A HYBRID METHOD FOR DOMAIN ONTOLOGY CONSTRUCTION FROM THE WEB

B. Frikh[1], A. S. Djaanfar[2] and B. Ouhbi[3]
[1]*B.P. 1796 Atlas Fès, Ecole Supérieure de Technologie, Route d'imouzer, Fès, Morocco*
[2]*Laboratoire LISQ, Faculté des Sciences, Dhar El Mahraz, Fès, Morocco*
[3]*Ecole Nationale Supérieure d'Arts et Métiers, Marjane II, Béni M'Hamed, B.P. 4024, Meknès, Morocco*

Keywords: Domain ontology, CHIR-statistic, Mutual information, Hybrid approach, Web mining.

Abstract: This paper describes a hybrid statistical and semantic relationships among model concepts for ontology construction. The implementation of the model, called HCHIRSIM (Hybrid Chir-Statistic and Similarity), can be adapted to any domain ontology learning from the Web. It can be viewed as a combination of information from inference view of concepts by using the CHIR-statistic method and the semantic relationships among concepts from the Web by the mutual information measure. The experiments show that our hybrid approach outperforms both purely statistical and purely semantic relationships among concepts approaches. The successful evaluation of our method with different values of the weighting parameter shows that the proposed approach can effectively construct a cancer domain ontology from unstructured text documents.

1 INTRODUCTION

The goal of a domain ontology is to reduce(or eliminate) the conceptual and terminological confusion among the members of a virtual community of users who need to share electronic documents and information of various kinds. This is achieved by identifying and properly defining a set of relevant concepts that characterize a given application domain. The vision of the World Wide Web as a huge repository of machine-processable information may be realized in different ways. The first one is to relay on the large-scale use of semantic annotations that refer to entities defined in a formal ontologies language such as RDF (RDF, 2004). The other one is to try to automatically "re-construct" the knowledge presented in (unstructured) web documents. Several Web mining and information extraction techniques have been proposed to automate this task (Craven et al., 2000), (Etzioni et al., 2005), (Petasis et al., 2003), (Sanchez and Moreno, 2004), (Frikh et al., 2009). The techniques employed by different systems in ontology construction or ontology learning, may vary depending on the tasks to be accomplished. The techniques can generally be classified into statistics-based, linguistics-based, logic-based, or hybrid. The various statistics-based techniques for accomplishing the tasks in ontology learning are mostly derived from information retrieval, machine learning and data mining. Some of the common techniques include clustering (Wong et al., 2006), latent semantic analysis (Turney, 2001), co-occurrence analysis (Budanitsky, 1999), term subsumption (Fotzo and Gallinari, 2004), contrastive analysis (Velardi et al., 2005) and association rule mining (Strehl, 2002). The main idea behind these techniques is that the extent of occurrence of terms and their contexts in documents often provide reliable estimates about the semantic identity of terms.

Co-occurrence analysis is usually coupled with some measures to determine the association strength between terms or the constituents of terms. Some of the popular measures include dependency measures (e.g. mutual information (Church and Hanks, 1990)), log-likelihood ratios (Resnik, 1999) (e.g. chi-square test), rank correlations (e.g. Pearsons and Spearmans coefficient (Strehl, 2002)), distance measures (e.g. Kullback-Leiber divergence (Maedche et al., 2002)), and similarity measures (e.g. cosine measures (Senellart and Blondel, 2003)).

Some of the common relevance measures from information retrieval include the Term Frequency-Inverse Document Frequency (TF-IDF) (Salton and Buckley, 1988) and its variants, and others based on language modeling (Croft and Ponte, 1998) and probability (Fuhr, 1992).

Frikh B., S. Djaanfar A. and Ouhbi B..
A HYBRID METHOD FOR DOMAIN ONTOLOGY CONSTRUCTION FROM THE WEB.
DOI: 10.5220/0003667502850292
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2011), pages 285-292
ISBN: 978-989-8425-80-5

