Mining Domain Specific Texts and Glossaries to Evaluate and Enrich Domain Ontologies

Viral Parekh
Department of Computer Science
and Electrical Engineering,
University of Maryland,
Baltimore County
Baltimore, MD 21250
Email: viral1@umbc.edu

Jack Gwo
Department of Civil and
Environmental Engineering
University of Maryland,
Baltimore County
Baltimore, MD 21250
Email: jgwo@umbc.edu

Tim Finin
Department of Computer Science
and Electrical Engineering,
University of Maryland,
Baltimore County
Baltimore, MD 21250
Email: finin@umbc.edu

Abstract

Ontologies have been widely accepted as the most advanced knowledge representation model. They are among the most important building blocks of semantic web, hence, very crucial for the success of semantic web. This paper discusses a fast and efficient method to facilitate the evaluation and enrichment of domain ontologies using a text-mining approach. We exploit domain specific texts and glossaries or dictionaries in order to automatically generate g-groups and f-groups. These groups are sets of concepts/terms which have either taxonomic or non-taxonomic relationships among them. The domain expert ontology engineer reviews these generated groups and uses them to evaluate and enrich the domain ontology. We have developed an extensive and detailed ontology in the field of environmental science using this approach in interaction with domain expert. Empirical results show that our approach can support domain expert ontology engineers in building domain specific ontologies efficiently.

Keywords

ontology enrichment, text mining, clustering, feature groups

1. Introduction

Ontologies have been developed to capture the knowledge of a real world domain. "Ontology is defined as a formal and explicit specification of a shared conceptualization of a domain. They provide a shared and common understanding of a particular domain of interest" [24,10]. A domain ontology defines a vocabulary of concepts and their relationships for that given domain, thereby defining the domain semantics. Other benefits that could be derived from the use of ontologies apart from knowledge sharing are reusability of domain knowledge and separation of domain knowledge from operational knowledge [17]. Ontologies are viewed as the most advanced knowledge representation model.

The emergence of the Semantic Web has marked another stage in the ontology research field. "The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" [1]. Ontologies form the backbone and most essential ingredient for the success of semantic web. They provide shared domain models which are understandable to both humans as well as machines. The key uses of ontologies in the Semantic Web world are knowledge sharing, ontology based reasoning, information integration and interoperability. Hence, developing domain specific ontologies is very crucial for semantic web.

As we have seen, fast and efficient strategies for ontology development are much needed today. Huge effort is needed from the domain expert in order to construct ontologies manually, especially in case where the application domain is large such as ours. There is a need for semi-automatic approach in ontology building which will help the domain expert in constructing extensive domain ontologies efficiently.

In this paper, we describe how an existing seed ontology can be enriched and evaluated by the domain expert ontology engineer. We propose the use of text mining techniques, especially mining the domain specific texts and glossaries/dictionaries in order to find groups of concepts/terms which are related to each other. Such groups of related concepts/terms will enable the domain expert to either, evaluate and update the existing ontology in case those concepts are already defined in the ontology, or to enrich the existing ontology in case those concepts are not defined. This is an iterative refinement process with the newly available knowledge bases and/or domain specific texts or glossaries.

Several representation languages for ontologies are currently proposed. Recently, W3C announced Resource Description Framework (RDF) [27] and Web Ontology Language (OWL) [26] as W3C recommendations. We are using OWL to describe our ontology. We are developing an ontology for the domain of environmental science and engineering using the approach presented in this paper in interaction with environmental science domain expert. We chose this domain due to the absence of any formal ontology in this field.

Section 2 describes the related work in the area of ontology learning and enrichment. Section 3 describes our approach – mining the glossaries domain text to help the domain expert to evaluate and update the existing ontology. In Section 4, we discuss our results. Section 5 presents our conclusions.

2. Related Work

Several disciplines have contributed to facilitate and expedite the construction of ontologies, especially natural language processing (NLP), data and text mining, clustering and machine learning.

