Learning Ontologies of Appropriate Size

Elias Zavitsanos[1,2], Sergios Petridis[1], Georgios Paliouras[1],
and George A. Vouros[2]

[1] Institute of Informatics and Telecommunications, NCSR "Demokritos", Greece
[2] University of Aegean, Department of Information and Communication Systems Engineering, Artificial Intelligence Laboratory, Samos, Greece
{izavits,petridis,paliourg}@iit.demokritos.gr, georgev@aegean.gr

Abstract. Determining the size of an ontology that is automatically learned from text corpora is an open issue. In this paper, we study the similarity between ontology concepts at different levels of a taxonomy, quantifying in a natural manner the quality of the ontology attained. Our approach is integrated in a recently proposed method for language-neutral learning of ontologies of thematic topics from text corpora. Evaluation results over the Genia and the Lonely Planet corpora demonstrate the significance of our approach.

1 Introduction

Ontology learning is a relatively new field of research, aiming to support the continuous and low-cost development and maintenance of ontologies, especially in fast evolving domains of knowledge. These tasks, when performed manually, require significant human effort. Thus, automated methods for ontology construction are very much needed.

Ontology learning is commonly viewed [1, 5, 13, 14] as the task of *extending* or *enriching* a seed 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 concept properties/relations. Furthermore, we may classify existing ontology learning approaches to be either of the linguistic, statistical, or machine learning type, depending on the specific techniques employed.

While much work concentrates on enriching existing ontologies, few approaches deal with the construction of an ontology without prior knowledge. Among the difficulties of such an endeavor, is the determination of the appropriate depth of the subsumption hierarchy, given the text collection at hand. The benefit of being able to determine the appropriate depth of a taxonomy is that the hierarchy captures accurately the domain knowledge provided by the texts, reducing the extent of overlap among concepts and providing a coherent representation of the domain. The determination of the appropriate hierarchy depth prohibits both over-engineered representations and generic ones, since it constitutes a criterion for a well-structured hierarchy. However, there is a strong dependence of such a method to the corpus, since an imbalanced corpus could lead to a misleading decision for the appropriate depth.

J. Darzentas et al. (Eds.): SETN 2008, LNAI 5138, pp. 327–338, 2008.

In this paper, we propose an automated statistical approach to ontology learning, without presupposing the existence of a seed ontology, or any other type of external resource, except the corpus of text documents. The proposed method tackles the tasks of concept identification and subsumption hierarchy construction. Moreover, it tries to optimize the size of the learned ontology for the given text collection.

In the proposed method, concepts are identified and represented as multinomial distributions over terms in documents[1]. Towards this objective, the Markov Chain Monte Carlo (MCMC) process of Gibbs sampling [9] is used, following the Latent Dirichlet Allocation (LDA) [4] model. To discover the subsumption relations between the identified concepts, conditional independence tests among these concepts are performed. Finally, statistical measures between the discovered concepts at different levels of the hierarchy are used to optimize the size of the ontology. The statistical nature of the approach guarantees language independence.

In what follows, section 2 states the problem, refers to existing approaches that are related to the proposed method, and motivates our approach. In section 3, we present the new method, while section 4 describes the derivation of a criterion for determining the appropriate depth of the hierarchy according to the corpus. Section 5 presents experiments and evaluation results, and finally, section 6 concludes the paper sketching plans for future work.

2 Problem Definition and Related Work

2.1 Problem Definition

In this paper we address three major problems related to the ontology learning task:

1. The discovery of the concepts in a corpus.
2. The ordering of the discovered concepts by means of the subsumption relation.
3. The determination of the depth of the subsumption hierarchy.

In other words, assuming only the existence of a text collection, we aim to (a) discover the concepts that express the content of documents in the corpus, independently of the terms' surface appearance, i.e. without taking into account simple TF/IDF values or the order of words in the texts, (b) form the ontology subsumption hierarchy backbone, using only statistical information concerning the discovered concepts, and (c) explore how deep in the subsumption hierarchy the text collection allows us to go, by measuring the similarity between the discovered concepts.

[1] "Terms" does not necessarily denote domain terms, but words that will constitute the vocabulary over which concepts will be specified. In the following, we use "terms" and "words" interchangeably.