In these approaches, the assumption normally made is that words that co-occur with sufficient frequency in the documents of a collection are in fact related to each other. Our approach is different in that our technique considered an hybrid approach. This new statistical data can describe the term category dependency more accurately than the statistics used in the paper of Sanchez and Moreno (Sanchez and Moreno, 2003), since CHIR keeps only terms relevant to the categories. The main assumption in their work is that words that are near to the specified keyword are closely related. To relax this hypothesis, we introduce a hybrid statistical and semantic relationships among model concepts for ontology construction. The implementation of the algorithm, called HCHIRSIM (Hybrid chir-statistic similarity), can be adapted to any domain ontology learning from the Web. It extended an earlier work of the authors (Djaanfar et al., 2010) in the sense that it can be viewed as a combination of information from inference view of concepts by using the CHIR-statistic method and the semantic relationships among concepts from the Web by the mutual information measure. The experiments show that our hybrid approach outperforms both purely statistical and purely semantic relationships among concepts approaches. The successful evaluation of our method with different values of the weighting parameter shows that the proposed approach can effectively construct a cancer domain ontology from unstructured text documents. The paper is organized as follows. Based on the CHIR-statistic and the similarity measure, we first propose our hybrid model, give then technical details of the HCHIRSIM-algorithm, and finally present evaluation results for different values of the weighting parameter.

2 DEFINITION OF THE MODEL

Motivated by the performance of text clustering by selecting the words that help to distinguish the documents into different clusters (Li et al., 2008) and the fact that two terms are considered similar if their mutual information with all terms in the vocabulary are nearly the same (cf. (Brun et al., 2002)), we propose a hybrid model that is able to identify and properly define a set of relevant concepts that characterize a given application domain then captures the semantic relationships among concepts from the Web.

2.1 Term Selection based on the CHIR Statistic

To measure the degree of dependency between a term and a specific category, the χ^2 statistic tests the hypothesis that the term and the category are statistically independent of each other. The χ^2 statistic is defined as:

$$\chi^2_{w,c} = \sum_i \sum_j \frac{(O(i,j) - E(i,j))^2}{E(i,j)}, \qquad (1)$$

where $O(i,j)$ is the observed frequency of the documents that belong to category j and contain w. $E(i,j)$ is the expected frequency of category j and term i.
In text mining and information retrieval studies, the χ^2 statistic has been used for feature selection (Salton and Buckley, 1988). As we have said in the introduction the CHIR statistic is an extended variant of the χ^2 statistic for term-category independence test to measure the degree of dependency between a term w and a category C of documents. This method can improve the performance of text clustering by selecting the words that help to distinguish the documents into different clusters (Li et al., 2008). Despite to the χ^2 statistic, the CHIR statistic select only relevant terms that have strong positive dependency on certain categories in the corpus and remove the irrelevant and redundant terms. The new term-category dependency measure $R_{w,c}$ is defined by:

$$R_{w,c} = \frac{O(w,c)}{E(w,c)}, \qquad (2)$$

where $O(w,c)$ is the number of documents that are in the category c and contain the term w and $E(w,c)$ is the expected frequency of the category c to contain the term w. If there is positive dependency then the observed frequency should be larger than 1. If there is negative dependency, $R_{w,c}$ should be smaller than 1. In the case of the no-dependency between the term w and the category c, the term-category dependency measure $R_{w,c}$ should be close to 1. When $R_{w,c}$ is larger than 1, the dependency between w and c is positive, otherwise, the dependency is negative. To get better information about the dependency between a term and a category, (Li et al., 2008) use a combining formula of $\chi^2_{w,c}$ and $R_{w,c}$ and define the term goodness of a term w in a corpus with m classes as:

$$r_{\chi^2}(w) = \sum_{j=1}^{m} p(R_{w,c_j})\chi^2_{w,c_j} \qquad with \qquad R_{w,c_j} > 1, \qquad (3)$$

where

$$p(R_{w,c_j}) = \frac{R_{w,c_j}}{\sum_{j=1}^{m} R_{w,c_j}} \qquad with \qquad R_{w,c_j} > 1, \tag{4}$$

is the weight of χ^2_{w,c_j} in the corpus in terms of R_{w,c_j}. A bigger $r_{\chi^2}(w)$ value indicates that the term is more relevant.

When there is positive dependency between the term w and the category c_j, this new term-goodness measure is the weighted sum of χ^2_{w,c_j}.

