

Mapping product descriptions to a large ontology

Henrik Oxhammar

Department of Linguistics

Stockholm University

116 63 Stockholm

Sweden

henrik.oxhammar@ling.su.se

Abstract

In this paper we describe an information retrieval approach for mapping online business information texts to concepts in a large ontology. We adopt the traditional vector space model by representing the texts as queries and the concept labels in the ontology as documents.

Because of the size of the ontology and the fact that concept labels are very sparse and generic, we conducted additional experiments for *reducing* the set of concepts, as well as the *enrichment* and *enlargement* of concept labels.

The documents in our collection were of too poor quality for this task, and although we show that our enrichment technique did provide us with an ontology with good overall similarity to our query collection, individual concepts did not include enough terms for our method to achieve good results.

1 Introduction

In this paper we describe our first attempts in building a system that assigns product and service concept labels in a large ontology, to online business information texts. Our method can be viewed as an information retrieval approach to standard text categorization.

The task of mapping such business information is a difficult one because of the number of concepts (over 8300) and the fact that the *labels* that describe these concepts are very sparse. Further, only a small percentage of the terms that these labels include, are also present in the vocabulary used on corporate

websites. We therefore also describe a number of techniques for improving the performance of a baseline system, including a technique for reducing the number of candidate concept labels for a document and two techniques for *enlarging* and *enriching* sparse concept labels.

2 The Common Procurement Vocabulary

The Common Procurement Vocabulary (CPV) (European Union and European Parliament, 2002) is a standardized vocabulary developed by the European Union. It is a classification scheme for public procurement and its purpose is to help procurement personnel and contracting authorities describe and classify their procurement contracts.

The CPV ontology defines products and services as concepts in a strict taxonomic relationship structure. Each concept consists of a unique eight-digit code and a *concept label* that describes the concept in natural language. E.g.,

```
18000000 "Clothing and accessories"  
18500000 "Leather clothes"  
18510000 "Leather clothing accessories"  
18512000 "Leather belts and bandoliers"  
18512100 "Belts"  
18512200 "Bandoliers"
```

The ontology defines 8323 unique concepts of this kind. By a concept's code, it is possible to derive a number of useful facts. First, we can determine at what level the concept is defined. E.g., 18512000 "Leather belts and bandoliers" resides on level five. Leaf concepts, i.e., concepts that have the finest granularity, make up almost 68% (5644) of the total number of concepts. (Warin et al., 2005)

The ontology is a strict taxonomy, i.e., concepts are related by the hyponymy/hyperonymy (sub/super) relationships. Therefore, it is pos-

sible to also derive a concept's *parent(s)*, *descendant(s)* and *sibling(s)*. E.g., the direct parent of 18512100 "Belts" is 18512000 "Leather belts and bandoliers", and 18512200 "Bandoliers" is its sibling since 18512200 shares the same parent as 18512100. 18512100 "Belts" has no descendants since it is a leaf concept.

3 The vector space model

Our method of mapping product descriptions to concepts in the CPV is based on the classical information retrieval approach for querying documents.

The *vector space model* (VSM) (Salton et al., 1975) assigns weights to document- and query terms and the model represents each document and query as multi-dimensional feature vectors. By computing the cosine angle between the vectors the similarity of a document and a query can be established. The smaller this angle is the more similar the query and the document are said to be. The model then usually returns a ranked list that includes the most similar documents to the query given.

The vector model has several advantages, including its partial matching ability, its simplicity, its quick retrieval times and the fact that it can allow queries to be of any size. The vector space model has proved to be superior or as good as other retrieval models. (Baeza-Yates and Ribeiro-Neto, 1999)

The task for the system we present in this paper was to suggest a number of concepts in the ontology that the model regarded being most similar, given a business information text as query. We further regard the *concept labels* that describe the concepts in the ontology as documents. The model returns the suggested concepts in ranked list according to relevance, with the most relevant concept first.

Our implementation of the vector model included stemming, term weighting by document frequency and vector normalization by unit length.