2.2 Related Work

Towards the automated learning of ontologies, much work concerns concept identification and taxonomy construction. In this paper we are interested in statistical techniques, and thus, we discuss here related approaches.

On the task of concept identification with statistical techniques, the authors in [2] extend an ontology with new concepts considering words that co-occur with each of the existing concepts. The method requires that there are several occurrences of the concepts to be classified, so that there is sufficient contextual information to generate topic signatures. The work reported in [1] follows similar research directions. In [5], the authors apply statistical analysis on Web pages in order to identify word clusters that are proposed as potential concepts 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.

More sophisticated schemes include the use of TF/IDF weighting in conjunction with Latent Semantic Indexing (LSI) [6], 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) [10] extends LSI assuming that each document is a probability distribution over topics and each topic is a probability distribution over words. Although PLSI provides more accurate modelling than 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 [4]. Latent Dirichlet Allocation (LDA) [4] improves on PLSI, providing a model that samples topics for each word that appears in each document.

Hierarchical extensions have also been proposed to the above models. Hierarchical Probabilistic Latent Semantic Analysis (HPLSA) has been proposed in [7], in order to acquire a hierarchy of topics, by enabling data to be hierarchically organized based on common characteristics. Hierarchical Latent Semantic Analysis (HLSA) has been introduced in [12] to identify hierarchical dependencies among concepts by exploiting word occurrences among concepts (latent topics). This approach actually computes relations among topics, based on the words that they contain. Different topics might share common words, and therefore these words are collected at a higher level. Both of these methods, inherit known problems of PLSI, such as overfitting.

Moreover, the method of hLDA [3], a hierarchical extention of LDA, has been proposed to deal with the problem of the hierarchical organization of topics. However, this method assumes that each document is a mixture of topics along a path from the root topic to a leaf, making this way a document to comprise only one specific topic and its abstractions. Finally, the model of hPAM [11] deals with some limitations of hLDA. It allows multiple inheritance between topics, but on the other hand the fixed-depth hierarchy that produces and the need for predefining the number of topics are its basic limitations. In general, determining the appropriate depth of the hierarchy still remains to our knowledge an open issue.

In this paper we address the problems of concept identification and taxonomy construction using statistical and machine learning techniques. The statistical nature of the proposed method assures that the method is not dependent on the language of the corpus, but only on the statistical information that the corpus provides, i.e., the word frequencies. In addition, having no prior knowledge, we aim to determine statistically the depth of the hierarchy.

3 The Method

As depicted in figure 1, given a corpus of documents, the method first extracts the terms. The extracted terms constitute the term space, over which the latent topics are defined. In the second step, feature vectors are constructed for each document, based on term frequency. Next, the latent topics are generated as distributions over vocabulary terms according to the documents in the corpus and the terms observed. Through an iterative process, latent topics are discovered and organized into hierarchical layers until the criterion for appropriate depth is satisfied.

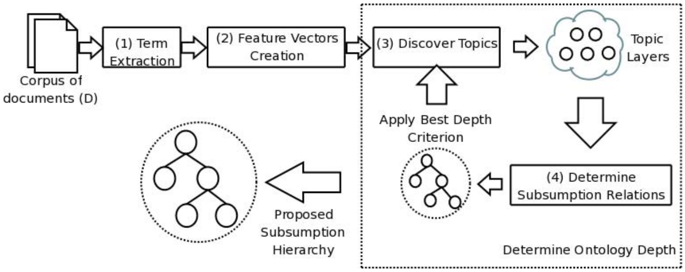
Corpus of documents (D)
(1) Term Extraction
(2) Feature Vectors Creation
(3) Discover Topics
Topic Layers
Apply Best Depth Criterion
(4) Determine Subsumption Relations
Proposed Subsumption Hierarchy
Determine Ontology Depth

Fig. 1. The proposed ontology learning method

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. The remaining words constitute the vocabulary, forming the term space for the application of the topic generation 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 topic generation model.

(3) *Discover Topics* - In this step, the iterative task of the learning method is initiated. To generate the topics we follow the Latent Dirichlet Allocation (LDA) [4] approach. LDA belongs in the family of Probabilistic Topic Models (PTMs). These models are based on the idea that documents are mixtures of thematic topics, which are represented by means of probability distributions over terms. PTMs are based on the bag-of-word assumption, assuming that words are independently and identically distributed in the texts, given the thematic

topics of each text. PTMs are generative models for documents: they specify a probabilistic procedure by which documents are generated as combinations of latent variables, i.e. topics. Generally, this procedure states that topics are probability distributions over a predefined vocabulary of words and according to the probability that a topic participates in the content of each document, words are sampled from the corresponding topic in order to generate the documents.

