2007 IEEE/WIC/ACM International Conference on Web Intelligence

Discovering Subsumption Hierarchies of Ontology Concepts from Text Corpora

Elias Zavitsanos[1,2]
izavits@iit.demokritos.gr

Georgios Paliouras[1]
paliourg@iit.demokritos.gr

George A. Vouros[2]
georgev@aegean.gr

Sergios Petridis[1]
petridis@iit.demokritos.gr

[1] Institute of Informatics and Telecommunications, NCSR "Demokritos", Greece
[2] Department of Information and Communication Systems Engineering, AI-Lab
University of Aegean, Greece

Abstract

This paper proposes a method for learning ontologies given a corpus of text documents. The method identifies concepts in documents and organizes them into a subsumption hierarchy, without presupposing the existence of a seed ontology. The method uncovers latent topics in terms of which document text is being generated. These topics form the concepts of the new ontology. This is done in a language neutral way, using probabilistic space reduction techniques over the original term space of the corpus. Given multiple sets of concepts (latent topics) being discovered, the proposed method constructs a subsumption hierarchy by performing conditional independence tests among pairs of latent topics, given a third one. The paper provides experimental results over the GENIA corpus from the domain of biomedicine.

1 Introduction

Ontologies have been proposed as the key element to shape, manage and further process information. However, the engineering of ontologies is a costly, time-consuming and error-prone task when done manually. Furthermore, in fast evolving domains of knowledge, or in cases where information is constantly being updated, making prior knowledge obsolete, the continuous maintenance and evolution of ontologies are tasks that require significant human effort. Thus, there is a strong need to automate the ontology development/maintenance in order to boost the rapid development/update of ontologies and to minimize the cost of their creation and evolution. For this reason, ontology learning has become an emerging field of research, aiming to help knowledge engineers to build and further extend ontologies with the help of automatic or semi-automatic machine learning techniques, exploiting several sources of information.

Ontology learning is commonly viewed ([3], [6], [18], [21]) as the task of *extending* or *enriching* an existing ontology with new ontology elements mined from text corpora. Depending on the ontology elements being discovered, existing approaches deal with the identification of concepts, subsumption relations among concepts, instances of concepts, or of concept properties/relations. However, the majority of the existing approaches need a seed ontology, i.e. a backbone or a generic ontology, that formalizes some of the concepts in a corpus. We may further classify existing ontology learning approaches to be either of the linguistic, statistical, or machine learning type, depending on the specific techniques that they use. Most of the existing approaches depend on the language of the corpus, using, for instance, language-dependent lexico-syntactic patterns.

In contrast to the majority of the existing work on ontology enrichment, this paper proposes an automated statistical approach to ontology learning, without presupposing the existence of a seed ontology, or any other type of external resource, other than a corpus of text documents. The proposed method tackles both the concept identification and the subsumption hierarchy construction tasks: Specifically, concepts are identified and represented as multinomial distributions over terms in documents[1]. Towards this objective, the method uses the Markov Chain Monte Carlo (MCMC) process of Gibbs sampling [10], following the Latent Dirichlet Allocation (LDA) [5] model. To discover the subsumption relations between the identified concepts, the method uses conditional independence tests among the dis-

[1] By "terms" we do not mean domain terms, but the words that will constitute the vocabulary over which concepts will be specified. In the following we use "terms" and "words" interchangeably.

0-7695-3026-5/07 $25.00 © 2007 IEEE
DOI 10.1109/WI.2007.55

covered concepts. The statistical nature of this approach assures the language-independence of the proposed method.

In what follows, section 2 states the problem, refers to existing approaches that are closely-related to the proposed method, and motivates our approach. Section 3 provides background information concerning probabilistic topic models and the LDA model. Section 4 describes the proposed method and section 5 presents experiments and evaluation results. Finally, section 6 concludes the paper by pointing out the advantages and drawbacks of the proposed method, sketching plans for future work.