4 Resources

4.1 Query collection

A collection of 739 documents crawled from various corporate web sites was at our disposal. We briefly inspected the texts for them

to at least include *some* information of the company's activities. The size of the documents varied from about 20 Bytes to about 400 Kbytes. Below is an example of a text in our query collection:

"Electronic Assembly Wire and Cable
The variety of electrical cable applications, provided by EDEC Kabel bv, is widely spread. Edec Kabel bv is a Sales Office specialised in electrical cable. We represent and or cooperate with several cable manufactures like : E&E GmbH Germany, GmbH Germany and others. In addition to this, EDEC Kabel bv, also supplies a comprehensive line of standard wire and cable to offer a total cable product range"

Far from all texts were of this quality. Because of the limited number of documents, we however decided to keep documents with less information in our query collection.

4.2 Gold standard

To evaluate our method, we had manually made mappings, between each of the 739 documents in the query collection and a number of concepts in our ontology. For example, the document in the previous section had been associated with:

28400000 "Cable, wire and related products"
31300000 "Insulated wire and cable"
31330000 "Coaxial cable"

The number of associated concepts ranged from one to 60 in one case. The and the total number of relevant concepts for our query collection was 3348, giving an average number of 4.5 associated concepts in our gold standard.

Unfortunately, many of the associated concept labels did not describe their respective document as well as the example above indicate.

5 Evaluation measures

The most commonly used measures for evaluating information retrieval systems are *precision* and *recall*. Precision gives the percentage of the number of correctly retrieved documents among all documents retrieved, while recall reflects the percentage of relevant documents retrieved by a query among all relevant documents for this query.

However, recall and precision are not appropriate measures when a system returns answers to a query in a ranked list according to

their relevance, simply because precision and recall (by default) do not take the ranking into account. A proper evaluation measure for such systems is instead a measure known as *interpolated precision averages at 11 standard recall levels*. This measure shows the overall quality and effectiveness of the system, by taking the ranking into account. (Baeza-Yates and Ribeiro-Neto, 1999)

Since it still can be difficult to compare two systems' precision averages at various recall levels, another *single* measure can be used that provides a *summary* of a system's performance over all relevant documents. *Average precision* is a single-valued measure that calculates the average of precision values after each relevant document has been retrieved, that enables two systems to be compared by a single value. (Baeza-Yates and Ribeiro-Neto, 1999)

When we evaluated the experiments described in section 6, we only did so according to *exact* matches. As mentioned in section 2, there are several relationships that the ontology describe, including the parent-child and sibling relations. Since these concepts are closely related, we could regard the mapping of a document to its correct concept's sibling, parent or child concept, not as *incorrect* but, instead, as *partially correct*. E.g., suppose a document is associated with 29521321 "Drilling machinery". This tells us that, the company described in the document does not manufacture drills for, say, *home use*, but drills used for things such as *mining*, since the parent of 29521321 is 29521320 "Oil drilling equipment". So if the system suggests 29521320 "Oil drilling equipment", then we can regard the system to be correct to some extent.

We can further induce from an associated concept that a document can also belong to the associated concept's children. E.g., a document originally associated with e.g., 15500000 "Dairy products" can therefore also belong to any of 15500000's children concepts, including 15510000 "Milk and cream, 15550000 "Assorted dairy products" and 15551310 "Unflavored yoghurt".

Two approaches, for adopting standard performance measures to hierarchically ordered categories, has been proposed by e.g., Sun and Lim (2001). They show that these extended precision and recall measurements, by including

Recall level	Precision Average
0.00	0.1485
0.10	0.1303
0.20	0.1051
0.30	0.0865
0.40	0.0661
0.50	0.0596
0.60	0.0412
0.70	0.0372
0.80	0.0365
0.90	0.0356
1.00	0.0356
Average precision:	0.0662

Table 1: Interpolated precision-recall values and average precision for baseline experiment

category similarity and *category distance*, do contribute positively compared to the standard precision and recall measurements.

