J. Basic. Appl. Sci. Res., 3(1)94-98, 2013
© 2013, TextRoad Publication
ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

Automatic Ontology Construction*

Rasoul Dezhkam[1], Marziyeh Khalili[2*]

[1]Department of Computer Engineering , Meymeh Branch, Islamic Azad University, Meymeh, Iran
[2]Software technology engineering student, Department of Computer Engineering , Meymeh Branch, Islamic Azad University, Meymeh, Iran

ABSTRACT

In recent years the uncontrolled development of internet applications and data growth on the Web has highlighted the complexity of data managing and mining. The utmost purpose of Semantic Web is to add the Semantic Knowledge base to Web pages in order to generate a deep potent search. Since the manual generation of ontology and data base is not cost and time-effective, it would be considered as a hindrance to the evolution of Semantic Web activities. Hence, researchers have been making efforts to devise automatic means of ontologies for the web world, in order to achieve the aims of the Semantic Web[10]. One difficulty encountered is that web documents, unlike structured databases, contain unstructured and semi-structured data. Our hypothesis is that creating ontologies to describe the semantics of web data is the key to bridging the gap between semi-structured data and structured databases, and hence facilitating the application of database techniques. We extract an ontology (or conceptual schema) from a set of web pages in a particular application domain automatically. Moreover, a number of Methods ought to be invented to evaluate auto-generated ontologies on the basis of its strengths and weaknesses, and to invent much better methods. In this study, we describe a tool we have developed. The tool is using dual approach method to extract ontology information and to generate ontology automatically by means of browsing on the search engine on Pubmed website. First of all, the most significant knowledge presented in the review of literature is extracted and stored by this method and then, the extracted and stored data will be analyzed to generate the ontology. At last, a comprehensive assessment plan is utilized to compare auto-generated ontologies with corresponding ontologies which are generated manually. Our evaluating method is taken from evaluating techniques in natural language applications.

KEYWORDS: Ontology, Natural Language Processing, Semantic Web, Data Mining, Text Mining.

INTRODUCTION

There exist a few human-made ontologies such as Word Net, Cyc. Theseinclusive ontologies embody few scientific concepts rendering them the least useful for most scientific searches. The UMLS system (unified medical language system) contains more than 620,000 medical concepts. However, no reliable ontologies or knowledge bases exist for other fields. Furthermore, manual ontology construction demands a lot of time and effort from domain experts and ontology engineers.World Net, CYC and UMLS projects took many person-years to come into existence[1].

Recently, there have been efforts for building ontologies semi-automatically from scientific documents[7,9]. Advances in text mining have improved the automatic ontology construction process[5]. In this study, however, relying on text mining technologies, and automatic ontology construction have been much facilitated. We have developed a tool which automatically generates ontology from the existing documents on the Web and then, with a standard assessment procedure, we have analyzed the efficiency of ontology.

1. Ontologies

Ontologies have proven to be the basic component of a large number of applications. To put differently, they have been used in Agent Syestems and Knowledge Management Systems[2]. Today, ontologies are being widely used in computer operating systems. The most significant current setback of ontology engineering is the need to construct different ontologies with different details, focusing on different knowledge with different processing and reasoning capabilities for different domains and applications[6,8,11]. On the other hand, a complete and comprehensive ontology which encompasses all human knowledge does not exist and if it exists, it might be limited to be employed for the particular areas and applications. Although there are differences between ontologies, the following generral aggrements about ontologies exist:

1) there are objects in the world.
2) objects have properties which could take values.
3) objects could be linked together.
4) characteristics and relationships could change over time.
5) there are events that could occure at different times.

*This article is taken from a research project entitled "Automatic Building Ontology"

***Corresponding author:** Marziyeh Khalili, Software technology engineering student, Department of Computer Engineering, Meymeh Branch, Islamic Azad University, Meymeh, Iran. E-mail: marziye.khalili@gmail.com

6) there are phases for participation and events.
7) the world and its objects could be in different positions.
8) the events could cause other events, or influence the situations.
9) objects could have parts.