In the field of NLP based ontology learning methods, Semi-Automatic Domain Ontology Acquisition Tool (SOAT) acquires relationships using a predefined knowledge representation framework, that integrates linguistic, commonsense and domain knowledge [29]. [11] uses pattern matching approach to learn new relationships between concepts in an ontology while [19] focuses on verb patterns to enrich the existing ontology.

[12] aims to convert dictionary to a graph structure where each node is a headword from the dictionary and arcs between nodes represent the use of other headwords for the definition of one particular node i.e. headword. Their approach uses algebraic extraction techniques to output a set of related terms. [7] mines the WWW and enriches the ontology based on the comparison between statistical information of word usage in the corpus and structure of the ontology itself.

Different conceptual clustering methods have been used for the semiautomatic construction of ontologies [2, 4]. Concepts are grouped according to the semantic distance between them. Text-To-Onto tool [14, 15] developed at the AIFB Institute in the University of Karlsruhe uses association rule mining and statistical approach to find relations among concepts. [30] uses Wordnet [8] and domain specific texts to discover relationships based on the WordSpace model. Most of these methodologies are based on the idea that frequent co-occurrence of concepts suggests relationships among them. Our approach of mining the domain specific texts lies in this area.

Many of these methods including our approach assume that most concepts to be included in the ontology will be present in the corpus to be analyzed. Our approach supplements domain texts with glossaries for ontology enrichment and evaluation.

3. Methodology

3.1 Overview

Fig 1 illustrates our approach of ontology building. The ontology engineer builds an initial seed ontology containing the relevant domain concepts and the relationships among the ontology concepts. He/She manually collects

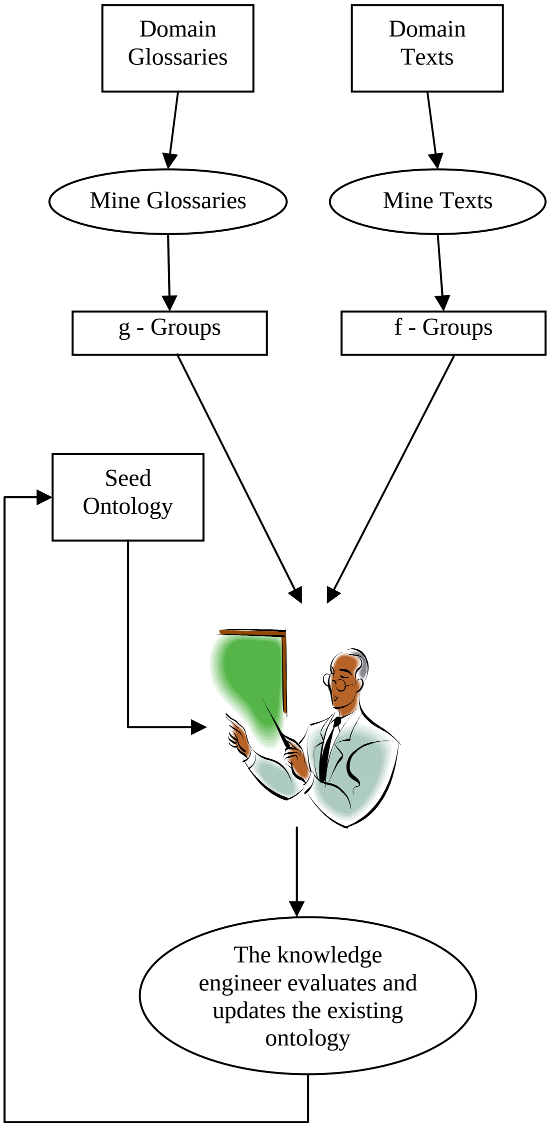

Fig 1. Ontology Building

glossaries/dictionaries and relevant domain specific texts. These documents contain text relevant to the knowledge represented in the seed ontology as well as the entire domain knowledge to be modeled. These documents form the inputs for the two mining components, each utilizing a specific technique to mine the specific form of text and output a set of groups of concepts/terms, specifically g-groups and f-groups. In general, both g-groups and f-groups refer to sets of concepts/terms which are related to each other. The distinction between them will be made clear in the following sub-sections. One such f-group could be {porous medium, hydraulic conductivity, water, porosity} clearly indicating that the terms are related. The ontology engineer reviews the groups and updates the seed ontology using these groups and his/her own expertise in the domain. This process is iterative and can be repeated with the newly available knowledge bases and domain texts or glossaries.