6 Experiments

In the experiments we describe here, we provided each of the 739 documents described in section 4.1 to the model as queries. We used stop word removal, stemming and weighting by document frequency. We evaluated each experiment using interpolated precision-recall and average precision as we described in section 5.

6.1 Baseline

We set up a baseline for mapping each of the documents in the query collection to *any* of the concept labels in our ontology. This meant that for any given query, the system needed to find the correct concept(s) for this query among all 8323 sparsely described concepts in the ontology.

6.1.1 Results

Table 1 displays the interpolated precision-recall values and average precision for the baseline. As we had expected, the performance of the baseline was poor, only achieving an average precision of 6.62% over all queries.

We believed that the reason for the poor performance had to do with the large number of concepts in the ontology as well as the obvious fact that the concept labels in our ontology were so sparse and too selective, meaning that very few of them included the same terms as our queries.

Therefore, we set up three additional experiments that would 1) reduce the set of concepts, 2) enlarge parent concept labels with related terms from the ontology itself and 3) enrich the leaf concept labels with semantically similar terms from WordNet. We describe the outcomes of these three experiments in the following sections.

6.2 Varying levels

The baseline tried to find the correct concept(s) for a given query among all the 8323 concepts, the ontology defines. Since the ontology is taxonomically structured, an alternative approach was to map our queries only according to concepts on a certain level or with certain *granularity*. In this case, we could measure how good the model was at finding the correct *branch* of the ontology and not necessarily the correct concept.

6.2.1 Experiment setup

We set up this experiment as follows. First we needed to collect a subset of concepts against which the documents in the query collection should be mapped. Let us call this subset the *concept collection*. Since the idea was to cut the original ontology at eight different levels, the technique omitted all concepts below the selected level and only included concepts *on*, and *above* this level. As we select deeper levels to map against, the number of concepts in the concept collection increases, each time with number concepts on previous levels plus the number of concepts on the selected level. E.g., if we are mapping our documents according to level three we have 487 (12+97+378) concepts in the concept collection. The model then returns the most relevant *branches* to which that document belongs.

Table 2 shows the number of concepts on each level as well as the number of concepts in each concept collection.

It is clear that as we choose to map documents according to deeper levels, the task becomes more difficult since we get more and more possible concepts for each document. Mapping according to level 8 is therefore the same as the baseline task in section 6.1.

Since a document can be associated with a concept that is located on a lower level than we are mapping to, and since the technique omitted all these concepts from the concept

Level	No. of Concepts	Concept Collection
1	12	12
2	97	109
3	378	487
4	1022	1509
5	2048	3557
6	2420	5977
7	1636	7613
8	710	8323

Table 2: Number of concepts at various levels and above

collection, the system would not suggest these concepts. Therefore, we replaced all concepts in the gold standard that were below the selected level with their ancestor concept on the level selected. E.g., if a document had originally been associated with 29131000 and the level was set to two, this document would now instead be associated to 29000000 which is 29131000's ancestor on level two. If several of the associated concepts had the same ancestor (i.e., associated concepts were located in the same branch) then the document would only be associated with this parent once. Associated concepts that resided on higher levels than the selected level we left unchanged.

These preliminary steps resulted in eight modified gold standards and eight different concept collections that we mapped each of the 739 documents in the query collection against. No propagation of children label terms to parent labels was used in this experiment. We describe that experiment in section 6.4.

6.2.2 Results

Figures displayed in table 3 show the results we obtained for these experiments.

We thought that, by reducing the number of concepts in this fashion, it would be easier for our model to do the correct mappings. This was not the case. We were surprised about the numbers the model returned. Although we could see some small improvements, still, the performance was poor.

Early on when we started implementing this system, we realized that the labels describing the concepts in our ontology were too uninformative for our method to correctly map the business information texts to.

Level	Average Precision
1	0.1874
2	0.1048
3	0.0950
4	0.0838
5	0.0748
6	0.0716
7	0.0692
8	0.0662

Table 3: Interpolated Precision-recall values and Average Precision for various levels

In the following subsection we describe two techniques we hoped would bridge the gap between the terms in our queries and the terms in concept labels.