2 Problem Definition and State of the Art

2.1 Problem Definition

It is widely agreed that an ontology is a formal specification of a conceptualization of a domain. Ontology elements comprise concepts, individuals and properties. In this paper we deal with concepts and the subsumption relation among concepts.

Ontology learning deals with discovering and acquiring new ontology elements, and integrating them in an existing ontology in a consistent and coherent way. The objective is to facilitate seamless to other human activities, rapid, effective (in terms of precision and recall), and low-cost ontology evolution. Specifically, given some sources of information (usually text corpora), the learning task aims to identify concepts, properties and/or individuals that capture knowledge in a specific domain. Furthermore, by exploiting sources of information available, a learning method may also aim to discover relations among ontology elements.

In this paper, we deal with two major problems related to the ontology learning task: (1) The discovery of the concepts in a corpus, and (2) The ordering of the discovered concepts by means of the subsumption relation. Specifically, assuming only the existence of a corpus of text documents, this paper aims to answer the following questions:

(1) Is it possible to discover the concepts that express the content of documents in the corpus, independently of the terms' surface appearance?

(2) Is it possible to form the ontology subsumption hierarchy backbone, using only statistical information concerning the discovered concepts?

(3) Is it possible to devise a language-neutral ontology learning method?

2.2 Concept Identification

Aiming at ontology learning from texts, many approaches use statistical techniques for identifying concepts. The work in [6] apply statistical analysis on web pages in order to identify words, which are then grouped into clusters that are proposed to the knowledge engineer. In this case, the ontology enrichment task is based on statistical information of word usage in the corpus and the structure of the original ontology.

The authors in [4] extend an ontology with new concepts, taking into account words that co-occur with each one of the existing concepts. The method requires that there are several occurrences of the concepts to be classified, so that there is enough contextual information to generate topic signatures. The work reported in [3] follows similar research directions.

Regarding the linguistic techniques for concept identification, the use of pattern matching on noun phrases ([14], [16]) is a widely-used approach that matches regular expressions with part-of-speech (POS) tags in order to derive noun phrases that indicate possible concepts. In addition, the internal structure of words can be exploited in order to identify domain-specific terms ([12]). Small domain-specific units (i.e. morphemes or suffices) can also indicate terms related to the domain of interest. These techniques require text preprocessing and depend heavily on linguistic aspects. Thus, their major drawback is language dependence.

All of these approaches assume that the surface appearance of terms (or the patterns of term appearances) in documents provide sufficient information for concept discovery. However, we aim to uncover latent topics in the corpus, emphasizing on the generative process of documents. Doing so, we assume that latent topics correspond to ontology concepts.

Concerning the concept identification task from the perspective of statistical and machine learning techniques, the TF/IDF [19] weighting scheme has been used in conjunction with Latent Semantic Indexing (LSI) [7] towards revealing latent topics in a corpus of documents. A classification task assigns words to topics making each topic a distribution over words. Probabilistic Latent Semantic Indexing (PLSI) [13] extends LSI assuming that each document is a probability distribution over topics and each topic is a probability distribution over words. Although PLSI outperforms LSI, it must be pointed out that this model is prone to overfitting (being corpus specific), involving a large number of parameters that need to be estimated [5]. This number grows linearly with the size of corpus: PLSI treats the weights of topic contributions to expressing the content of each document as a set of individual parameters that are explicitly linked to the corpus.

Similarly to these approaches, we aim to uncover latent topics (represented as probability distributions over terms) that mediate knowledge on documents' contents. This approach is based on the assumption that the topics represent ontology concepts. Towards this target, we improve on previous approaches towards avoiding overfitting and large sets of parameters, by using the Latent Dirichlet Allocation model.

2.3 Taxonomy Construction

Linguistic approaches regarding the taxonomy construction task usually apply lexico-syntactic patterns (Hearst patterns [11] are the most widely used) on text corpora. These patterns are of the form $NPsuchasNP, NP, .., andNP$ and aim to find subsumption relations between noun phrases that usually serve as concepts in an ontology. Thus, in addition to the need of a seed ontology, these approaches are limited by the fact that these patterns do not occur frequently enough in texts: this may result to low recall of subsumption relations.