3.2 Collecting data and Pre-processing

Relevant domain specific texts may be manually collected by the domain expert. In our case, the domain expert collected around 100 relevant documents in the domain of environmental science from resources such as U.S. Geological Survey (USGS) [23], U.S. Environmental Protection Agency (EPA) [22], Oak Ridge National Laboratory Environmental Sciences Division Research & Development WWW Site [18] and so on. We believe that in such a large domain specific homogenous text collection, related concepts/terms can be found in similar context [21].

Apart from domain specific text, the domain expert needs to collect several glossaries/dictionaries. The reason for using glossaries/dictionaries is that they are widely used by domain experts to manually build ontologies. The content of glossaries/dictionaries follow a definite structure with the headwords and their definition being separate, and a definite visible relationship between the words used in the definition and the headword of the definition. This organized text structure facilitates the task of mining these glossaries/dictionaries and improves the results. Domain specific ontologies are designed to be shared by people and reused by a wide community. Hence it is necessary for an ontology to be acceptable and useful to various audience in the community. This led us to the use of not just one, but many available glossaries/dictionaries. We believe that mining several glossaries/dictionaries will capture the view of many domain experts who have written these glossaries/dictionaries, thereby making the ontology more extensive and complete. However, the final design decision will be in the hands of the ontology engineer. This approach guarantees that he/she will be aware of the view of several domain experts before he/she takes any decisions. In our case, the domain expert collected 5 glossaries/dictionaries from resources such as USGS [23], EPA [22], WEF [28] and other web resources.

Before mining any text or glossaries, we perform certain pre-processing operations:

1. removal of stop words
2. change the upper case characters to lower case
3. perform stemming
4. removal of irrelevant or generic terms by comparing the distribution of terms within this domain corpus and in a more general collection of large set of random documents [25]. Thus, we use generic corpus as a filter to perform this task.

We now discuss our approach for mining each of the two forms of text – glossaries and domain text.

3.3 Mining the glossaries/dictionaries

Glossaries/dictionaries are more organized and structured when compared to domain text. We are interested in automatically generating the g-groups (glossary groups) of concepts/terms after mining the glossaries. A g-group is associated with each headword that is included in the glossaries/dictionaries and is defined as a set of concepts/terms which are frequently used to define that headword. Due to the use of several glossaries/dictionaries, there will be many definitions for each headword. This g-group for each headword is generated after considering all the definitions of that headword among the collected glossaries/dictionaries. In order for any concept/term to qualify as a member of a g-group for a particular headword, it has to be present in at least two third of the definitions for that headword among all the glossaries/dictionaries. We came up with this value after experimenting with different values and reviewing the results in consultation with the domain expert ontology engineer.

The algorithm for this process is as follows:

1. Parse all the collected glossaries and dictionaries, and prepare a list of unique headwords that are defined in them.
2. For each headword present in the list do the following:
   2.a. Scan all the definitions of that particular headword.
   2.b. Prepare a list of terms that occur in at least two third of these definitions.

Here we assume that there is a clear distinction between the headword and its definition, and the text parser is aware of this. The extracted list for each headword now becomes the g-group for that headword. We also include the headword in its g-group as the first member of the g-group. This identifies the g-group. For example, consider a g-group *{**porous medium**, porosity, permeability, void ratio}*. The first bold member of the g-group is the headword associated with the g-group. The other 3 members in this g-group are the representative terms for the headword *porous medium*. From our algorithm, these 3 terms are present in this g-group since they are observed to be frequently used in the definitions of the headword *porous medium* and they are present in at least two third of all the definitions found for *porous medium* in all the collected glossaries/dictionaries.

