Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 2, February 2015, pg.14 – 17

RESEARCH ARTICLE

Different Resources of Taxonomy Building Methodology in Web Mining

Geetanjali Panda[1], Bhavani Shankar Panda[2], B.Giridhar[3]

[1]M.Tech in CSE, Utkal University, Odisha, India

[2,3]Asst.Prof in CSE Dept, Centurion University, Odisha

[2] chintupanda@gmail.com; [3] giridhar.bcse@gmail.com

Abstract—Web Mining has become an important research area. Web Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. In this paper, a Survey of Web Mining techniques and applications have been s presented. The obtained results can be very useful for easing the access to web resources of any medical domain or creating ontological representations

Keywords— Web Mining, ontological, taxonomy, Lexocations, Disease domain

I. INTRODUCTION

Web is one of the central largest sources of data in the world. However, the extraction of information from web is a difficult task due to its unstructured definition, its un-trusted sources and its dynamically changing nature. The number of health information websites and online services is increasing day by day. Some of the services provided by these sites are: information about symptoms and cause of any disease, information regarding the cure and medicine for the particular disease, information about the practitioners and specialists for a particular disease and information about the health care centers and hospitals in particular area. Much more information may also be provided by these websites in addition to above.

The World Wide Web is an invaluable tool for researchers, information engineers, health care companies and practitioners for retrieving knowledge. However, the ex-traction of information from web resources is a difficult task due to their unstructured definition, their entrusted sources and their dynamically changing nature. So, in this paper

We present a methodology to extract knowledge from the Web to build automatically a taxonomy of concepts and web resources for a given domain. Moreover, if named entities for a discovered concept are found, they are considered as instances. During the building process, the most representative web sites for each subclass or

instance are retrieved and categorized according to the specific topic cov-ered. With the final hierarchy, an algorithm for discovering different lexicalizations and synonyms of the domain keywords is also performed to find lexicalizations or synonyms, which can be used to widen the search. A prototype has been implemented and tested in several knowledge areas (results for the Disease domain are included).

We base our proposal is the redundancy of information that characterizes the Web, allowing us to detect important concepts and relationships for a domain through a statistical analysis. Moreover, the exhaustive use of web search engines represents a great help for selecting "representative" resources and getting global statistics of concepts; they can be considered as our particular "experts" in all domains. The rest of the paper is organized as follows: section 2 describes the methodology developed to build the taxonomy and classify web sites. Section 3 describes a new approach for discovering lexicalizations and synonyms of terms. Section 4 explains the way of representing the results and discusses the evaluation with respect to other systems. The final section contains the conclusions and proposes lines of future work.

II. TAXONOMY BUILDING METHODOLOGY

The algorithm is based on analyzing a significant number of web sites in order to find important concepts by studying the neighbourhood of an initial keyword. Concretely, in the English language, the immediate anterior word for a keyword is frequently classifying it (expressing a semantic specialization of the meaning), whereas the im-mediate posterior one represents the domain where it is applied [5]. More concretely, it queries a web search engine with an initial keyword (e.g. dis-ease) and analyses the obtained web sites to retrieve the previous (e.g. heart disease) and posterior words (e.g. disease treatment) for the specific keyword. Previous words are analyzed to distinguish if they will be considered as future subclasses or concrete instances into the taxonomy. In order to perform this differentiation, we take into consideration the way in which named entities are presented: English language distinguishes named entities through capitalization [5]. Concretely, the method analyses the first web sites returned by the searchengine when querying the instance candidates are processed to count the number of times that they are presented with upper and lower letters. Those which are presented in capital letters in most cases will be selected as instances (see some examples in Table 1).

Table 1. Examples of instances found for several classes of the obtained taxonomy for the *Disease* domain (100 web documents evaluated for each candidate).

Class	Instance	Full name	Conf.
Chronic	Wisconsin	*Wisconsin Chronic Disease Program*	92.59
Lyme	American	*American Lyme Disease Foundation*	81.69
Lyme	California	*California Lyme Disease Association*	83.33
Lyme	Connecticut	*Connecticut Lyme Disease Coalition*	100.0
Lyme	Yale	*Yale University Lyme Disease Clinic*	87.23
Mitochondrial	European	*European Mitochondrial Disease Network*	100.0
Parkinson	American	*American Parkinson Disease Association*	87.5

Regarding to subclass candidates, a statistical analysis is performed, taking into consideration their relevance in the whole Web for the specific domain. Concretely, the web search engine is queried with the new concepts and relationships found (e.g. "heart disease") and the estimated amount of hits returned is evaluated in relation to the initial keyword measure (e.g. disease). This web scale relevancy statistic combined with other local values obtained from the particular analysis of individual web resources are considered to obtain a relevance measure for each candidate and reject extreme cases (common words or misspelled ones). Only the most relevant candidates are finally selected (see Fig. 1). The selection threshold is automatically computed in function on the domain's generality itself. For each new subclass, the process is repeated recursively to create deeper-level subclasses (e.g. coronary heart disease). So, at each iteration, a new and more specific set of relevant documents for the subdo-main is retrieved. Each final component of the taxonomy (classes and instances) stores the set of URLs from where they have been selected.

Regarding to the posterior words of the keyword they are used to categorize the set of URLs associated to each class. For example, if we find that for a URL associated to the class heart disease, this keyword is followed by the word prevention, the URL will be categorized with this domain of application.