6.3 WordNet-enriched leaf concepts

We had previously developed a method (Warin et al., 2005), for enriching the *leaf concepts* in our ontology using WordNet (Fellbaum, 1998). This technique adds synonyms and introduce a broader terminology to these concepts. The method uses semantic similarity measures as disambiguation techniques for adding synset descriptions (glosses) to leaf concepts, including synonyms and other, less selective terms. E.g., the method enriched concept 15811200 "Rolls" with: "small rounded bread either plain or sweet bun, roll".

In this method, terms from leaf concepts and their parents are looked up in WordNet that returns their senses. The semantic similarity is then computed for the pair of word senses. The leaf sense with the highest (total) score is assumed to be a good candidate for enriching the leaf concept label. The outcome of this process provided us with a new ontology with 5366 out of the 5644 leaf concepts enriched.

6.3.1 Results

Neither in this experiment could we see any improvements compared to the baseline. The average precision only increased to about 6.74%. A possible explanation we found to this, as we also discuss in section 8, was that for the documents we looked at, few of the associated concepts in gold standard tended to be leaf concepts.

6.4 Term propagation

The next thing we tried was to enlarge parent concept labels with related terms from their children. By propagating children concept label terms upward in the ontology, parent concepts will become larger, and include all of their children concept labels. Thus, the parent concept labels will include a cluster of highly related terms and in effect, describe a complete branch in one single concept. Again, we envisioned that this would increase the probability that individual concepts would include terms also present in our queries.

6.4.1 Experiment setup

The technique we used for adding children's concept label terms to parent concept labels was straightforward. In a bottom-up fashion, starting at the finest granularity of the ontology, each concept label was added to each of that concept's parent label. The procedure then added these propagated parent labels to each of their parent labels and so on. This meant that each child's label was added to all of its ancestor labels. The idea was that this would constitute a sort of weighting for terms occurring on lower levels by the fact that on upper levels, these terms would become frequent.

E.g., the branch 29566000 originally look like this:

```

29566000 "Machinery for the treatment of
          sewage"
    29566100 "Comminutors"
        29566110 "Macerators for the
                  treatment of sewage"
    29566200 "Sewage presses"
    29566300 "Scrapers"
    29566400 "Mixer units"
    29566500 "Sewage screens"
    ...
    29566900 "Sludge-processing
              equipment"
    
```

After these concepts had been propagated, 29566000 and 29566100 now included all children label terms as well:

```

29566000 "Machinery for the treatment
          of sewage Macerators for the
          treatment of sewage Sedimentation
          beds Scrapers Sewage presses
          Precipitators Sewage screens
          Sludge-processing equipment
    
```

Comminutors Macerators for the treatment of sewage Oxygenation equipment Mixer units"

29566100 "Comminutors Macerators for the treatment of sewage"

The other concepts in the example were unaffected by the propagation since these are leaf concepts.

Again, we queried each of the documents in the query collection according to the 8323 concept labels, many of which we now had enlarged with related terms from their children concept labels.

6.4.2 Results

Again the results were poor, and the propagation technique was unable to improve the results significantly. Although this model achieved the highest average compared all previous experiments (6.93%) it was still very low. We explain this further in section 8.

7 Related work

The work most similar to ours in the literature is the work done by Ding et al. (2002). They introduce 'GoldenBullet', a system that classifies product descriptions according to the UN-SPSC classification scheme, which is similar to our ontology. They also view this as an information retrieval problem, treating product descriptions as queries and category labels in UN-SPSC as documents. Although they also use more sophisticated methods that are able to achieve fairly good results, they report of a classification accuracy of less than 1% for the method most comparable to ours. (Ding et al., 2002)