Our results show that many of these extracted g-groups are quite useful to the ontology engineer in updating the ontology. These g-groups represent the terms defined in the collected glossaries, thereby avoiding the domain expert ontology engineer to manually go through these glossaries to build ontologies.

3.4 Mining the domain text

Our goal in this module is to generate efficient f-groups (feature groups). We define a f-group as a set of features (concepts/terms) which are related to each other based on their lexical co-occurrence within similar contexts. The context information is used to group related features (concepts/terms) because in large homogenous text corpus related features can be found in similar context [21]. We define context of a feature *f* as the set of features that occur in all sentences containing *f*.

The text corpus is free of most of the irrelevant or noise words due to our pruning step using a generic corpus, as discussed in the Pre-processing section 3.2 above. We now divide the entire text corpus into a set of sentences and each sentence into a set of features. Now, the context of a feature *f* is the set of features that occur in sentences along with *f*. We associate a context vector c(*f*) = {c_1, c_2, ..…, c_m} with each extracted unique feature *f* (m is the number of unique features). The i-th coordinate c_i of context vector c(*f*) denotes the number of sentences in the entire text corpus that contain both features *f* and f_i. Let {s_1, s_2, ..…, s_n} be the set of sentences in the text corpus. We now build a "feature by sentence matrix S" as an (m x n) matrix whose entry S_{ij} is the number of times the feature f_i occurs in the sentence s_j. The i-th column of the symmetric matrix SS^T corresponds to the context vector of feature f_i. Thus, we now have the context vectors of each unique feature extracted from the text corpus.

Feature grouping can then be performed by using a clustering technique on these context vectors. We use a Singular Value Decomposition (SVD) based clustering technique called Principal Direction Divisive Partitioning (PDDP). This clustering technique is memory efficient and fast. The details of the PDDP clustering technique and its use to generate feature groups can be found in [20, 6, 3]. When we applied this technique to the domain text and generated the f-groups, the domain expert ontology engineer was able to discover several interesting relationships as will be seen in the Results section.

3.5 Ontology evaluation and enrichment

The domain expert ontology engineer now reviews the seed ontology prepared by him/her or any existing domain ontology and uses the automatically generated g-groups and f-groups in order to evaluate the ontology and update it in this process. The ontology engineer may include new domain concepts into the ontology or assign new attributes to existing concepts or update the concept hierarchy. Both f-groups and g-groups may include concepts/terms that have either taxonomic or non-taxonomic relationships among them. Certain groups may also suggest specific individuals of the concept classes defined in the ontology. Overall, these groups facilitate the ontology engineer's task of making the ontology complete and extensive.

4. Results

In this section, we discuss some of the results of our experiments with the environmental science and engineering domain ontology. Our experiments in consultation with the domain expert ontology engineer proved that the automatically generated g-groups and f-groups were quite useful to him in building the domain ontology. In order to prove the effectiveness of our results, we will take a particular sub-section of the seed ontology and show how the domain expert ontology engineer used the automatically extracted groupings to update that sub-section of the seed ontology.

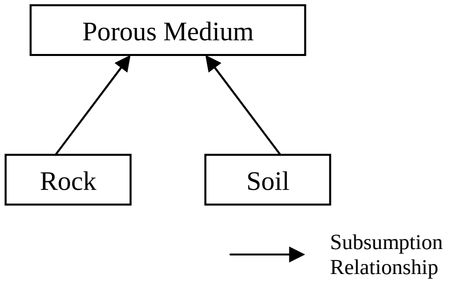

Fig 2. A small sub-section of the seed ontology

Category	Group
g-group	{**porous medium**, porosity, natural, artificial, void-ratio}
g-group	{**sedimentary rock**, rock, sediment, shale, limestone, organic, chemical}
g-group	{**igneous rock**, solidify, molten, volcanic, plutonic, surface}
g-group	{**metamorphic rock**, heat, pressure, rock}
f-group	{porous medium, hydraulic conductivity, water, porosity}
f-group	{moisture content, porous medium, permeability, water}

