for the mapping of documents to domain ontology concepts, is taken for granted.

2.3 Thesauri, concepts, instances and IsA relations

The THESUS architecture makes use of a hierarchical thesaurus in order to find the best possible mapping, according to an extension of the Wu and Palmer distance measure (Wu and Palmer (1994)), between a set of keywords describing a Web document and a set of concepts of the given domain ontology. The hierarchical thesaurus used in the case of THESUS is Wordnet. So, the role of Wordnet in THESUS is to act as the graph in which both the domain ontology concepts and Web document keywords will be mapped, so that a distance between such two sets can be found, based on their mappings distances in Wordnet. The task of characterizing a Web document with concepts from a given domain ontology is thus reduced to finding the mapping of ontology concepts and document keywords to Wordnet senses (a word sense disambiguation task), such that the extended Wu and Palmer measure described in Nuygen (2003) gives the minimum distance between the two sets of senses.

The hierarchical thesaurus plays another significant role in the proposed ontology evolution method. It is used as a guide to discriminate between potential new ontology concepts and potential new ontology instances. Though ontology concepts and instances share the same meaning as in the ontology model proposed in Motik et. al. (2002), we have further extended the nature of both, with regards to their existence or non-existence in a given hierarchical thesaurus. We thus consider ontology concepts as ontology nodes pertaining to the domain described by the ontology, and exist in a hierarchical thesaurus (like WordNet). The ontology instances are considered to be words that relate to a specific concept and do not exist in a hierarchical thesaurus (i.e. company names, places, etc.). Finally the ontology IsA relations describe the parent - children relations of the concepts existing in the ontology.

3 Ontology evolution method

The method commences with an initial set of several URLs relating to the user's domain of interest. Furthermore, an initial domain ontology is needed as input. A

crawler then fetches a large set of Web documents, commencing the search from the initial URLs, utilizing the incoming and outgoing hyperlinks of the documents. The documents are then characterized with representative keywords arising from their content. The keywords and the ontology concepts are mapped to Wordnet. A mapping of the Web documents to the ontology concepts is then made through Wordnet. The Web documents are now characterized by ontology concepts, and DB-Scan clustering algorithm is applied. The results of the clustering are utilized to derive the proposed ontology changes. Four types of suggestions are being provided to the user: new concept insertion, new instance insertion, concept merging and concepts declining. The changes to be committed must be finally approved by a domain expert.

3.1 Method description

In this part we informally elaborate on our method, using concepts that are formally defined in the next section. For *new concepts insertion* the candidate concepts are keywords that appear in more than a given percentage of documents of the collection and can be found in Wordnet. For each keyword being a new candidate concept we find the closest concept of the ontology to it, using a hierarchal thesaurus, and its relation with this ontology concept (parent, child, sibling). Based on this relation, the suggested placement of the new candidate concept is generated. A confidence value that ranges from 0 to 1 for each proposed new concept is also provided. This value expresses the "amount" of confidence with which the insertion of this new concept to the ontology is suggested to the domain expert. In the case of a concept c_1 proposed as a sibling to ontology node c_2 , c_1 is added to the ontology at the same level as c_2 . In case c_1 is suggested as a parent of c_2 , the ontology node is moved one level downwards, and we add the new concept as its parent. In case c_1 is suggested as a child, it is simply added as c_2 's child.

New instances suggested are keywords that appear in more than a given percentage of clusters in the clusters' collection and cannot be found in Wordnet. Example of instances, could be companies' or a products' names. In the case of the new concepts the frequency of appearance in the collection of the documents is taken into consideration, while in the case of new instances the frequency of appearance in the collection of *clusters* is being considered. The reason for this decision is two-fold: 1) In the case of new instances, there is no hint on the placement in the ontology. Since instances do not exist in a hierarchical thesaurus, a relation between them and the concepts of the ontology cannot be derived. Thus, taking into account the frequency of their appearance in the level of the clusters collection we have a hint on suggested placements of these new instances, which derives from consulting the rest of the keywords characterizing those clusters, and 2) We would like to diminish the possibility that an unimportant word that has "escaped" from the THESUS stopwords removal and stemming mechanism would be suggested as a new instance in the ontology. If we had taken into consideration such a word's frequency in the document collection, this possibility would be high.

Concepts merging can be derived from ontology concepts that are frequently found to co-participate in clusters' labels. In the THESUS clustering module, each cluster is attached with a label that consists of those terms of the ontology that best characterize the cluster (Varlamis et. al. (2003)). These clusters' labels are used to extract information about the relationships between the ontology concepts. More specifically, ontology concepts that are frequently found to co-participate in clusters' labels are assumed to have a relation, defined as correlation between them.

Finally, the *candidate concepts declining* are defined as the concepts of the ontology that do not appear in any document's keywords and do not have any children in the ontology.

3.2 Definitions

Let a_1 and a_2 be probability thresholds, O_N be an ontology, H_T be a Hierarchical Thesaurus, D_C be a web document collection pertaining with O_N and C_C be the document clusters collection.