Although our method can be regarded as an information retrieval approach to automatic text categorization (cf. (Sebastiani, 2002)), the task we are facing is also related to work conducted within *hierarchical* text categorization. Most effort in this area has been put in to classifying text according to *web directories* (Labrou and Finin, 1999; Choi and Peng, 2004; Pulijala and Gauch, 2004), the *Reuters collection* (Sun and Lim, 2001; Weigend et al., 1999), as well as the automatic assignment of Gene Ontology (GO) terms to medical articles (Seki and Mostafa, 2005; Raychaudhuri et al., 2002; Kir-

Ontology	Relevance to Query Collection
Original	0.47
Propagated	0.45
WordNet Enriched	0.54

Table 4: Ontology and query collection similarity scores

itchenko et al., 2004; Ruiz and Srinivasan, 1999).

8 Qualitative study

The results that we obtained for the baseline were, although not satisfactory, not surprising. However, the results we obtained using the propagated and WordNet enriched ontologies were disappointing and puzzling.

In order to understand why the performance of our model was so low, and why neither the term propagation nor the WordNet enrichment technique provided us with any improvements, we did a small qualitative study of our data and the results that we had obtained.

The first thing we measured was the coverage between each ontology version and the documents in the query collection. Next, we did a closer inspection of the results to get an idea of how many of the labels in our gold standard did include terms that also were present in their associated document. If associated concept labels included terms that were not present in their associated document, even after term propagation or WordNet enrichment, it would also explain why the performance was low and why there was no significant improvements to the baseline.

The overall similarity between the ontologies we experimented on and the query collection is shown in table 4. We obtained these figures using the same vector model as we had used in our experiments. In this case we provided the model with the complete query collection as a single query to compare against the complete ontology. We adopted this method from Brewster et al. (Brewster et al., 2004) who describe this method in the framework of ontology evaluation for measuring the 'fit' between an ontology and a domain corpus it is suppose to model.

The figures in table 4 show that for all ontologies, there was a clear coverage of concept label terms in the query collection. Interestingly,

the coverage of the propagated ontology was in fact smaller than for the original ontology, although only slightly. We were glad to see that our WordNet-enrichment technique did provide us with an ontology that more closely resembled the query collection.

The reason for propagated ontology having a similar coverage as the original ontology is not so surprising. The label terms in the propagated ontology are the same as those that describe concepts in the original ontology, and since the propagation technique did not add any new terms to the ontology, it did not provide us with an ontology which resembled our query collection more closely than the original ontology.

The figures in table 4 do not seem to explain why the performance was so poor for the experiments that we did. We therefore turned to the gold standard to see how well the label terms in our gold standard covered the terms in their associated documents. This would give us an indication of how well individual concept labels covered terms in our queries.

To clarify: what we would like is, of course, that each term in an associated label also is present in its respective document. A good example is the document 362680 displayed in section 4.1 and its associated concept labels we showed in section 4.2. In this example, each associated label includes terms that also frequently occur in the text. When we inspected the results for 362680, the model had accurately suggested all three correct concepts at position, 3, 4, and 14 in the ranked list:

```
...
31300000 Insulated wire
           and cable
28400000 Cable, wire and
           related products.
...
31330000 Coaxial cable.
```

However, we found that when we ran the experiments on the original ontology, 153 documents in our query collection did not include any of the terms in the labels with which they had been associated. E.g., document 171651 includes the following text: "on the unique experiences of this leading company of greenhouse Climate and better crops".

According to the gold standard, 171651 should be mapped to the concept 45211350 "Multi-functional buildings".

Secondly, it seemed that the terms that the propagation and enrichment techniques had added the associated concept labels with, in a number of cases, simply were not present in the document. E.g., concept 45211350 "Multi-functional buildings" was enriched with "the occupants of a building building". In this case, the only term introduced was occupant.

Also with the propagated ontology, cases like this could be observed. E.g., document 755728, originally associated with "Beauty products" and "Perfumes and toiletries", included only the label term perfume. After the ontology had been propagated, the following terms were added to 755728's associated labels: toiletries, shaving, preparations, shampoos, manicure or pedicure, preparations, toilet, waters, hair, preparations, beauty, skin-care, antiperspirants, deodorants, make-up, preparations, oral, dental, and hygiene. Not a single one of these related terms were present in the document.