Definition of ontology[12]

The ontology is difined with a quaternary (C,R,F,A) where,

- C is the collection modelled from the existing concepts in the world;
- R is[a] set of relations between concepts which is divided into two individual subsets: R_N and R_T
  - R_T is the set of hierarchical relations between concepts which create the inclusion hierarchy and they are binary
  - R_N is the set of non-class relations…… $(1 \leq n)$

$$R = R_T \cup R_N$$
$$R_T \cap R_N = \Phi$$
$$R_T = \{ r_n \mid (r_n \subseteq C \times C)\}$$
$$R_N = \{ r_{Ni} \mid (r_{Ni} \subseteq C \times C \times \ldots C)\}$$

- F is the set of existing specifications in ontology about their relations and concepts, and is divided into two subsets: F_T and F_N
  - F_T is the set of existing specifications in ontology about hierarchical relations between concepts. In other words, it represents the hierarchy of inclusion.
  - F_N is the set of existing specifications in ontology about non- hierarchical relations between concepts.

$$F : R \rightarrow C^n$$
$$F = \{(r, c_1, c_2, \ldots, c_n) \mid (r \in R) \wedge (c_i \in C) ; 1 < i \leq n \}$$
$$F = F_T \cup F_N$$
$$F_T \cap F_N = \Phi$$
$$F_T : R_T \rightarrow C \times C$$
$$F_N : R_N \rightarrow C \times C \times \ldots C$$

- A is the set of axiomatic principles, existed in ontology which expressed with a formal language such as logic. As an example, figure 1 represents a small and simple ontology:

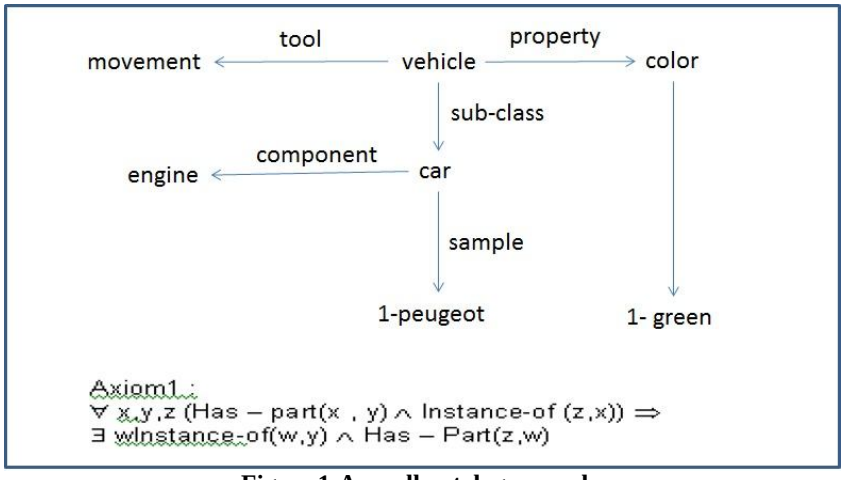

Figure 1:A small ontology sample

The components of the ontology are as follows:

C = {it has component, it has property, car, sub-class, sample, tool, movement, vehicle, color, 1-green, 1-Peugeot, engine}

R_T= {sample, sub-class}

R_N= {tool, color, it has component, it has property}

F_T= {sample (1-green, color), sample (1- Peugeot, car), sub-class (car, vehicle)}

F_N= {it has property (vehicle, color), it has component (car, engine), color (1- Peugeot, 1-green), tool (movement, vehicle)}

A= {Axiom 1}

2. The Introduction of ontology-biology construction tool

To simplify and reduce processing, we have merely used the abstracts of papers instead of enjoying the whole text of papers. The tool which we have constructed is capable of creating a preliminary ontology in the field of biology. Connecting to the internet, of course, the tool constantly and automatically keeps updating its ontology. With respect to the tool's utmost aim, four distinct algorithmic sections can be identified.

- An algorithm for searching on the Web in order to detect articles, corresponding to the query word, and to receive them from the Web.
- An algorithm for extracting and storing articles data in the system
- An algorithm for the analysis of texts based on the query word and displaying results in a readable manner
- An algorithm for displaying the data of the articles to the users

Previous studies suggest that a large number of papers have examined their desired method, using PubMed searching engine Website. Similarly, we enjoy the same search engine. At first, it is required to determine the standard form of the articles. Then, we consider the abstracts as the input of the system.