Definition 1. A *new ontology concept* C_N is a keyword K_N if $K_N \in H_T$. and if K_N appears in N documents (with $N \leq |D_C|$) so that the following holds: $(\frac{N}{|D_C|}) \geq a_1$.

Definition 2. Given a keyword $K_N \in H_T$ having already been characterized as a *new concept* for an ontology O_N , the ontology concept $O_i \in O_N$ with which K_N should be related if K_N was inserted to the ontology is the one for which the $\text{sim}(K_N, O_i)$, according to the extended Wu and Palmer measure (Nguyen (2003)), becomes maximum. The relation suggested between them in the ontology would be the same with the one found in H_T . The *confidence* with which this suggestion is made is equal to the $\text{sim}(K_N, O_i)$ and it belongs in $[0,1]$.

Definition 3. A *new ontology instance* C_N is a keyword K_N if $K_N \notin H_T$ and if K_N appears in N_C clusters, so that the following holds: $(\frac{N_C}{|C_C|}) \geq a_2$.

Definition 4. Let C be the set of all the concepts C_i of a given ontology and $P'(C)$ the power set of C , let alone the empty set and all the unary sets. If for all $S_i \in P'(C)$ the times of occurrence (TOO) are computed with respect to C_C , then we define the *correlation* of all the concepts C_i belonging to S_i as follows: $\text{Cor}(C_i \in S_i) = \frac{\text{TOO}(S_i)}{\max_j \{\text{TOO}(S_j \in P'(C))\}}$

Definition 5. Given a threshold t and the correlations of all the $S_i \in P'(C)$ (as those defined above), and a set $S \in P'(C)$ containing concepts C_i , in order that *concepts C_i merging* should be suggested if the following holds: $\text{Cor}(C_i \in S_i) \geq t$.

4 Experimental evaluation

For the purposes of the experimental evaluation, we have implemented our method as an add-on to the THESUS system. The domain ontology used is shown in Figure

1. The experimental set up involved the evolution of the given domain ontology, so that it better describes Web documents pertaining with the domain of the heavy industry.

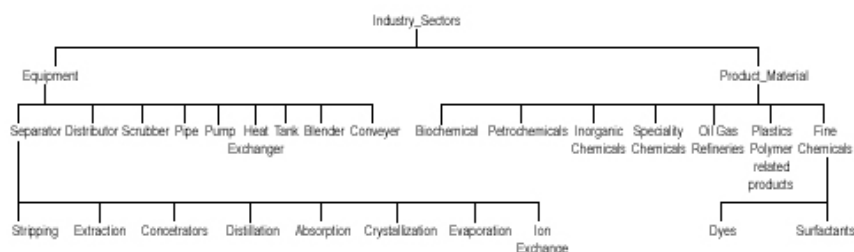


Fig. 1. Domain Ontology used, pertaining with the heavy industry.

Initially, given 10 URLs by a domain expert, an execution of our method took place with default values for a_1 , a_2 and t being $a_1 = a_2 = t = 20\%$, and for the number of hops in the crawling being 2. The fetched Web documents were mapped to the initial ontology and characterized with its concepts.

ChangeNumber	Change Type	Change description
1	Addition of new concept	Add <i>Engineering</i> as a parent concept of <i>Tank</i> , with confidence 0.8.
2	Addition of new concept	Add <i>Chemical</i> as a parent concept of <i>Biochemical</i> , with confidence 0.7777.
3	Addition of new instance	Add <i>BP</i> as a new instance of <i>Oil_Gas_Refineries</i> (thus as its child node), with 21% frequency of appearance in the clusters collection.
4	Concepts Merging	Merge <i>Product_Material</i> , <i>Biochemical</i> and <i>Tank</i> with confidence 0.8571

Table 1. Suggested Ontology changes that were approved by the domain expert.

In parallel from the suggested ontology changes (17 in total), only the changes shown in Table 1 were approved by the domain expert to be committed in the ontology. Having the two ontologies (the original and the evolved one), we aimed at collecting a large set of URLs pertaining with the domain and then feed them into THESUS twice: once for characterizing these URLs with the initial ontology's concepts and once more for characterizing them with the evolved ontology's concepts.

The evaluation was now reduced to providing the URLs and the two different set of characterizations to blind testers and asking them to evaluate with a score varying from 1 to 5 (with 1 being totally irrelevant and 5 being totally relevant) each characterization. The results of the blind testing are shown in Figure 2.