Moving towards machine learning and statistical methods, an extension of PLSI, named Hierarchical Probabilistic Latent Semantic Analysis (HPLSA) has been used in [8], in order to acquire a hierarchy of topics, by enabling data to be hierarchically grouped based on its characteristics. Specifically, this approach has been used for document categorization. This does not warantee that the resulting hierarchy will be (or even reflect) a subsumption hierarchy among ontology concepts. In addition, since it has a strong relation with PLSI, the drawbacks of PLSI mentioned in the previous section are inherited by this method.

Hierarchical Latent Semantic Analysis (HLSA) has been applied in [17] to introduce hierarchical dependencies among concepts by exploiting word occurrences among concepts (latent topics). This approach actually computes relations among topics in terms of the contained words. However, such an approach depends heavily on the surface appearance of words.

In this paper we aim to overcome the problem of the linguistic techniques, following a purely statistical approach to subsumption relation discovery. Actually, the proposed approach does not depend on the language or the annotation of the corpus. Instead it uses conditional independence tests on the latent topics discovered, in order to identify subsumption relations. It must also be pointed out that, given the latent topics, the proposed method may compute more than one subsumption hierarchies of no predefined depth.

3 Background on Probabilistic Topic Models

Probabilistic Topic Models (PTMs) [20] are based on the idea that documents are mixtures of topics, where a topic can be thematic and is represented by means of a probability distribution over terms. PTMs follow the bag-of-words assumption, assuming that words are independently and identically distributed in the texts. Topic models are generative models for documents: they specify a probabilistic procedure by which documents are generated. They are based on probabilistic sampling rules that describe how documents are generated as combinations of latent variables (i.e. topics). Figure 1 illustrates the generative process: topics (clouds) are probability distributions over a predefined vocabulary of words (puzzle pieces). According to the probability that a topic participates to the content of each document, the process samples words from the corresponding topic in order to generate the documents.

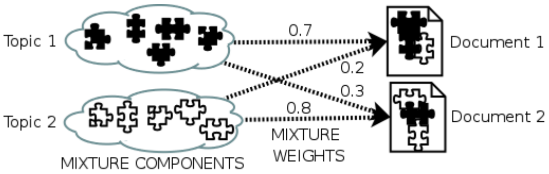

Figure 1. The generative process: Documents are mixtures of topics. Topics are probability distributions over words (puzzle pieces). The probability of participation of a topic in a document is defined by the mixture weights. (Inspired from [20])

In this paper, we are not interested in the generative process per se, but rather in the inverse process. Documents are known and words are observations towards assessing the topics of documents, as combinations of words. For this purpose, we use the LDA model described below.

The probabilistic generative process that is used in LDA states that topics are sampled repeatedly in each document. Specifically, given a predefined number of topics K, then for each document:

1. Choose $N \sim$ Poisson(ξ).

2. Choose $\theta \sim$ Dirichlet(α).

3. For each of the N words w_n:

   - Choose a topic $z_n \sim$ Multinomial(θ)
   - Choose a word w_n from p($w_n \mid z_n, \beta$), a multinomial probability distribution conditioned on the topic z_n.

$p(z_n = i)$ stands for the probability that the i^{th} topic was sampled for the n^{th} word and indicates which topics are important (in terms that they reflect the content) for a particular document. $p(w_n \mid z_n = i)$ stands for the probability of the occurrence of word w_n given the topic i and indicates the probability of word occurrence for each topic.

As already pointed, in our case, where the objective is to discover concepts and order them in a subsumption hierarchy, the documents are known, and the observations are the terms appearing in the documents. So, we aim to infer the topics that generated the documents and then organize these topics hierarchically. The proposed method uses the Markov Chain Monte Carlo (MCMC) process of Gibbs sampling [10]. The reader is referred to [9] for a detailed explanation of this process.