2.2 Term Similarity Metric

The most widely measure used, in statistics and information theory, is the mutual information. In mutual information, the co-occurrence frequencies of the constituents of complex terms are utilized to measure their dependency. Two terms are considered similar if their mutual information with all terms in the vocabulary are nearly the same cf. (Brun et al., 2002). In this paper, the metric we use for measuring the similarity of two terms is that given by (Dagan et al., 1999). This similarity measure is defined by:

$$\begin{aligned} Sim(w,w') &= \frac{1}{2|V|}\sum_{i=1}^{|V|}\left(\frac{min(I(z_i,w),I(z_i,w')}{max(I(z_i,w),I(z_i,w')}\right. \\ &+ \left.\frac{min(I(w,z_i),I(w',z_i)}{max(I(w,z_i),I(w',z_i)}\right), \end{aligned} \tag{5}$$

where V is the vocabulary and $I(z_i,w)$ is the mutual information of terms z_i and w. We use it to identify terms which are semantically relevant in a domain ontology construction. This measure is based on the mutual information calculated for a window of d terms. It's nature is essentially semantic than syntactic. The mutual information is defined as follows:

$$I(z_i,w) = P_d(z_i,w) log \frac{P_d(z_i,w)}{d^2 P(z_i)P(w)}, \tag{6}$$

where d is the withdrawal, $P(z_i)$ and $p(w)$ are the a priori probability of term z_i and w respectively. $P(z_i,w)$ is the probability of succession of terms z_i and w in the window observation. This probability can be estimated by ratio of the number of times that z_i is followed by w within the window and by the cardinal of the vocabulary.

$$\hat{P}(z_i,w) = \frac{f_d(z_i,w)}{|V|}, \tag{7}$$

where $f_d(z_i,w)$ is the number of times that z_i is followed by w.

2.3 The Hybrid Proposed Model

The main objective behind the proposed model is to propose a hybrid model that is able to identify and properly define a set of relevant concepts that characterize a given application domain then captures the semantic relationships among concepts from the Web. The basic idea is to reorder the retrieved concepts based on the hybrid model. It comes from the following observations. For a domain application start with a keyword, w_{rep} that has to be representative enough for the domain. The similarity between the initial keyword w_{rep} and a candidate concept in the Web is measured based on the weighting model combining a component estimated from the CHIR-statistic and one from the similarity measure by linear interpolation:

$$S(w) = \lambda * r_{\chi^2}(w) + (1-\lambda) sim(w, w_{rep}), \tag{8}$$

where λ is a weighting parameter between 0 and 1. This can be viewed as a combination of information from inference view of concepts by using the CHIR-statistic method and the semantic relationships among concepts from the Web by the mutual information measure. Concepts with the highest similarities are returned as the retrieval results. Since relevant concepts convey semantically similar information with respect to the initial keyword, it is likely that previous Web pages have already judged them as co-relevant through relevance feedback. Therefore, it is reasonable to assume that concepts having strong score are likely to be relevant concepts to the initial keyword. With each retrieved concepts a new keyword is constructed joining the new concept and the initial one. This process (cf. Figure1) is repeated recursively until a selected depth level is achieved or no more results are found.

3 HCHIRSIM ALGORITHM

In this section, the proposed algorithm used to discover and select representative concepts and websites for a domain and construct the final ontology is described. This algorithm is based on analyzing a large number of web sites in order to find important concepts for a domain by introducing an initial keyword. The candidate concepts are processed in order to select the most adequate ones by performing HCHIRSIM analysis. The selected classes are finally incorporated into the ontology. For each one, the main websites from where it was extracted are stored, and the process is repeated recursively in order to find new terms and build a hierarchy of concepts. This algorithm is described by the following steps:

- Perform a k-means clustering algorithm on the set of all documents and get initial clusters.
- Start with a keyword that has to be representative enough for the domain and a set of parameters that constrain the search and the concept selection (cf.(Djaanfar et al., 2010)).
- Extract all the candidate concepts by analyzing the neighborhood of the initial keyword; select the anterior words and posterior words as candidate concepts.
- For each candidate concept, calculate its score $S(w)$ measure by using (8).
- Sort the terms in descending order of their $S(w)$ measure.
- Select the top l terms from the list.
- The l concepts extracted are incorporated as classes or instances in the taxonomy and the URLs, from where they are extracted, are stored.
- For each concept incorporated in the taxonomy, a new keyword is constructed joining the new concept and the initial one. This process is repeated recursively until a selected depth level is achieved or no more results are found.
- Finally, a refinement process is performed in order to obtain a more compact taxonomy and avoid redundancy.