The LDA model specifies a generative process, according to which, topics are sampled repeatedly in each document. Specifically, given a predefined number of topics K, 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, i.e., reflect the content of 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 significance of each word for each topic.

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. Thus, we aim to infer the topics that generated the documents and then organize these topics hierarchically. In order to infer the latent topics, the proposed method uses the Markov Chain Monte Carlo (MCMC) process of Gibbs sampling [8].

At each iteration of this step, sets of topics, that we call layers, are generated by the iterative application of LDA. Starting with one topic and by incrementing the number of topics in each iteration, layers with more topics are generated. A layer comprising few topics attempts to capture all the knowledge of the corpus through generic concepts. As the number of topics increases, the topics become more focused, capturing more detailed domain knowledge. Thus, the method starts from “general” topics, iterates, and converges to more “specific” ones.

(4) *Determine Subsumption Relations* - In each iteration, the method identifies the subsumption relations that hold between topics of different layers. The discovered topics are arranged in a hierarchical manner according to their conditional independencies, determined by the following condition:

$$|\hat{P}(A \cap B \mid C) - \hat{P}(A \mid C)\hat{P}(B \mid C)| \leq th. \tag{1}$$

Equation (1) is best explained through an information theoretic framework. Specifically, since the generated topics are random variables, e.g. A and B, by measuring their mutual information we obtain an estimate of their mutual dependence. Therefore, given a third variable C that makes A and B (almost) conditionally independent, the mutual information of topics A and B is reduced and C contains a large part of the common information of A and B, i.e., C is a broader topic than the others. In this case we may safely assume that C subsumes both A and B and the corresponding relations are added to the ontology. We should also point out, that C has been generated before A and B. Thus, it belongs in a layer that contains topics that are broader in meaning than the ones in the layer of A and B.

In addition, we search for a topic C that makes topics A and B as much conditionally independent as possible. Therefore, between two possible parent topics C_1 and C_2 we will choose the one that maximizes the difference of the following mutual informations:

$$\Delta = I(A, B) - I(A, B \mid C) \tag{2}$$

Therefore, equation (1) is not used as an absolute measure to judge the subsumption relations between concepts, but as a relative way of finding the best concept C that can be considered as the father of A and B.

In order to calculate the conditional independencies between topics, we use the document-topic matrix generated by the LDA model. Each entry of this matrix expresses the probability of a specific topic given a document. The estimation of the probabilities of equation (1) is explained in [15]. Moreover, the threshold th has been introduced to avoid small rounding problems at the calculations. Therefore, it has a very small value near zero.

(5) *Determine Ontology Depth* - A significant contribution of the proposed method is the determination of the appropriate depth of the hierarchy from the given corpus of documents. As already mentioned, the topics are probability distributions over the term space. We use a criterion based on the similarity of these distributions that indicates the convergence towards the appropriate depth. We thus improve on our recently proposed work [15] by proposing algorithm 1. The way in which the appropriate depth of the taxonomy is determined is explained in the following section.

4 Measuring Similarity between Concepts

In the proposed method, concepts are represented as multinomial distributions over terms in documents. In order to determine the depth of the subsumption hierarchy we define a criterion based on the symmetric KL divergence between concepts of different levels that participate in subsumption relations. The intuition behind this is that the symmetric KL divergence between concepts that belong in the top levels of the hierarchy should be higher than the KL divergence between concepts that belong in the lower levels of the hierarchy. This is due to the fact that the top concepts are broader in scope than lower ones and the "semantic distance" between them and their children is expected to be higher than this of more specific concepts and their children.