Table 1 A few of the g-groups and f-groups that are generated after mining the glossaries and domain text

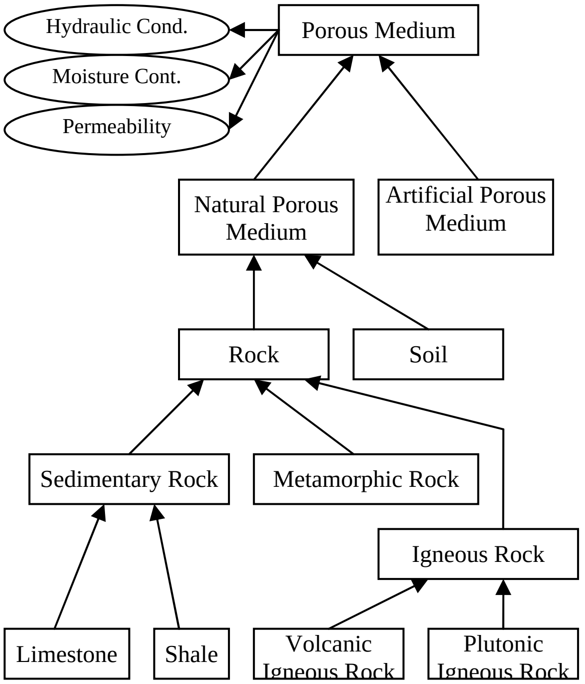

Fig 3. An enriched version of the seed ontology (Fig 2) using the groupings of Table 1

Figure 2 shows a small sub-section of the seed ontology. It defines 3 concepts and a taxonomy among them. Table 1 includes a few of the f-groups and the g-groups that were automatically generated after mining the domain specific text and the glossaries/dictionaries respectively. The domain expert ontology engineer used these groupings of concepts/terms and his own domain expertise in order to enhance and evaluate the seed ontology. Figure 3 shows an enriched version of the same sub-section of the seed ontology. New concepts have been added, the taxonomy has been updated and new attributes have been introduced. However, this is a very small sub-section of the extensive environmental science ontology that the domain expert ontology engineer was able to build using the generated f-groups and g-groups.

5. Conclusion

We have a system which uses certain text mining techniques to automatically extract groupings of related terms/concepts from domain specific texts and glossaries/dictionaries. The domain expert ontology engineer may then use these groups in order to enrich and evaluate an existing ontology. We have investigated the effectiveness of our approach and these automatically generated groups have helped the ontology engineer discover many important and interesting relationships in the domain of interest, thereby helping the ontology engineer to construct domain ontology efficiently.

6. Acknowledgements

We would like to thank Mr. Vishal Shanbag for making available his thesis and Dr. Jacob Kogan of University of Maryland, Baltimore County for his advice.

7. References

[1]Berners-Lee T., Hendler J., Lassila O., The Semantic Web, Scientific American, May 2001

[2]Bisson G, Nedellec C, Cañamero D. (2000) Designing Clustering Methods for Ontology Building. The Mo'K Workbench. Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence, ECAI'00, Berlin, Germany, August 20-25.

[3]Boley D., Principal Direction Divisive Partitioning, Data Mining and Knowledge Discovery, 2(4):325:344, 1998.

[4]Chaelandar G, Grau B. (2000) SVETLAN'- A System to Classify Words in Context. Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence ECAI'00, Berlin, Germany, August 20-25.

[5]Craven M., DiPasquo D., Fre-itag D., McCallum A., Nigam K., Mitchell T., and Slat-tery S. Learning to Construct Knowledge Bases from the World Wide Web. Artificial Intelligence, 1999.