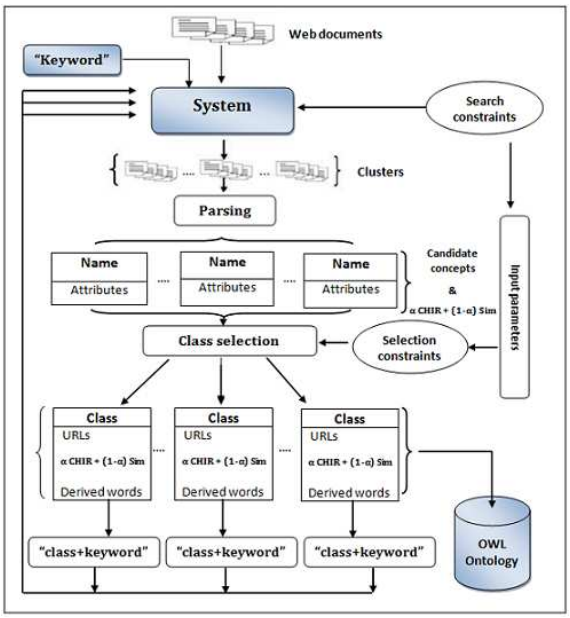

Figure 1: Taxonomy building methodology.

4 ONTOLOGY BUILDING AND REPRESENTATION

Ontologies consist of a set of classes representing the categories of the entities of interest in a domain and the relationships between those entities. In doing so, an ontology can be used to capture what it means to be one of those entities; that is, the semantics of the domain. In our study, we use a number of tools for ontology development. These include: the Web Ontology Language(OWL), a commonly used language for expressing ontologies, and Protégé-OWL, a leading ontology engineering environment, Lucene 3.0.1(a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform), HTML Parser 2.0, Stemmers 1.0(suitable for the morphologic root in English language), Jena 2.6.3(Framework Java).

4.1 Ontology Representation

The final ontology is edited by ”Protégé 4.1” in order to evaluate its precision then it is saved in OWL format. Its content is entirely written in the standard OWL language representation (OWL, 2004). The final domain ontology (the vocabulary of a domain and a set of constraints on how terms can be combined to model the domain), is presented to the user in a refined way and can be interrogated by the user.

4.2 Application

As an illustration application, we choose to use ”cancer” as the initial keyword. The program was executed on a collection of 52758 documents, indexed from 26 web sites of this domain. The maximum depth level has been fixed to 5. For each query, we select the top 11 terms from the list. The 11 concepts extracted are incorporated as classes or instances in the taxonomy and the URLs from where they are extracted are stored. The length of the window is fixed to 4. The resulting taxonomy is formally correct. We present in Figure 2, a part of the obtained ontology. The numbers represent the number of instances in the classes.

5 MEASUREMENT METHODS AND RESULTS

Due to the complex nature of ontologies, evaluation approaches can also be distinguished by the layers of

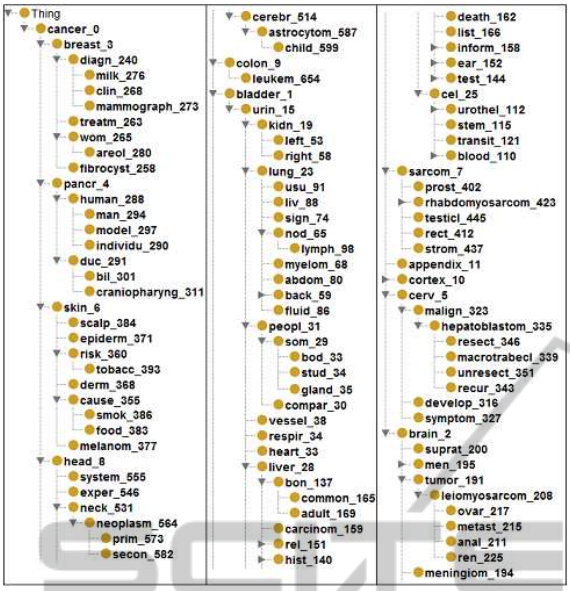

Figure 2: Cancer taxonomy visualized on Protégé 4.1: numbers are class identifiers.

