Knowledge Enhancement Through Ontology-Guided Text Mining

Muhammad Abulaish[1] and Lipika Dey[2,*]

[1] Department of Mathematics, Jamia Millia Islamia (A Central University),
New Delhi-25, India
mdabulaish@yahoo.com
[2] Department of Mathematics, Indian Institute of Technology,
New Delhi-16, India
lipika@maths.iitd.ernet.in

Abstract. In this paper we have proposed a system that performs both ontology-based text information extraction and ontology update using the extracted information. The system employs text-mining techniques to mine information from text documents guided by an underlying ontology. It also enhances the existing ontology with new concepts and their descriptors which may be precise and/ or imprecise, mined from the text. All extracted information related to concepts and concept descriptors are also stored in a structured knowledge base.

Keywords: Text mining, Information extraction, Ontology enhancement.

1 Introduction

Semantic analysis of texts with the help of structured domain knowledge can help in extracting information effectively from the documents. Ontologies store the key concepts and their inter-relationships thereby providing a shared common understanding of a domain [3]. Generally a concept in Ontology is defined in terms of its mandatory and optional properties along with the value restrictions on those properties. However, ontology-based text mining poses its own problems. Firstly, the concepts even if present in documents need not be present in conjunction with the same descriptors, rather may appear with new descriptors or even in association with qualifiers, which defines a fuzzy presence of the properties. Moreover, information extraction from web documents has to also account for imprecise concept descriptions presented by users while specifying a query [1]. A description is termed *imprecise* if it can have varying degrees of similarity with known descriptions. As of now, there is no general ontological framework for qualifying a property. Another bottleneck in designing ontology based information extraction systems arises from the fact that most of the existing domain ontologies are created and maintained manually by domain experts, which is a costly and time-consuming task. A critical issue in developing such systems therefore is the task of *identifying, defining* and *embedding* new concepts and concept descriptions into existing domain ontologies.

* Author for Correspondence.

S.K. Pal et al. (Eds.): PReMI 2005, LNCS 3776, pp. 601 – 604, 2005.

This paper explores the possibility of focused information extraction from web documents in an ontology-guided way. The proposed text-mining system parses documents syntactically and then starts looking for known concepts present in the domain seed ontology. Thereafter it applies semantic analysis for inferring related concepts. The mined concepts are integrated into the existing ontological structure to enhance it. The rest of the paper is organized as follows. We present a brief overview of some related works on ontology learning and ontology-based text processing systems in section 2. Section 3 presents the system overview followed by some results that are given in section 4. Finally, we conclude the paper in section 5.

2 Related Work

The use of ontological models to access and integrate large knowledge repositories in a principled way has an enormous potential to enrich and make accessible unprecedented amounts of knowledge for reasoning [2]. Ontology representation languages like DAML+OIL and OWL are based on Description Logics (DLs) thus enabling the knowledge representation systems to provide reasoning support as well [4]. A number of systems exist to help in the process of construction of domain ontologies, of which the most famous one is the Protégé[1]. In [6] a java-based tool that helps domain experts by providing a graphical interface for domain ontology creation and testing is proposed. This is then used to extract data from web documents and to store them in structured form. [8] reported a text-mining tool to identify, define, and enter concept descriptions into ontology structures. In [5] an abstract Web mining model for gathering and extracting concepts by analyzing approximate concepts hidden in user profiles on the Semantic Web has been proposed. All the above works assume that concept descriptions are precise. Answering imprecise queries over structured databases have been addressed in [7].

3 Proposed Framework

Fig. 1 presents the architecture of the proposed system, which consists of five major modules – *Document Processor, Concept Miner, Fuzzy Description Generator, Ontology Editor* and *Ontology Parser*. The *Document Processor* applies shallow parsing techniques to identify relevant segments in a document and stores them into a tree structured form. Both *Concept Miner* and *Fuzzy Description Generator* work on these tree structures to enhance the ontology with new concepts extracted from documents and to instantiate the structured knowledge base. Ontology update is implemented through the *ontology editor*. The structured knowledge base is created to store the extracted information in a structured format. The design of the knowledge base is based on the ontology and is output by the *ontology parser*. The document processor outputs two similar tree structures differing in the information components contained in their nodes. For the *Concept Miner*, the nouns, adjectives and verbs identified in the document are stored in the tree generated as follows:

[1] http://protégé.stanford.edu

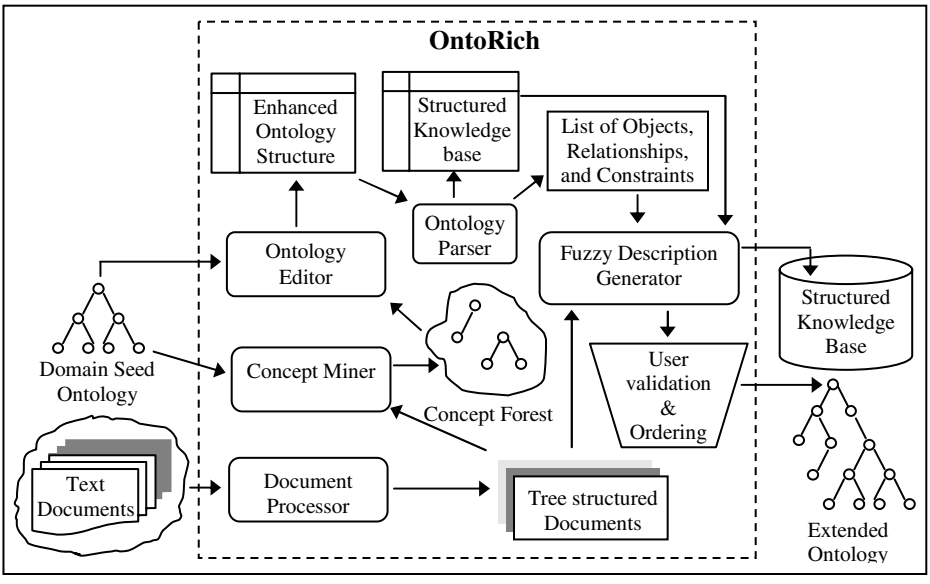

Fig. 1. Proposed system architecture for Ontology enhancement

Root (**R**): A node that contains the *verb phrase* of a segment.
Lchild (**L**): A node that contains *noun phrase* left of the verb phrase considered at R.
Mchild (**M**): A node that contains *noun phrase* right of the verb phrase considered at R
Rchild: points to the root of the sub-tree constructed from the next segment.

The equivalent context-free grammar for this is given as follows:

Document (D) $\rightarrow S^*$, $S \rightarrow LRMS \mid \in$, $L \rightarrow (N+J)^*$, $R \rightarrow V$, $M \rightarrow (N+J)^*$

where, S denotes an English sentence, N, J and V are noun, adjective and verb tags respectively assigned by the POS tagger. Similarly, the tree for the *Fuzzy Description Generator* is created using nouns, adjectives, verbs and adverbs from the documents. The *Concept Miner* and the *Fuzzy Description Generator* mine the respective trees to locate ontology concepts and organize the extracted information into a structured knowledge base. They work in an interactive way to enrich the ontology structures with the newly mined concepts, relations, concept descriptions and qualifier sets for concept descriptions. Imprecise concept descriptions are stored using a fuzzy ontology structure, in which a concept descriptor is a combination of a qualifier and a value. Some new descriptors that were learnt for wine by the system are *exquisite flavor*, *indigenous red color* etc.

4 Experiments and Results

We have conducted experiments using a number of different domain ontologies like Wine, and the GENIA ontology for categorizing biological substances and locations

Table 1. Precision and recall of concept recognition in text documents

Domain	# Relevant Concepts Extracted	# Non-relevant Concepts Extracted	# Relevant Concepts Missed	Precision	Recall
Wine	209	16	50	92.89 %	80.69%
Molecular Biology	137	41	10	76.97%	93.20%

(virus, mono-cell, body-part etc.) The GENIA ontology specifies only taxonomic relations among the basic classes and is enhanced by our proposed system to incorporate new biological relations like "*induce*", "*activate", "associate", "bind" etc*. Since a biological relation may be associated with different pairs of biological concepts each relation is accompanied by a fuzzy membership value to represent its association strength. The membership value is directly proportional to the frequency of co-occurrences of a relation and associated biological concept-pairs. Table 1 summarizes the precision and recall values for identifying and storing relevant concepts and their descriptions from text documents for the two domains.

5 Conclusions and Future Work

In this paper we have presented an ontology-based text-mining system to enhance domain seed ontologies by extracting relevant concepts as well as precise and imprecise concept descriptors from text documents. The extracted instances of concepts are used to upgrade the ontology and answer user queries efficiently.

References

1. Abulaish, M., Dey, L.: Using Part-of-speech Patterns and Domain Ontology to Mine Imprecise Concepts from Text Documents. In Proceedings of the 6th International Conference on Information Integration and Web Based Applications and Services (iiWAS'04) Jakarta, Indonesia, (2004) 91-100
2. Crow, L., Shadbolt, N.: Extracting focused knowledge from the semantic web. Intl. Journal Human-Computer Studies 54 (2001) 155-184
3. Fensel, D., van Harmelen, F., Horrocks, I., McGuinness, D. L., Patel-Schneider, P. F.: OIL: An ontology Infrastructure for the Semantic Web. IEEE Intelligent Systems, 16 (2), (2001) 38-45
4. Horrocks I., Sattler, U.: Ontology Reasoning in the SHOQ(D) Description Logic. In Proceedings of the 7th IJCAI'01, (2001) 99-204
5. Li, Y., Zhong, N.: Web Mining Model and its Applications for Information Gathering. Knowledge-Based Systems 17 (2004) 207-217
6. Liddle, S. W., Hewett, K. A., Embley, D. W.: An Integrated Ontology Development Environment for Data Extraction. In Proceedings of ISTA'03, (2003) 21-33
7. Nambiar U., Kambhampati, S.: Answering Imprecise Database Queries: A Novel Approach. In Proceedings of the 5th ACM CIKM Workshop on Web Information and Data Management (WIDM'03), New Orleans, (2003)
8. Velardi, P., Fabriani, P., Missikoff, M.: Using Text Processing Techniques to Automatically Enrich a Domain Ontology. In Proceedings of ACM Conference on Formal Ontologies and Information Systems, FOIS, (2001) 270-284