When dealing with polysemic domains (e.g. virus), a semantic disambiguation algorithm is performed in order to group the obtained subclasses according to the specific word sense. The methodology performs a cauterization process of those classes depending on the amount of coincidences between their associated URLs sets. Finally, a refinement is performed to obtain a more compact taxonomy: classes and subclasses that share a high amount of their URLs are merged because we consider them as closely related. For example, the hierarchy "jansky->bielschowsky"results in "jansky_bielschowsky", discovering automatically a multiwordterm.

III. LEXICALIZATIONS AND SYNONYMS DISCOVERY

A very common problem of keyword-based web indexing is the use of different names to refer to the same entity (lexicalizations and synonyms). In fact, the goal of a web search engine is to retrieve relevant pages for a given topic determined by a keyword, but if a text doesn't contain this specific word with the same spelling as specified it will be ignored. So, in some cases, a considerable amount of relevant resources are omitted. In this sense, not only the different morphological forms of a given keyword are important, but also synonyms and aliases.

We have developed a novel methodology for discovering lexicalizations and synonyms using the taxonomy obtained and, again, a web search engine. Our approach is based on considering the longest branches (e.g. atherosclerotic coronary heart disease) of our hierarchy as a contextualization constraint and using them as the search query omitting the initial keyword (e.g. "atherosclerotic coronary heart") for obtaining new webs. In some cases, those documents will contain equivalent words for the main keyword just behind the searched query that can be candidates for lexicalizations or synonyms (e.g. atherosclerotic coronary heart disorder). From the list of candidates, those that have been obtained from a significant amount of multiword terms are selected as they are commonly used in the particular domain. For the disease domain, some alternative forms discovered are: disease(s), disorder(s), syndrome and infection(s).

IV. ONTOLOGY REPRESENTATION AND EVALUATION

The final hierarchy is stored in a standard representation language: OWL. The Web Ontology Language is a semantic markup language for publishing and sharing ontologies on the World Wide Web [2]. Regarding to the evaluation of taxonomies, for the Disease domain, we have per-formed an evaluation of the first level of the taxonomy for different sizes of the search computing, in each case, the precision and the number of correct results ob-tained. These measures have been compared to the ones obtained by a human-made directory (Yahoo), and automatic classifications performed by web search engines (Clusty and AlltheWeb) as shown in Fig. 3. Comparing to Yahoo, we see that al-though its precision is the highest, as it has been made by humans, the number of results are quite limited. The automatic search tools have offered good results in rela-tion to precision, but with a very limited amount of correct concepts obtained.

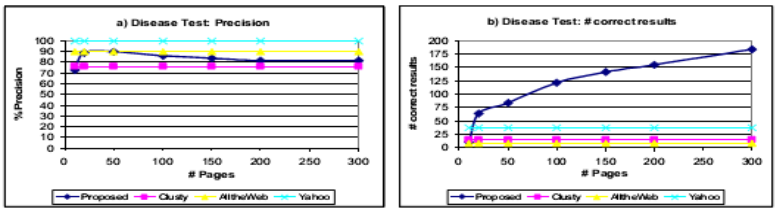

Fig 3. Evaluation of results for the *Disease* domain for different sizes of web corpus.

For the named-entities discovered, we have compared our results to the ones returned by a named-entity detection package trained for the English language that is able to detect with high accuracy some word patterns like organisations, persons, and locations. For illustrative purposes, for the Cancer, and Disease domains, we have obtained a precision of 100% and 91% respectively.

V. RELATED WORK, CONCLUSIONS AND FUTURE WORK

Several authors have been working in the discovering of taxonomical relationships from texts and more concretely from the Web. In [1, 6] authors propose approaches that rely in previously obtained knowledge and semantic repositories like. More closely related, in [4] authors try to enrich previously obtained semantic structures but using web-scale statistics. Other authors using this approach are [3], that learn taxonomic relationships, and [7], for synonym discovery. In contrast, our proposal does not start from any kind of predefined knowledge and, in consequence, it can be applied over domains that are not typically considered in semantic repositories. Moreover, the automatic and autonomous operation eases the updating of the results in highly evolutionary domains. In addition to the advantage of the hierarchical representation of web sites for easing the accessing to web resources, the taxonomy is a valuable element for building machine processable representations like ontologies. This last point is especially important due to the necessity of ontologies for achieving interoperability in many knowledge intensive environments like the Semantic Web [1].

As future work, we plan to study the discovering of non-taxonomical relationships through the analysis of relevant sentences extracted from web resources (text nug-gets), design forms of automatic evaluation of the results and study the evolution of a domain through several executions of the system in different periods of time to ex-tract high level valuable information from the detected changes in the results.

REFERENCES

[1]. Agirre, E., Ansa, O., Hovy, E., and Martinez, D.: Enriching very large ontologies using the WWW. WorkshoponOntology Construction(ECAI-00). 2000.
[2]. Berners-lee T., Hendler, J., Lassila O.: The semantic web. Available at:http://www.sciam.com72001 /050lissue/0501berners-lee.html
[3]. Cimiano, P. and Staab, S.: Learning byGoogling. SIGKDD, 6(2), pp. 24-33. 2004.
[4]. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A., Shaked, T., Soderland, S.and Weld, D.: Web Scale Information Extraction in KnowItAll. WWW2004, USA. 2004.
[5]. Grefenstette G.: SQLET: Short Query Linguistic Expansion Techniques. In: Information Extraction: A Multidisciplinary Approach toan Emerging Information Technology, vol-ume 1299 of LNAI, chapter 6, 97-114. Springer. SCIE-97. Italy, 1997.
[6]. Navigli, R. and Velardi, P.: Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. In Computational Linguistics, Volume 30, Issue 2. June 2004.
[7]. Turney, P.D.: Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. In Proceed-ings of the Twelfth European Conference on Machine Learning. 2001.