For cases like these, it is easy to explain why we observed only a small increase in performance of our model, after our ontology had been enriched or enlarged.

Further, for those cases where the enrichment or propagation technique had added terms to the associated concept label(s) and those terms were also *present* in the respective document, the study indicated that what we had added was *low frequency* terms. E.g., when the associated concept 29433000 "Bending, folding, straightening or flattening machines" had been enriched with

"any mechanical or electrical device that transmits or modifies energy to perform or assist in the performance of human tasks machine",

only the term mechanical, occurred in the document and with a frequency of *one*. Again we could observe similar patterns when experi-

menting on the propagated CPV. E.g., document 14043 was originally associated to:

```
29474000 "Parts and accessories for
          metal-working machine tools"
29462000 "Welding equipment"
52000000 "Retail trade services"
74700000 "Cleaning services"
29423000 "Milling machines"
74230000 "Engineering services"
```

Only *one* of these label terms occurred in the document namely, engineering. The propagation technique then enlarged the associated labels above with 251 (unique) children terms, resulting in 360 associated label terms for this document. Measuring the coverage again, now six additional label terms could be found in the document (item, transport, process, support, control, and design). So, not only was just a small fraction (6/351) of the terms the propagation technique had added actually in 14043, in addition, all these terms occurred only once or at most three times in the document.

We concluded from this study that the texts in our query collection were simply too uninformative for our model to achieve good results. Although our ontologies seemed to model the business information domain well *overall*, there was too big of a difference between *individual* concepts and texts. The number of documents in our gold standard that did not include *any* of its associated concept label terms were 153 for the original ontology. Although we saw that these cases decreased as we enriched or enlarged the ontology, the tendency was that only few of the added concept terms were actually in the individual documents, and for those that were, instead, they were too infrequent anyway.

Another explanation for why no real improvement could be seen using the propagated ontology was that, by propagating the terms from children labels to all parents, the technique introduced these terms to a large number of other concept labels. This distributed label terms across large portions of the ontology, that in effect made the concepts more similar to each other. In fact, mapping the documents in our query collection on to the propagated ontology generated 10.000 more answers than when we mapped them to the original ontology.

A positive outcome however was that we did get confirmation that the WordNet enrichment

technique did provide us with an ontology that more closely resembled the query collection.

It is important to note that we only enriched leaf concepts, and although leaf concepts make up the majority of the concepts, for this technique to have effect, not only must we have enriched the concepts with correct and useful terms, but also, the documents we are mapping need to be associated with these leaf concepts in the gold standard. We saw several cases were only a few of the associated labels for a document were leaf concepts. Similarly, since the term propagation only affects parent concepts, and leaf concepts are left unaffected and since they constitute the majority of concepts, it could explain why we saw only little improvement. As was true for the enriched concepts, for the propagation technique also to have a real effect, not only does the added concept terms need to be in the document, but the documents also needs to be associated with enlarged concepts (i.e., parent concepts).

9 Future work

To assess the pros and cons of our baseline method as well our other techniques, we need to do a much larger qualitative study than we did for these experiments. But before we do that, we need to collect texts that are more informative than those documents we currently have in our collection.

>From the small qualitative study we did, it is clear that terms describing the concepts in our ontology included few terms used in real business information texts. An appropriate next step will therefore be to enrich the ontology with terms from such real world business information texts.

The results we have reported here are based on exact matches. However, we have observed many cases where the model has either suggested a more general concept (parent) to a correct concept, but even more so a child concept to a correct concept. If the model suggests such closely related concepts, than it should count for something. In future versions of this system, we will regard cases like these as partially correct by giving them a penalty depending on their distance to the correct concept in the ontology.