4 The Proposed Method

As Figure 2 illustrates, given a corpus of documents, the method first extracts the terms. The extracted terms constitute the input for the LDA model described in section 3. In the second step, feature vectors are constructed based on the document frequency of the terms. In the third step, the latent topics are generated as distributions over vocabulary terms according to the documents in the corpus and the terms observed. Finally, assuming that the topics generated correspond to ontology concepts, we organize them in a subsumption hierarchy.

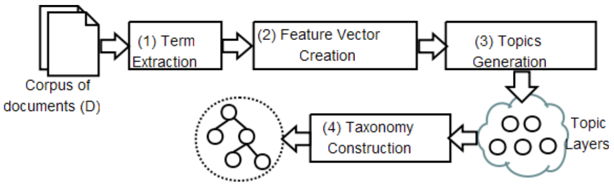

Figure 2. The ontology learning process.

More specifically, the stages followed by the proposed method are as follows:

(1) *Term Extraction* - From the initial corpus of documents, treating each document as a bag of words, we remove stop-words using statistical techniques in order to maintain the language-independence of the method. The remaining words constitute the vocabulary and form the term space for the application of LDA model.

(2) *Feature Vector Creation* - This step creates a Document - Term matrix, each entry of which records the frequency of each term in each document. This matrix is used as input to the LDA model.

(3) *Topic Generation* - Sets of topics are generated at this step by the iterative application of the LDA model for different values of the parameter K (number of topics). Therefore this step results in a multi-set of topics; each set being produced for a specific value of K. Starting from one topic, the method iterates and terminates when a predefined number of topics is reached. A small K forces a small number of topics to capture all the knowledge that the corpus contains, making the topics too generic in meaning. As K increases iteratively, the generated topics become more focused, capturing more detailed domain knowledge. Thus, the method starts from “general” topics, iterates and converges to more “specific” ones.

(4) *Taxonomy Construction* - Assuming the each of the computed topics corresponds to an ontology concept, the last step constructs the subsumption hierarchy of the discovered concepts: These are arranged in a hierarchical manner according their conditional independencies. The intuition behind this is as follows: since the generated topics are random variables, e.g. *A* and *B*, by measuring the mutual information we have a measure of their mutual dependence. Therefore, given a third variable *C* that makes *A* and *B* conditionally independent, the mutual information of the *A* and *B* topics approaches zero and *C* contains the information that both *A* and *B* variables contain (i.e. *C* is a broader topic than the others). In this case we may safely assume that *C* subsumes both *A* and *B*, which are formed as subsumees of *C*.

According to the iterative procedure of step 3, sets of “general” topics are being generated before the generation of sets of “specific” topics. In order to calculate the conditional independencies between topics, we take advantage of the document-topic matrix generated by the LDA model. Each entry of this matrix expresses the probability of a specific topic to participate in a specific document. In a more formal way, this is the probability of a topic, given a document. The process that generates the subsumption hierarchy is described by algorithm 1. It should be noted that this algorithm might generate more than one hierarchies.

```
Data: LDA output: Document - Topic matrix
Result: Subsumption hierarchy of topics
for every topic set L do
    for every topic i in topic set L do
        for every pair of topics (j, k) in topic set L+1
        do
            if (conditional independence of j and k
            given i is the maximum among other pairs)
            AND (satisfies a threshold th) then
                i is parent of j and k
            end
        end
    end
end
```

Algorithm 1: Taxonomy construction using conditional independence tests.

The algorithm starts from the first topic set that contains the most “general” topic and iterates across all topic sets generated with increasing values of K. Given the set of topics L_{i+1} generated for $K = i + 1$, the aim is to detect the pair of topics *(A,B)* whose independence is computed to be the maximum among the existing pairs of topics in L_{i+1}, given a topic *C* in L_i.