[6]Dhillon I.S., Kogan J. and Nicholas C. Feature Selection and Document Clustering, A Comprehensive Survey of Text Mining, Springer-Verlag, 73-100, 2003

[7]Faatz A. and Steinmetz R. (2002). Ontology enrichment with texts from the WWW. Semantic Web Mining 2nd Workshop at ECML/PKDD-2002, 20th August 2002, Helsinki, Finland

[8]Fellbaum C. ed: “Wordnet”, The MIT Press, 1998. URL: http://www.cogsci.princeton.edu/~wn/

[9]Gómez-Pérez A., Manzano-Macho A., Universidad Politécnica de Madrid, A survey of ontology learning methods and techniques,

[10]Gruber T., (1995), Towards Principles for the Design of Ontologies Used for Knowledge Sharing, International Journal of Human-Computer studies

[11]Hearst M. A. (1998), Automated Discovery of WordNet Relations. In Christiane Fellbaum (Ed.) WordNet: An Electronic Lexical Database, MIT Press, pp. 132--152.

[12]Jannink, J. & Wiederhold, G. (1999). Ontology maintenance with an algebraic methodology: A case study. In. Proceedings of AAAI workshop on Ontology Management, July 1999.

[13]Maedche A., Staab S., Discovering Conceptual Relations from Text, ECAI2000, pp.321-325 (2000)

[14]Maedche A, Staab S. (2001) Ontology Learning for the Semantic Web. IEEE Intelligent Systems, Special Issue on the Semantic Web, 16(2)

[15]Maedche, A. and Volz, R. (2001) The Ontology Extraction and Maintenance Framework Text-To-Onto. Proceedings of the ICDM Workshop on integrating data mining and knowledge management, San Jose, California, USA.

[17]Noy N. & McGuinness D. (), Ontology Development Guide 101: A guide to creating your first ontology

[18]Oak Ridge National Laboratory (ORNL) Environmental Sciences Division (ESD) Research & Development WWW Site.
URL: http://research.esd.ornl.gov/

[19]Roux C., Proux D., Rechermann F., and Julliard L. (2000). An ontology enrichment method for a pragmatic information extraction system gathering data on genetic interactions. Position paper in Proceedings of the ECAI2000 Workshop on Ontology Learning(OL2000), Berlin, Germany. August 2000.

[20]Shanbag V. (2003) Improving Feature Ranking in Document Clusters using Phrasal Compensation and Feature Grouping, Masters Thesis, University of Maryland, Baltimore County.

[21]Sch¨utze H. and Pedersen J. Information retrieval based on word senses. In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, pages 161–175, Las Vegas, NV, 1995.

[22]U.S. Environmental Protection Agency.
URL: http://www.epa.gov/

[23]U.S. Geological Survey.
URL: http://www.usgs.gov/

[24]Ushold M. & Gruninger M. (1996), Ontologies: Principles, methods and applications, in The Knowledge Engineering Review

[25]Vogel D., Using Generic Corpora to Learn Domain Specific Terminology. Workshop on Link Analysis for Detecting Complex Behavior, Washington, DC, USA, August 27, 2003

[26]W3C: OWL Web Ontology Language Semantics and Abstract Syntax.
URL:http://www.w3.org/TR/2004/REC-owl-semantics-20040210/, 2004.

[27]W3C: Resource Description Framework (RDF) Concepts and Abstract Syntax.
URL: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/, 2004.

[28]Water Environment Foundation.
URL: http://www.wef.org

[29]Wu S.H, Hsu W.L. (2002). SOAT: A Semi-Automatic Domain Ontology Acquisition Tool from Chinese Corpus. In the 19th International Conference on Computational Linguistics, Howard International House and Academia Sinica, Taipei, Taiwan

[30]Yamaguchi T., Acquiring Conceptual Relationships from Domain-Specific Texts, Proceedings of the Second Workshop on Ontology Learning OL'2001 Seattle, USA, August 4, 2001.