In order to validate this assumption, we have experimented with two golden standard ontologies and the corresponding corpora:

```
Data: Document - Term Matrix
Result: Subsumption hierarchy of topics
initialization;
start with number of topics K=1;
while Stop Criterion not achieved do
    Generate a pair of topic layers in parallel (for current value of K and for
    K+1);
    for every topic i in 1st topic layer of pair do
        for every pair of topics (j, k) in 2nd topic layer do
            if (conditional independence of j and k given i is the maximum
            among other pairs) AND (satisfies the threshold th) then
                i is parent of j and k
            end
        end
    end
    if Stop Criterion achieved then
        end;
    else
        increase number of topics;
    end
end
```

Algorithm 1. Constructing a subsumption hierarchy of appropriate depth

1. The Genia[2] ontology comprises 43 concepts connected with 41 subsumption relations, which is the only type of relation among the concepts. The corresponding corpus consists of 2000 documents from the domain of molecular biology.
2. The Lonely Planet ontology contains 60 concepts and 60 subsumption relations among them. The Lonely Planet corpus is a collection of about 300 Web pages from the Lonely Planet Web site[3], providing touristic information.

In order to measure the similarity of the concepts in the ontologies, we represented the concepts of each gold standard ontology as probability distributions over the term space of the corresponding corpus, as shown in figure 2. This representation allows the application of statistical measures concerning the similarity between concepts.

To represent each concept as distribution over terms we have to measure the frequency of the terms that appear in the context of each concept. In both corpora, the concept instances are annotated in the texts, providing direct population of the concepts in the golden standard ontologies with their instances. Since we have populated each concept with its instances, it is possible to associate each document to the concept(s) that it refers to, by counting the concept instances that appear in the document. Thus, we are able to create feature vectors based on the document in which each concept appears. These feature vectors actually form a two-dimensional matrix that records the frequency of each term in the context

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

[3] The Lonely Planet travel advise and information, http://www.lonelyplanet.com/

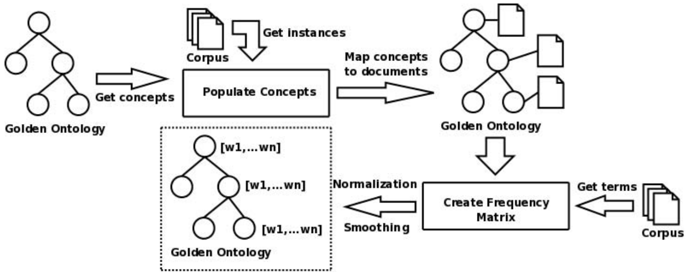

Fig. 2. The process of representing the golden ontology concepts as probability distributions over the term space

of each concept. That is, we have a "concept - term" matrix that represents each concept as a distribution over the term space of the text collection.

For each concept, frequencies are normalized giving a probability distribution over the term space. Since the goal is to measure the symmetric KL divergence between concepts that participate in subsumption relations, we also performed smoothing of the probability distributions to eliminate possible zero values of unseen terms. For this purpose, we applied the Laplace law (3) on the probability distribution of each concept.

$$\hat{P}_L(w_i) \doteq \frac{\hat{P}(w_i)+1}{N+1}, \forall i, \tag{3}$$

where N is the vocabulary size.

In order to measure the symmetric KL divergence between two concepts p and q that are related with a subsumption relation, we used the following formula:

$$D_{KL} = \frac{1}{2}[\sum_i P(w_i) log \frac{P(w_i)}{Q(w_i)} + \sum_i Q(w_i) log \frac{Q(w_i)}{P(w_i)}], \tag{4}$$

where $P(\cdot)$ and $Q(\cdot)$ are the distributions corresponding to concepts p and q. Small values of KL divergence indicate high similarity between concepts. Figure 3 depicts the results obtained by measuring the similarity between concepts that participate in subsumption relations, in the case of the Genia and the Lonely Planet gold standard ontologies.

Figure 3 confirms our assumption that concepts at the lower levels of the hierarchy are more similar to their children than concepts at higher levels of the hierarchy. For both corpora, KL divergence is minimized at the leaf level of the ontologies.