The conditional independence between two topics *A* and *B*, given a topic *C* is tested according to equation (1), where th is a threshold measuring the independence of the two topics given topic *C*.

$$|P(A \cap B \mid C) - P(A \mid C)P(B \mid C)| \leq th \qquad (1)$$

In order to compute equation (1) we need the probability of a topic *A* to participate in the corpus *D*, given that a topic *C* participates in the corpus. This is provided by equation (2):

$$P(A \mid C) = \frac{P(A \cap C)}{P(C)}. \qquad (2)$$

The probability of a topic *C* to participate in the corpus is given by the equation (3):

$$P(C) = \sum_{i=1}^{|D|} P(C \mid d_i)P(d_i), \qquad (3)$$

where $|D|$ is the number of documents in the corpus and $P(d_i) = \frac{1}{|D|}$ is the probability of a document in the corpus. Accordingly, the joint probability of topics *A* and *B* to participate in the corpus, given that a topic *C* participates in the corpus, is given by equation (4):

$$P(A \cap B \mid C) = \frac{P(A \cap B \cap C)}{P(C)}. \qquad (4)$$

Given the above probabilities, the mutual information between pairs of topics can be measured by equality (5):

$$I(A \cap B) = \sum_{a \in A} \sum_{b \in B} p(a,b) log \frac{p(a,b)}{p(a)p(b)}. \qquad (5)$$

By maximizing the independence of two topics given a third one, we minimize their corresponding mutual information:

$$I(A \cap B \mid C) = \sum_{a,b \in A,B} \sum_{c \in C} p(a,b \mid c) log \frac{p(a,b \mid c)}{p(a \mid c)p(b \mid c)}, \qquad (6)$$

which is true, since in this case the following equation holds:

$$log \frac{p(a,b \mid c)}{p(a \mid c)p(b \mid c)} = log1 = 0 \qquad (7)$$

It is worth-noting, that since the algorithm searches for conditional independences between pairs of topics, it is not able to infer subsumption relations in the case where a topic subsumes only one other topic.

5 Experiments

We have evaluated the proposed method on the GENIA corpus, which contains 2000 documents from the domain of biomedicine and is accompanied by the GENIA ontology, comprising 54 concepts and 45 subsumption relations between them. These resources are available from the GENIA project [1]. The computation of the latent topics has been done with a stand-alone Java application making use of the Gibbs sampling implementation [2]. The parameters involved are the maximum number of topics (K) and the threshold (th) value involved in equation (1). Since Algorithm 1 performs an exhaustive search to find the best solution, its complexity is $O(K^3)$. Although there is still space for making this algorithm more efficient, the implemented algorithm needed only 4 minutes to compute the hierarchies of our experiments on a standard Pentium 3.0 GHz PC.

Although the maximum number of topics K affects the number of iterations of the algorithm described in section 4, it must be pointed that it does not affect the depth of the produced hierarchy, leaving this choice to algorithm 1. The depth of the hierarchy depends on the inclusion relations that are discovered between topics in different layers.

We have experimented with various values of the parameter K and th. The results that are provided in this section have been produced for $K = 54$ topics and $th = a * 10^{-6}$, where $1 < a < 9$ is a constant. We have experimented with different values of the threshold parameter, forcing it to be as low as possible. As is the case for the value of K, the value of the threshold also does not affect the subsumption hierarchy construction. The algorithm removes the topics that are not related to any other topic and presents the final hierarchy (or hierarchies) computed.

Figure 3 depicts a part of the produced ontology, as well as a part of the GENIA ontology. The latter serves as the "gold" standard for evaluating the results of the proposed method.

The GENIA ontology contains two distinct hierarchies. The proposed method manages to distinguish these two hierarchies successfully and infer results that are close to the gold standard. However, the method was not able to identify some very specific concepts of the GENIA ontology. This is due to the nature of the LDA model, which ignores topic correlations, assuming that the produced topics are independent to each other, where they are not. Thus, its ability to discover a large number of fine-grained, tightly-coherent topics is reduced [15].