an ontology (e.g. term, concept, relation) they evaluate (Porzel and Malaka, 2004). More specifically, evaluations can be performed to assess the (1) correctness at the terminology layer, (2) coverage at the conceptual layer, (3) wellness at taxonomy layer, and (4) adequacy of the non-taxonomic relations. The focus of evaluation at the terminology layer is to determine if the terms used to identify domain-relevant concepts are included and correct. Some form of lexical reference or benchmark is typically required for evaluation in this layer. Typical precision and recall measures from information retrieval are used together with exact matching or edit distance (Maedche and Staab, 2002) to determine performance at the terminology layer. The lexical precision and recall reflect how good the extracted terms cover the target domain. Lexical Recall (LR) measures the number of relevant terms extracted $e_{relevant}$ divided by the total number of relevant terms in the benchmark $b_{relevant}$, while Lexical Precision (LP) measures the number of relevant terms extracted $e_{relevant}$ divided by the total number of terms extracted e_{all}. LR and LP are defined as (Sabou et al., 2005):

$$LP = \frac{e_{relevant}}{e_{all}}, \tag{9}$$

$$LR = \frac{e_{relevant}}{b_{relevant}}, \tag{10}$$

The precision and recall measure can be also combined to compute the corresponding F_β-score. The general formula for a non-negative real β is:

$$F_\beta = \frac{(1+\beta^2)precision \times recall}{\beta^2 \times precision + recall}, \tag{11}$$

Evaluation measures at the conceptual level are concerned with whether the desired domain-relevant concepts are discovered or otherwise. Lexical Overlap (LO) measures the intersection between the discovered concepts (Cd) and the recommended concepts (Cm). LO is defined as:

$$LO = \frac{|\,Cd \bigcap Cm\,|}{|\,Cm\,|}, \tag{12}$$

Ontological Improvement (OI) and Ontological Loss (OL) are two additional measures to account for newly discovered concepts that are absent from the benchmark, and for concepts which exist in the benchmark but were not discovered, respectively. They are defined as (Sabou et al., 2005):

$$OI = \frac{|\,Cd - Cm\,|}{|\,Cm\,|}, \tag{13}$$

$$OL = \frac{|\,Cm - Cd\,|}{|\,Cm\,|}, \tag{14}$$

Evaluations at the taxonomy layer is more complicated. Performance measures for the taxonomy layer are typically divided into local and global. The similarity of the concepts positions in the learned taxonomy and in the benchmark is used to compute the local measure. The global measure is then derived by averaging the local scores for all concept pairs. One of the few measures for the taxonomy layer is the Taxonomic Overlap (TO) (Maedche and Staab, 2002). The computation of the global similarity between two taxonomies begins with the local overlap of their individual terms. The semantic cotopy, the set of all super- and sub-concepts, of a term varies depending on the taxonomy. The local similarity between two taxonomies given a particular term is determined based on the overlap of the terms semantic cotopy. The global taxonomic overlap is then defined as the average of the local overlaps of all the terms in the two taxonomies.

In order to evaluate our methodology, we have used the MeSH Browser. "MeSH is the National Library of Medicine's controlled vocabulary thesaurus. It consists of sets of terms naming descriptors in a hierarchical structure that permits searching at various levels of specificity. The MeSH Browser is an online vocabulary look-up aid available for use with MeSH (Medical Subject Headings). It is designed to help quickly locate descriptors of possible interest and to show the hierarchy in which descriptors of interest appear. Virtually complete MeSH records

are available, including the scope notes, annotations, entry vocabulary, history notes, allowable qualifiers, etc. The browser does not link directly to any MIDLINE or other database retrieval system and thus is not a substitute for the PUBMED system. The MeSH Browser points to the newest version of MeSH and so it will also find new Supplementary Concepts as these are added and updated weekly. The MeSH Browser may be used to find descriptors, qualifiers, or Supplementary Concepts of interest and see these in relationship to other concepts. The browser is part of the MeSH Web pages. It finds descriptors of interest without assuming knowledge of the often complex vocabulary structure and rules" (Mesh, 2010).

We have submitted the query "cancer" to the MeSH Browser. For each candidate concept, we have extracted all related entries then stocked it in a text file. To exploit the resulted concepts, we have decomposed them in simple terms. All terms under three characters were rejected. Then we extracted the morphological root of the rest terms. This different extracted morphological roots are then considered as relevant terms for the benchmark lexical. The total, 881 different morphological roots were used to recover the domain cancer. We have then conducted the evaluation. For the number of terms retained in the ontology construction, we use concepts and instances. To evaluate F_β-score, we take β equal to 1. Results are presented in tables Table.1- Table.5.