10 Conclusions

In this paper, we have described a system that adopts the vector space model in the framework of automatically assigning concept labels to business information texts, by mapping these texts to a large ontology defining a wide range of products and services. We envisioned the task to be a difficult one, because of the number of concepts in the ontology and the sparse labels that describe them. However, the task proved to be more challenging than we had anticipated.

It became clear that the business information texts on which we tested our model, were of too poor quality for our task, something we simply could not do anything about, regardless of the experiment we conducted. Either the texts were too short (too uninformative) or included too much non-sense text. We tried several techniques for improving on the baseline, including the *reduction of concepts*, *enlargement of concept labels with related terms*, and the *enrichment of new terms* with the help of WordNet. We were able to show that the latter technique, by introducing many *non-present* terms to the ontology, did yield an ontology that more closely resembled the texts we tried to map overall, and that our improvement techniques did allow our model to achieve a higher accuracy. Still, the individual concepts labels in our enlarged and enriched ontologies included terms that rarely occurred in our queries.

To bridge the gap between the selective, generic vocabulary describing concepts in our ontology and the specific terminology used in online business information texts, in the future, we intend to develop an accurate method that instead enriches the ontology with such vocabulary.

References

- Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern information retrieval*. ACM Press, New York.
- Christopher Brewster, Harith Alani, Srinandan Dasmahapatra, and Yorick Wilks. 2004. Data-driven ontology evaluation. *Proceedings of the 4th International Conference on Language Resources and Evaluation*.
- Ben Choi and Xiaogang Peng. 2004. Dynamic and hierarchical classification of web pages. *Online Information Review*, Vol. 28, No. 2:139–147.
- Ying Ding, Maksym Korotkiy, Borys Ome-layenko, Vera Kartseva, Volodymyr Zykov, Michel Klein, Ellen Schulten, and Dieter Fensel. 2002. Goldenbullet in a nutshell. *Proceedings of the Fifteenth International Florida Artificial Intelligence Research Society Conference*, pages 403–407.
- European Union and European Parliament. 2002. European union and european parliament. on the common procurement vocabulary (cpv). *Regulation (EC) no 2195/2002 of the European Parliament and of the Council*.
- Christiane D. Fellbaum. 1998. *WordNet, an electronic lexical database*. MIT Press.
- Svetlana Kiritchenko, Stan Matwin, and Fazel Famili. 2004. Hierarchical text categorization as a tool of associating genes with gene ontology codes. *Proceedings of the Second European Workshop on Data Mining and Text Mining for Bioinformatics (held at ECML-04)*, pages 26–30.
- Yannis Labrou and Tim Finin. 1999. Yahoo! as an ontology: using Yahoo! categories to describe documents. *Proceedings of CIKM-99, 8th ACM International Conference on Information and Knowledge Management*, pages 180–187.
- Ashwin Pulijala and Susan Gauch. 2004. Hierarchical text classification. *International Conference on Cybernetics and Information Technologies, Systems and Applications: CITSA 2004, Vol. 1.*, pages 257–262.
- Soumya Raychaudhuri, Jeffrey T. Chang, Patrick D. Sutphin, and Russ B. Altman. 2002. Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Research*, 12(1):203–214, January.
- Miguel E. Ruiz and Padmini Srinivasan. 1999. Hierarchical neural networks for text categorization. *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, pages 281–282.
- Gerard Salton, A. Wong, and C.S. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.
- Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Kazuhiro Seki and Javed Mostafa. 2005. An application of text categorization methods to gene ontology annotation. *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 138–145.

Aixin Sun and Ee-Peng Lim. 2001. Hierarchical text classification and evaluation. *Proceedings of ICDM-01, IEEE International Conference on Data Mining*, pages 521–528.

Martin Warin, Henrik Oxhammar, and Martin Volk. 2005. Enriching an ontology with wordnet based on similarity measures. *Proceedings of the MEANING-2005 Workshop*.

Andreas S. Weigend, Erik D. Wiener, and Jan O. Pedersen. 1999. Exploiting hierarchy in text categorization. *Information Retrieval*, 1(3):193–216.