Based on this approach of measuring the KL divergence of subsumed concepts, we define a relative criterion that indicates how deep the hierarchy should be according to the information provided by the corpus of documents. This criterion, which corresponds to the stop criterion of Algorithm 1, is defined as follows:

$$1 - \frac{KL_{bottom}}{KL_{top}} < \varepsilon. \tag{5}$$

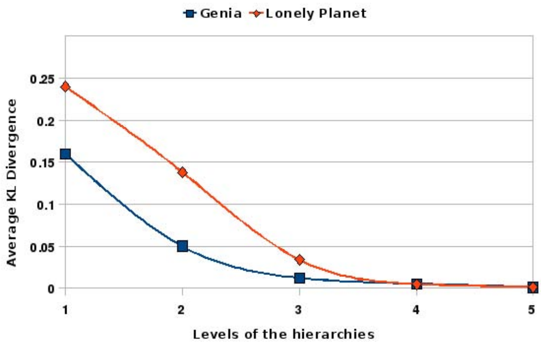

Fig. 3. Average KL Divergence of subsumed concepts in the Genia and the Lonely Planet gold standard ontology

In equation (5), KL_{top} corresponds to the average symmetric KL divergence between the concepts of level l and the concepts of level $l+1$. KL_{bottom} is the average symmetric KL divergence between the concepts at level $l+1$ and the concepts of level $l+2$. Values close to 0 indicate that the new level of concepts added does not differ much from the parent concepts. Thus we are reaching maximum "specificity" and therefore optimal depth. Values near 1 indicate that the hierarchy can go deeper. Actually, the parameter ε, does not depend on the application. It has a very small value very close to zero to avoid small rounding errors during the computations.

5 Evaluation

We have evaluated the proposed method on both corpora introduced in section 4. The parameters that have been introduced in this paper are the parameter ε in the stop criterion (5) and the threshold th for the significance of subsumption relations. Both parameters are introduced in order to provide control over the process, although the method is robust to the values of the parameters. Typically, one would choose very small values for these parameters, independent of the particular application.

The evaluation procedure that we followed uses the representation of the golden standard concepts as probability distributions over the term space of the documents, as explained in section 4. In addition, the concepts of the produced hierarchy have exactly the same representation. They are probability distributions over the same term space. We can, thus, perform a one-to-one comparison of the golden concepts and the produced topics. More specifically, a topic is matched to a concept if their corresponding distributions were the "closest" compared to all the other and their KL divergence (4) was below a fixed threshold th_{KL}. Obviously, small values of KL divergence indicate high similarity between golden concepts and discovered topics.

The quantitative results have been produced using the metrics of *Precision* and *Recall*. 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. 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.

The choice of threshold th_{KL} affects the quantitative results, since a strict choice would force few topics to be matched with golden concepts, while a loose choice would cause many topics to be matched with golden concepts. We have chosen a value of $th_{KL} = 0.2$ for the purposes of our evaluation, as we observed relative insensitivity of the result for values between 0.2 and 0.4 and we opted for the more conservative value in this plateau. Tables 1 and 2 depict the experimental results in the case of the Genia and Lonely Planet corpora respectively.

Table 1. Evaluation results for the Genia corpus

Concept Identification		
Precision	Recall	F-measure
94%	76%	84%

Subsumption Hierarchy Construction		
Precision	Recall	F-measure
93%	75%	83%

Table 2. Evaluation results for the Lonely Planet corpus

Concept Identification		
Precision	Recall	F-measure
62%	36%	44%

Subsumption Hierarchy Construction		
Precision	Recall	F-measure
53%	35%	42%

In order to obtain a more detailed picture of the performance of the method, we replaced the stopping criterion of Algorithm 1 with predefined depths for the learned hierarchy and we experimented in both corpora. Figures 4 and 5 present the evaluation results in terms of the F-measure for various depths of the hierarchy, using the same configuration ($th_{KL} = 0.2$) for the evaluation method.

Figure 4 depicts that for a predefined depth of 8 levels of the produced hierarchy, the F-measure is maximized compared to the Genia gold standard. Respectively, in the case of the Lonely Planet corpus, the F-measure is maximized for a predefined depth of 10 levels of the produced hierarchy (figure 5). Tables 1 and 2

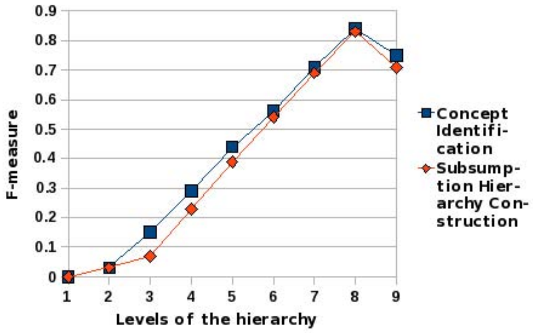

Fig. 4. F-measures for Concept Identification and Subsumption Hierarchy Construction for the Genia corpus