Table 1: Performance of HCHIRSIM algorithm in terms of precision, recall, lexical overlap, ontological improvement and ontological loss for $\lambda = 0.5$.

$\lambda = 0.5$	MeSH	Documents collection
Selected	881	827
Rejected	463	102
Total	1344	929
LR		93,87%
LP		89,02%
F_β		91,38%
LO		86,83%
OI		7,04%
OL		13,17%

Table 1 summarizes the performance of the HCHIRSIM algorithm for $\lambda = 0.5$ Our new measure achieved a 93,87%, 89,02% for recall, precision, respectively. One will notice that our new algorithm achieved a lexical overlap of 86,83% and ontological improvement of 7,04%. However the ontological loss is 13,17%. This validate the performance of the HCHIRSIM algorithm for $\lambda = 0.5$.

5.1 Experimental Results

In order to determine the best value for the weighting parameter of our HCHIRSIM model, we have conduct experiments, under the same conditions as in the previous subsection, on the values of the weighting parameter λ. Concretely, we use $\lambda \in \{0, 0.2, 0.5, 0.8, 1\}$. Results are presented in tables $1-5$:

Table 2: Performance of HCHIRSIM algorithm in terms of precision, recall, lexical overlap, ontological improvement and ontological loss for $\lambda = 0$.

$\lambda = 0$	MeSH	Documents collection
Selected	881	614
Rejected	463	213
Total	1344	827
LR		69,69%
LP		74,24%
F_β		71,90%
LO		66,97%
OI		2,72%
OL		33,03%

Table 3: Performance of HCHIRSIM algorithm in terms of precision, recall, lexical overlap, ontological improvement and ontological loss for $\lambda = 0.2$.

$\lambda = 0.2$	MeSH	Documents collection
Selected	881	809
Rejected	463	113
Total	1344	922
LR		91,83%
LP		87,74%
F_β		89,74%
LO		85,36%
OI		6,47%
OL		14,64%

Table 4: Performance of HCHIRSIM algorithm in terms of precision, recall, lexical overlap, ontological improvement and ontological loss for $\lambda = 0.8$.

$\lambda = 0.8$	MeSH	Documents collection
Selected	881	796
Rejected	463	117
Total	1344	913
LR		90,35%
LP		87,19%
F_β		88,74%
LO		84,34%
OI		6,02%
OL		15,66%

Table 5: Performance of HCHIRSIM algorithm in terms of precision, recall, lexical overlap, ontological improvement and ontological loss for $\lambda = 1$.

$\lambda = 1$	MeSH	Documents collection
Selected	881	528
Rejected	463	256
Total	1344	784
LR		59,93%
LP		67,35%
F_β		63,42%
LO		57,89%
OI		2,04%
OL		42,11%

5.2 Discussion

Taking $\lambda = 1$ corresponds to the Chir-Statistic technique which is an improved version of the chi-statistic test (Resnik, 1999). While taking $\lambda = 0$ corresponds to the mutual information dependency measure (Church and Hanks, 1990). Instead of only exploiting mutual information measure or only using statistic technique or sequential method given by (Djaanfar et al., 2010) which is time consuming, we use a hybrid measure. The performance evaluation is conducted by setting different weights. The experimental results are presented in tables 1-5. In general, the hybrid method yields better performance than statistics-based or linguistics-based. Moreover, the performance of the hybrid method for the weight $\lambda = 0.5$ ($LR = 93,87\%, LP = 89,02\%, F_\beta = 91,38\%, LO = 86,83\%, OI = 7,04\%, OL = 13,17\%$) is much higher than that of the other weights.

6 CONCLUSIONS

This paper presented a hybrid method combining statistical and semantic approaches for automating the ontology construction process by retrieving and extracting data from Web resources. The obtained algorithm called HCHIRSIM can be adapted to any domain ontology learning from the Web. The experiments show that our hybrid approach outperforms both purely statistical and purely semantic relationships among concepts approaches. The successful evaluation of our method with different values of the weighting parameter shows that the proposed approach can effectively construct a cancer domain ontology from unstructured text documents.

REFERENCES

Brun, A., Smaili, K., and Haton, J.-P. (2002). Wsim : une mthode de dtection de thme fonde sur la similarit entre mots. In *9me conf. fran. TALN'2002, Nancy, France*.