It must be pointed that inherently, the proposed method does not label the concepts of the produced ontology. It presents the topics as distributions over words. However, the GENIA corpus is annotated with tags that indicate the instances of the concepts that appear in the documents. It should be stressed that this information was not exploited at all by the ontology learning procedure. The annotation of the corpus was taken into account only at the evaluation stage. Topics were labelled automatically by aligning each topic with the GENIA concept with which it shares

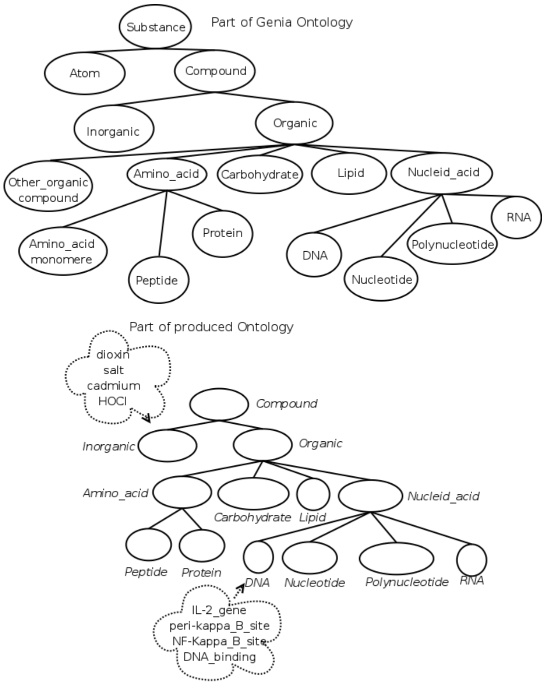

Figure 3. Parts of the produced ontology and the GENIA ontology. In clouds: important terms that participate in the corresponding topics, which are also concept instances of the GENIA ontology.

the largest number of instances, as it is illustrated in figure 3. Specifically, we created the lists of the high-probability terms participating in each topic. These terms all share a common characteristic: their probability values are greater than the mean value of their corresponding distribution, plus its standard deviation. By comparing the lists of terms in topics we label the topics according to the instances of concepts that they contain. In the case where a topic contains instances of more than one concept, then this topic is named by the concept which contributes the majority of instances.

We have evaluated the proposed method in terms of *precision* and *recall* metrics: regarding the concept identification, we define *precision* as the ratio of the number of concepts correctly detected to the total number of concepts detected, and *recall* as the ratio of the number of concepts correctly detected to the number of concepts in the gold standard. The *F-measure* is a combined metric defined as follows:

$$Fmeasure = \frac{2 * precision * recall}{precision + recall}. \qquad (8)$$

Accordingly, for the subsumption relations (SRs): *precision* is the ratio of the number of SRs correctly detected to the total number of SRs detected, and *recall* is the ratio of the number of SRs correctly detected to the number of SRs in the gold standard. Table 1 provides the evaluation results.

Table 1. Evaluation results.

Configuration: $K = 54, th = 3 * 10^{-6}$		
Concept Identification		
Precision	Recall	F-measure
74%	85%	79%
Subsumption Hierarchy Construction		
Precision	Recall	F-measure
97%	82%	88%

Overall the results are very encouraging, showing that the method was able to construct a very large part of the original ontology, based purely on machine learning from the corpus. Its weakest result is the precision in the identification of concepts. This is partly due to the method's inability to identify very specific concepts, stated above and also due to the labeling procedure used in the evaluation phase, which failed to label some topics, as it was not clear which concept's instances they were containing. Concerning recall, the 15% of missing concepts is due to the fact that not all the GENIA concepts are covered by the corpus.

Finally, one should bear in mind that the evaluation of ontologies when these ontologies are produced by an automated learning procedure is an open field of research. The research community has not established a standard methodology for automating ontology evaluation. Especially when the evaluation is done against a gold standard ontology, it seems that we cannot judge objectively the result, since the gold standard was created by humans probably in a subjective or a biased manner. Particularly, in cases where the ontology has been learned from scratch and it is not the result of a seed-ontology enrichment, the evaluation is even more difficult.