2.1. The extracted information

The most important data, stored in the articles of PubMed website are the title and the PMID value of the article. Furthermore, we have to extract a summary from each individual article, which is done by extracting the main concepts of each article. Many extracting information researchers focused on extracting pre-defined components. To do this, we assume that the target words of interest are the ones that have more frequent occurrences. In reference # 3, they wanted to recognize Protein-Protein interactions by analyzing the texts. Therefore, they used a number of predefined protein names and a limited set of verbs which indicated interactions. In the meantime, we do not seek protein names, however, the paper gives us the idea that a number of important verbs of each article should be kept, since they can describe some favorite cases like an interaction. In addition, among the verbs of an article, "to be" and "to have" verbs appear in different forms which lack useful information. Thus, we decided to remove the various forms of the verbs "to be" and "to have". In an article, the words that have more frequent occurrences, are often not interested, because, most of them are personal pronouns, possessive pronouns, possessive, conjunctions, adverbs or common English words. So we consider the words which have the most frequent occurrences and are not part of the above words. Therefore, the author of the article repeats the words which represent the main idea of the article in the body or the abstract of the article. Considering that the words in the title of the article represent the content of the article, we give more weight to the words of the title. To eliminate common English words, in the reference #4, they have been employed the TF.IDF family of metrics, a well-known term weighting scheme. The reference set used was the British National Corpus (BNC) collection. Terms that appear frequently in a document (TF= Term Frequency), but rarely in the reference set (IDF= Inverse Document Frequency) are more likely to be specific to the document. Terms with a high TF*IDF value or absent from the BNC collection are retained for further processing. Common English words are thus eliminated. Computing TF*IDF criteria for all the terms in each article requires a word by word analysis of the article. This process is applicable to eliminate common English words which are not of our favor. However, the consuming and processing costs are too high. Using a POS (part of speech) tagger, we have done the process at lower costs and similar results. A POS tagger, indeed, follows the whole body of the article, then for each individual term in the article a specific tag is selected. These tags identify different components of the article and the components include nouns, pronouns, verbs, adverbs, possessives and etc. Using a simple parser based on Penn Treebank tag set, we separate our intended set of terms which include: nouns, verbs and adjectives. Therefore, we can ensure that less useful common English words which are not of our interest such as pronouns, possessives and adverbs have been eliminated. Common English words like nouns and adjectives, surely, exist in the texts, which are not interesting. Therefore, with regard to their low occurrences, we can eliminate the words which occur less than the minimum value by determining a minimum value.

2.2. The Extracted Relations

In the previous section, six concepts were extracted from each article which are used to construct ontology. Considering that we used statistical analysis method to extract concepts' data related to query term, there is a statistical relationship between each query term and related concepts. However, we are seeking structural and conceptual relationships like IS-A and Alias. Given the fact that structural and conceptual relations between the words are closely related to semantic relations between the words, we are looking for specific structural patterns to obtain these types of relations[1].

IS-Arelations: IS-A relations are extracted in two ways. First, we search glossary definitions for the query term. Since each query term includes a number of words, so for each word we find the entry definition in the glossary. In the entry, we look for common patterns of IS-A relation represented in reference #1. Some of these patterns are as follows:

- $noun_0$wich …
- a {kind | type | category} of $noun_0$ …
- a {term | concept} { [used] to verb|for verb-ing} …

These relations are of supper class type.

Second, we search structural patterns for query term throughout the glossary and texts retrieved from PubMed search engine. Some of these patterns are as follows:

- $noun_1$ is {a|the} $noun_0$
- $noun_1$ is a term { |used| to verb|for verb-ing} $noun_0$
- such $noun_0$ as $noun_1$, $noun_2$, …, {and|or} $noun_n$
- $noun_0$ {including|especially} $noun_1$, ..., {and|or} $noun_n$
- $noun_1$, $noun_2$, …, $noun_n$, …, {and|or} other $noun_0$
- $noun_0$ except $noun_1$, $noun_2$, …, {and|or} $noun_n$
- $noun_0$, for example $noun_1$, $noun_2$, …, {and|or} $noun_n$

These relations are of sub-class type.