Budanitsky, A. (1999). Lexical semantic relatedness and its application in natural language processing. In *Technical Report CSRG-390*. Computer Systems Research Group, University of Toronto.

Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicograph. *Computational Linguistics*, 16(1).

Craven, M., Dipasquo, D., Freitag, D., McCallum, A., Mitchell, T., Nigam, K., and Slattery (2000). Learning to construct knowledge bases from the world wide web. *Artificial Intelligence*, 118(1):69–113.

Croft, B. and Ponte, J. (1998). A language modeling approach to information retrieval. In *21st International Conference on Research and Development in Information Retrieval*.

Dagan, I., Lee, L., and Pereira, F. C. N. (1999). Similarity-based models of word co-occurrence probabilities. *Machine Learning*, 34:43–69.

Djaanfar, A. S., Frikh, B., and Ouhbi, B. (2010). A domain ontology learning from the web. In *M. Saadi et al. (eds), Studies in Comp. Intel., Vol(315), 201-208,*. Springer-Verlag.

Etzioni, O., Cafarella, M., Downey, D., Popescu, A., Shaked, T., Soderland, S., , Weld, D., and Yates, A. (2005). Unsupervised named-entity extraction from the web: an experimental study. *Artificial Intelligence*, 165(1):91–134.

Fotzo, H. and Gallinari, P. (2004). Learning generalization/specialization relations between concepts application for automatically building thematic document hierarchies. In *The 7th International Conference on Computer-Assisted Information Retrieval (RIAO)*. RIAO Vaucluse, France.

Frikh, B., Djaanfar, A. S., and Ouhbi, B. (2009). An intelligent surfer model combining web contents and links based on simultaneous multiple-term query. In *The seventh ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-2009), IEEE Computer Society*.

Fuhr, N. (1992). Probabilistic models in information retrieval. *The Computer Journal*, 35(3):243–255.

Li, Y., Luo, C., and Chung, S. M. (2008). Text clustering with feature selection by using statistical data knowledge and data engineering. *IEEE Transactions on Know and Data Eng.*, 20(5):641–651.

Maedche, A., Pekar, V., and Staab, S. (2002). *Ontology learning part one-on discovering taxonomic relations from the web*. Springer-Verlag.

Maedche, A. and Staab, S. (2002). Measuring similarity between ontologies. In *European Conference on Knowledge Acquisition and Management (EKAW), Madrid, Spain*.

Mesh (2010). *Medical Subject Headings*. National Library of Medicine's controlled vocabulary thesaurus.

OWL (2004). *Web Ontology Language*. W3C Recommendation 10 February.

Petasis, G., Karkaletsis, V., and Spyropoulos, C. (2003). Cross-lingual information extraction from web pages: The use of a general-puepose text engineering platform. In *the 4th International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria*.

Porzel, R. and Malaka, R. (2004). A task-based approach for ontology evaluation. In *The 16th European Conference on Artificial Intelligence*. Valencia, Spain.

RDF (2004). *Resource Description Framework*. W3C Recommendation 10 February.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11(1):95–130.

Sabou, M., Wroe, C., Goble, C., and Mishne, G. (2005). Learning domain ontologies for web service descriptions: an experiment in bio-informatics. In *The 14 International Conference on World Wide Web*.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Sanchez, D. and Moreno, A. (2003). Web-scale taxonomy learning. In *Tech. Rep. of Dep. Computer Science and Mathematics*. University Rovira i Virgili.

Sanchez, D. and Moreno, A. (2004). Creating ontologies from web documents. *Recent Advances in Artificial Intelligence Research and Development*, 113:11–18.

Senellart, P. and Blondel, V. (2003). *Automatic discovery of similar words*. Springer-Verlag.

Strehl, A. (2002). *Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining*. PhD thesis, University of Texas at Austin.

Turney, P. (2001). Mining the web for synonyms: Pmi-ir versus lsa on toefl. In *The 12th European Conference on Machine Learning*. ECML, Freiburg, Germany.

Velardi, P., Navigli, R., Cucchiarelli, A., and Neri, F. (2005). *Evaluation of ontolearn, a methodology for automatic learning of ontologies*. IOS Press.

Wong, W., Liu, W., and Bennamoun, M. (2006). Featureless similarities for terms clustering using tree-traversing ants. In *The International Symposium on Practical Cognitive Agents and Robots*. (PCAR), Perth, Australia.