6 Conclusion and Future Work

In this paper we have proposed a fully-automated method for learning ontologies. The proposed method uses the Latent Dirichlet Allocation model for the discovery of topics that represent ontology concepts. According to this method, topics are represented as multinomial distributions over document terms. Then, a method that performs conditional independence tests among topics is applied to arrange concepts in a subsumption hierarchy.

The major advantage of this approach is its statistical nature, which is based on probabilistic topic models. This allows the computation of topics in a language-neutral way, revealing those topics that express the contents of documents, and thus, the concepts that express the knowledge that documents mediate. This makes the method very generic, tackling at the same time both problems of concept identification and hierarchy construction.

The proposed method was evaluated on the GENIA ontology and associated corpus. The results that we obtained were very encouraging, showing that the method can reconstruct a large part of the GENIA ontology from the analysis of the corpus. One weakness of the method that was identified in the experiment was the difficulty of LDA to detect some very specific topics. An additional issue that requires further work is the automated concept labelling procedure that we introduced for the evaluation of our method.

Finally, further work includes the use of the TF/IDF measure instead of word frequencies at the feature vector creation step for the LDA model, in conjunction with the variational inference [10] to infer the latent topics.

Acknowledgments

The research described within this paper was supported be the research and development project ONTOSUM[2], which is in turn funded by the Greek General Secretariat for Research and Technology.

References

[1] The genia project, http://www-tsujii.is.s.u-tokyo.ac.jp/genia.

[2] Lda java implementation, http://www.arbylon.net/projects.

[3] E. Agirre, O. Ansa, E. Hovy, and D. Martinez. Enriching very large ontologies using the www. In *ECAI'00 Workshop on Ontology Construction*.

[4] E. Alfonseca and S. Manandhar. An unsupervised method for general named entity recognition and automated concept discovery. In *International Conference on General WordNet*, 2002.

[5] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*.

[6] A. Faatz and R. Steinmetz. Ontology enrichment with texts from the www. In *Semantic Web Mining Workshop ECML/PKDD'02*.

[7] B. Fortuna, D. Mladevic, and M. Grobelnik. Visualization of Text Document Corpus. In *ACAI 2005*.

[8] E. Gaussier, C. Goutte, K. Popat, and F. Chen. A hierarchical model for clustering and categorising documents. In *BCS-IRSG 2002*.

[9] T. Griffiths and M. Steyvers. A probabilistic approach to semantic representation. In *Conference of the Cognitive Science Society*, 2002.

[10] T.L. Griffiths and M. Steyvers. Finding scientific topics. In *National Academy of Science*, 2004.

[11] M.A. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *International Conference on Computational Linguistics*, 1992.

[12] U. Heid. A linguistic bootstrapping approach to the extraction of term candidates from german text. 1998.

[13] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, 1999.

[14] J.S. Justeson and S.M. Katz. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1995.

[15] Wei Li and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *ICML 2006*.

[16] D.I. Moldovan and R.C. Girju. An interactive tool for the rapid development of knowledge bases. *Journal on Artificial Intelligence Tools*, 2001.

[17] G. Paaß, J. Kindermann, and E. Leopold. Learning prototype ontologies by hierarchical latent semantic analysis. In *Knowledge Discovery and Ontologies*.

[18] C. Roux, D. Proux, F. Rechermann, and L. Julliard. An ontology enrichment method for a pragmatic information extraction system gathering data on genetic interactions. In *ECAI'00 Workshop on Ontology Learning*.

[19] G. Salton and M.H. McGill. *Introduction to Modern Information Retrieval*.

[20] M. Steyvers. *Probabilistic Topic Models*.

[21] A Wagner. Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis. In *ECAI'00 Workshop on Ontology Learning*.

[2] See also *http://www.ontosum.org/*