Alias: this relation specifies alternative names for a concept. Abbreviations are the most common examples of this relation. We are looking for Alias relations with the following pattern which are represented in reference #1:

- Zooming, formerly known as 311C90
- 3, 4-methylenedioxymethamphetamine (also known as "Ecstasy")

2.3. The Algorithm Description

Initially, in the first phase the main words of articles are extracted (6 words for each article). These words are just nouns, adjectives or verbs, because, before tagging the article, unwanted words like possessives and etc have been eliminated.

In the second phase, a number of occurrences of the six words in the whole articles are considered. Of these main concepts, then, new concepts are derived. These concepts can be unique words or several words related to each other. Eventually, these new concepts are used to construct clusters. The words with the most frequent occurrences are used to construct new concepts, assuming that the words which have the most frequent occurrences in the whole articles render a fine concept to define cluster. In the next phase, we extract related words and the type of relations from the dictionary and related articles to the query term by using a syntactic parser and according to syntactic patterns which have been presented in the previous section.

3. Testing and Evaluation of Automatic ontology construction

The tool have been tested with a number of query terms. On average, the tool extracted 10 percent of concepts and 25 percent of relations for the concepts throughcomparing it with GO ontology. In fact, the efficiency of the tool in constructing ontology is determined to be 2.5 percent.

Conclusion

The system is able to grow, i.e. connecting to the internet, the system is capable of updating and developing its database. With respect to the results obtained with a small number of articles, higher efficiency can be achived by increasing the number of articles. To evaluate system performance, we just compared the results with GO which is a manually constructed ontology. More reliable assessment could be found if there exists the possibility of comparing results with other manually constructed ontologies. This system does not support any particular ontology language like OWL either.

REFRENCES

1. Youngia Park, Roy J .Byrd and Branimir K. Boguraev, "Towards Ontologies On Demand", Proceedings ofWorkshop on Semantic Web Technologies for Scientific Search and Information Retrieval, 2004.

2. M He, NR Jennings, H Leung, "Ontology Languages for the Semantic Web", IEEE Transactions on Knowledgeand Data Engineering, 2003.

3. Ono, T., Hishigaki, H., Tanigami, A., Takagi, T. , "Automated extraction of information on protein-protein

4. Interactions from the biological literature", Bioinformatics, Vol 17, No. 2, pages 155-6, 2001.Iliopoulos, I., Enright, A.J., Ouzounis, C.A. , "Textquest: Document Clustering Of Medline Abstracts ForConcept Discovery In Molecular Biology", In Proceedings of the Sixth Annual Pacific Symposium onBiocomputing (PSB 001), 2001, pages 384-395, 2001.

5. V¨olker, J.; Vrande˘ci´c, D.; and Sure, Y., "Data-driven change discovery", evaluation. Technical report,Universit ¨at Karlsruhe. SEKT Deliverable 3.3.2, 2006.

6. F. D'Antonio, M. Missikoff, F. Taglino, "Formalizing the OPAL eBusiness ontology design patterns with OWL",in: Third Interna- tional Conference on Interoperability for Enterprise Applications and Software, I-ESA, 2007.

7. Dahab, M. Y. Hassan, H., and Rafea, A.., "TextOntoEx: Automatic ontology construction from natural Englishtext", Expert Systems with Applications, doi:10.1016/j.eswa.2007.01.043, 2007.

8. El-Beltagy, S.R., Maryam, H., and Rafea, "Ontology Based Annotation of Text Segments", To appear inProceedings of the 2007 ACM symposium on Applied computing, Seoul, Korea, 2007.

9. BassemMakni, KhaledKhelif, Rose Dieng-Kuntz and HaceneCherfi, "Semi-automatic Construction of anOntology and of Semantic Annotations from a Discussion Forum of a Community of Practice", KnowledgeEngineering: Practice and Patterns Springer Berlin / Heidelberg, 2008.

10. Iain Coleman, "Saving lives with the Semantic Web", NeSC News, January/February 2009.

11. AntonioDeNicola, MicheleMissikoff, RobertoNavigli, "A software engineering approach to ontology building", Information Systems 34, 2009.

12. Shamsfard, Mehrnoush. " Designing an ontology learning model", PhD thesis, Computer Engineering, Department of computer engineering, Amirkabir University of Technology, 2002.