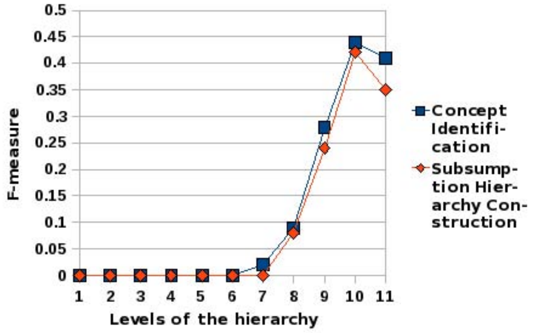

Fig. 5. F-measures for Concepts Identification and Subsumption Hierarchy Construction for the Lonely Planet corpus

confirm that the proposed method, using the stop criterion that we derived in section 4, managed to achieve the best results in both corpora. Therefore, the method determined correctly the appropriate depth in both corpora.

Concerning the quantitative results, in the case of the Genia corpus, where the golden concepts were instantiated sufficiently in the documents, i.e. the texts contain many concepts instances, the numerical results were higher than the ones in the case of the Lonely Planet corpus, where half of the golden concepts had only one instance and generally most of the concepts were insufficiently instantiated. The difficulty of the model to retrieve some very specific concepts in the Lonely Planet corpus is due to this fact.

6 Conclusions

In this paper, we have presented a method for concept identification and taxonomy construction that determines automatically the appropriate size of the subsumption hierarchy. We improved our recently proposed method that relies on conditional independence tests between thematic topics, by incorporating a statistical criterion that determines the appropriate depth of the produced hierarchy. We have also experimented with two corpora, where we have showed that the presented method managed to determine the most appropriate size for the subsumption hierarchy, producing the best quantitative results.

Future work includes further experiments to validate the proposed method on new ontologies and corpora.

References

1. Agirre, E., Ansa, O., Hovy, E., Martinez, D.: Enriching very large ontologies using the www. In: ECAI 2000 Workshop on Ontology Construction (2000)
2. Alfonseca, E., Manandhar, S.: An unsupervised method for general named entity recognition and automated concept discovery. In: International Conference on General WordNet (2002)
3. Blei, D., Griffiths, T., Jordan, M., Tenenbaum, J.: Hierarchical topic models and the nested chinese restaurant process. In: NIPS (2004)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. In: Journal of Machine Learning Research (2003)
5. Faatz, A., Steinmetz, R.: Ontology enrichment with texts from the www. In: Semantic Web Mining Workshop ECML/PKDD (2002)
6. Fortuna, B., Mladevic, D., Grobelnik, M.: Visualization of Text Document Corpus. In: ACAI (2005)
7. Gaussier, E., Goutte, C., Popat, K., Chen, F.: A hierarchical model for clustering and categorising documents. In: BCS-IRSG (2002)
8. Griffiths, T., Steyvers, M.: A probabilistic approach to semantic representation. In: Conference of the Cognitive Science Society (2002)
9. Griffiths, T.L., Steyvers, M.: Finding scientific topics. In: National Academy of Science (2004)
10. Hofmann, T.: Probabilistic latent semantic indexing. In: SIGIR (1999)
11. Mimno, D., Li, W., McCallum, A.: Mixtures of hierarchical topics with pachinko allocation. In: Proceedings of the 24th International Conference on Machine Learning (2007)
12. Paaß, G., Kindermann, J., Leopold, E.: Learning prototype ontologies by hierarchical latent semantic analysis. In: Knowledge Discovery and Ontologies (2004)
13. Roux, C., Proux, D., Rechermann, F., Julliard, L.: An ontology enrichment method for a pragmatic information extraction system gathering data on genetic interactions. In: ECAI Workshop on Ontology Learning (2000)
14. Wagner, A.: Enriching a lexical semantic net with selectional preferences by means of statistical corpus analysis. In: ECAI Workshop on Ontology Learning (2000)
15. Zavitsanos, E., Paliouras, G., Vouros, G.A., Petridis, S.: Discovering subsumption hierarchies of ontology concepts from text corpora. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence - WI 2007. Springer, Heidelberg (2007)