The total number of the URLs fetched for this blind testing was 108, and the fetching commenced with 10 (different from the initial used in evolution) URLs

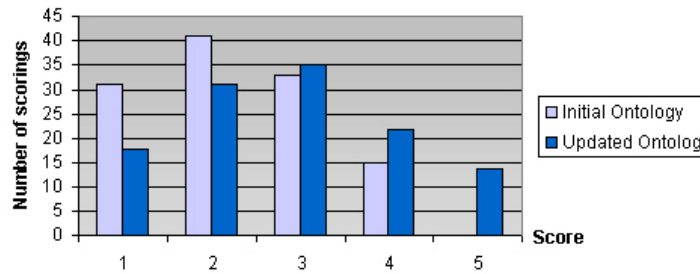


Fig. 2. Results of blind testing scorings.

provided by the domain expert. The percentage of the score characterizations represents the mean value of the scores given by all blind testers. When comparing the bad document characterizations arising from both ontologies' concepts, Figure 2 suggests that the initial ontology provided us with more scorings between 1-2, than the evolved one. Furthermore, commenting on the good characterizations (scores 3-5), these are more in the case of the updated ontology, let alone the fact that scores of 5 were only assigned when the updated ontology was used. In conclusion, as it is shown by this early experimentation, the characterization of the Web pages was improved after applying the suggested changes to the ontology arising from the use of our method.

5 Related work

An approach that exploits WWW content in order to enrich ontologies is presented in Agirre et. al. (2000). It suggests a technique to link document collections from the Web to concepts in WordNet. Concepts are linked to topically related words that form the topic signature for each concept in the hierarchy. They experiment with clustering the concepts that belong to a given word. Binary hierarchical clustering is used on the retrieved concepts. An enrichment process based on the statistical information of word usage is presented in Faatz and Steimnetz (2002). It presents a method for ontology enrichment by comparing the distance measures of collocational information in a text corpus. It suggests that better results are achieved using a more specialized corpus compared to pages retrieved from a more general corpus such as Google. Roux et.al. (2000) propose a system to categorize terms with the joint utilization of an existing ontology and verb patterns defined as small conceptual graphs. The extraction mechanism is built upon conceptual graph architecture, in conjunction with a domain specific ontology. The proposed method differs from Aggirre et. al. (2000) in the fashion that it uses a clustering algorithm to cluster the fetched Web documents and not the concepts. Furthermore, it differs from Faatz and Steimnetz (2002) and Roux et.al. (2000) in that it makes use of clustering techniques.

6 Conclusions and future work

In this paper we have presented a method for ontology evolution by mining the WWW. Our evaluation shows that the characterization of Web pages improves after applying the ontology updates based on our suggestions. Thus the proposed changes generated can enrich the conceptualization of a domain represented by a given ontology. Part of our future work is to tackle with other types of composite changes that could be taken into account when providing suggestions to the users. These include, for example, the splitting of a concept into several subconcepts, and the removal of ontology concepts that might have children. Furthermore, it is essential to notice that one ontology change can trigger other changes. Due to that reason, before applying a change to the ontology, a list of all implications to the ontology should be generated and presented to the user, as well as an algorithm that lists the proposed changes according to an importance measure should be injected in our method. Finally, experiments in a large scale should take place, and a framework for better and automatic evaluation of the results should be constructed.

References

- AGIRRE, E., ANSA, O., HOVY, E. and MARTINEZ, D. (2000): Enriching very large ontologies using the WWW. *In Proc. of the Workshop on Ontology Learning, ECAI'00*.
- FAATZ, A. and STEIMNETZ, R. (2002): Ontology Enrichment with Texts from the WWW. *In Proc. of Semantic Web Mining Workshop at ECML/PKDD-2002*.
- MAEDCHE, A., MOTIK, B., STOJANOVIC, L., STUDER, R. and VOLZ, R. (2003): Ontologies for Enterprise Knowledge Management. *IEEE Intelligent Systems Vol. 18, No. 2*.
- MAEDCHE, A., MOTIK, B., STOJANOVIC, L., STUDER, R. and VOLZ, R. (2003): An Infrastructure for Searching, Reusing and Evolving Distributed Ontologies. *In Proc. of WWW2003*.
- MOTIK, B., MAEDCHE, A. and VOLZ, R. (2002): A Conceptual Modeling Approach for building semantics-driven enterprise applications. *In Proc. of ODBASE-2002*.
- NGUYEN, B., VAZIRGIANNIS, M., VARLAMIS, I. and HALKIDI, M. (2003): Organizing Web Documents into Thematic Subsets using an Ontology. *VLDB journal, special issue on "Semantic Web", vol. 12, number 4, pp. 320-332*.
- ROUX, C., PROUX, D., RECHENMANN, F. and JULLIAR, L. (2000): An Ontology Enrichment Method for a Pragmatic Information Extraction System gathering Data on Genetic Interactions. *In Proc. of OL'2000 Workshop*.
- STOJANOVIC, L., MAEDCHE, A., MOTIK, B. and STOJANOVIC, N. (2002): User-driven Ontology Evolution Management. *In Proc. of EKAW2002*.
- VARLAMIS, I., VAZIRGIANNIS, M., HALKIDI, M., and NGUYEN, B. (2003): THE-SUS: Effective Thematic Selection And Organization Of Web Document Collections Based On Link Semantics. *IEEE Transactions on Knowledge and Data Engineering Journal, vol. 16, No. 6, pp. 585-600*.
- WU, Z. and PALMER, M. (1994): Verb Semantics and Lexical Selection. *In Proc. of ACL-1994*.