

Ontology-Based Document and Query Representation May Improve the Effectiveness of Information Retrieval

Mauro Dragoni and Célia da Costa Pereira and Andrea G.B. Tettamanzi¹

Abstract. This paper presents a vector space model approach to representing documents and queries, which considers concepts instead of terms and uses WordNet as a light ontology. This representation reduces information redundancy with respect to conventional semantic expansion techniques. Experiments carried out on the MuchMore benchmark and on the TREC-7 and TREC-8 Ad-hoc collections demonstrate the effectiveness of the proposed approach.

1 Introduction

This paper presents an ontology-based approach to the conceptual representation of documents. Such an approach is inspired by a recently proposed idea presented in [8], and uses an adapted version of that method to standardize the representation of documents and queries. The proposed approach bears some similarity to the classic query expansion technique. However, additional considerations have been taken into account and some improvements have been applied as explained below.

Query expansion is an approach used in Information Retrieval (IR) in order to improve the system's performance. It consists of the expansion of the content of the query by adding the terms that are semantical correlated with the original terms of the query. Several works demonstrated the enhanced performance of IR systems that implement query expansion approaches [22] [3]. However, the query expansion approach has to be used carefully because, as demonstrated in [7], expansion might degrade the performance of some individual queries. This is due to the fact that an incorrect choice of terms and concepts for the expansion task might harm the retrieval process by drifting it away from the optimal correct answer [1].

Document expansion applied to IR has been recently proposed in [2]. In that work, a sub-tree approach has been implemented to represent concepts in documents and queries. However, when using a tree structure, there is a redundancy of information because more general concepts may be represented implicitly by using only the leaf concepts they subsume.

This paper presents a new representation for documents and queries. The proposed approach exploits the structure

of the well-known WordNet machine readable dictionary in order to reduce the redundancy of information generally contained in a concept-based document representation. The second improvement is the reduction of the computational time needed to compare documents and queries represented using concepts. This representation has been applied to the *ad-hoc* retrieval problem. The approach has been evaluated on the Muchmore benchmark and on the TREC-7 and TREC-8 Ad-hoc collection, and the results demonstrate its viability.

The paper is organized as follows. In Section 2 an overview of the environment in which ontology has been used is presented. Section 3 presents the tools used for this work. Section 4 illustrates the proposed approach to represent information, while Section 5 compares this approach with other two well-known approaches used in conceptual representation of documents. In Section 6 the experimental results are discussed. Finally, Section 7 concludes.

2 Related Works

An increasing number of recent information retrieval systems make use of ontologies to help the users clarify their information needs and come up with semantic representations of documents. Many ontology-based information retrieval systems and models have been proposed in the last decade. An interesting review on IR techniques based on ontologies is presented in [10]. Model for the exploitation of ontology-based knowledge bases are presented in [6] and [19]. The aim of these models is to improve search over large document repositories. Both models include an ontology-based scheme for the annotation of documents, and a retrieval model based on an adaptation of the classic vector-space model.

The implementation of ontology models has been also investigating using fuzzy models, an approaches is presented in [5].

Ontology-based semantic retrieval is very useful for specific-domain environments. A general IR system to facilitate specific domain search is illustrated in [13]. The system uses fuzzy ontologies and is based on the notion of information granulation, a novel computational model is developed to estimate the granularity of documents. The presented experiments confirm that the proposed system outperforms a vector space based IR system for domain specific search.

In IR, the user's input queries usually are not detailed enough, so the satisfactory query results can not be brought

¹ Università degli Studi di Milano, Dipartimento di Tecnologie dell'Informazione, Via Bramante 65, I-26013 Crema (CR), Italy, email: mauro.dragoni,celia.pereira,andrea.tettamanzi@unimi.it

back. Query expansion of IR can help to solve this problem. However, the common query expansion in IR cannot get steady retrieval results. Ontologies play a key role in query expansion research. A common use of ontologies in query expansion is to enrich the resources with some well-defined meaning to enhance the search capabilities of existing web searching systems.

In [21] the authors propose and implement query expansion method which combines domain ontology with the frequent of terms. Ontology is used to describe domain knowledge; logic reasoner and the frequency of terms are used to choose fitting expansion words. This way, higher recall and precise can be gotten as user's query results.

A natural language processing approach is presented in [14]. In this work the authors have developed ontology-based query processing to improve the performance of design information retrieval. In [9] the authors present an approach to expand queries that consists in searching terms from the topic query in an ontology in order to add similar terms.

One of the vital problems in the searching for information is the ranking of the retrieved results. Users make typically very short queries and tend to consider only the first ten results. In traditional IR approaches the relevance of the results is determined only by analyzing the underlying information repository. On the other hand, in the ontology-based IR the querying process is supported by an ontology. In [17] a novel approach for determining relevance in ontology-based searching for information is presented.

3 Preliminaries

The roadmap to prove the viability of a concept-based representation of documents and queries consists in two main tasks: to choose a method that permits to represent all documents terms using the same set of concepts and to implement an approach that permits to index and to evaluate each concept, in both documents and queries, with an appropriate weight.

To represent documents, the method described in Section 4 has been used, combined with the use of the WordNet Machine-Readable Dictionary (MRD). From the WordNet database, the set of terms that do not have hyponymy has been extracted, each term is named "base concept". A vector, named "base vector", has been created and, to each component of the vector, a base concept has been assigned. This way, each term is represented using the base vector of the WordNet ontology.

The representation described above has been implemented on top of the Apache Lucene open-source API.²

In the pre-indexing phase, each document has been converted in its ontology representation. After the calculation of the importance of each concept in a document, only concepts with a degree of importance higher than a fixed cut-value have been maintained, while the others have been discarded. The cut-value used in these experiments is 0.01. This choice has a little drawback, namely that an approximation of representing information is introduced due to the discard of some minor concepts. However, we have experimentally verified that this approximation does not affect the final results.

During the evaluation activity, queries have been also converted into the ontological representation. This way, weights

have to be assigned to each concept to evaluate all concepts with the right proportion. One of the features of Lucene is the possibility of assigning a boost to each term of the query. Therefore, to each element present in the concept-based representation of the query, the assigned boost value is the concept weight.

4 Document Representation

Conventional IR approaches represent documents as vectors of term weights. Such representations use a vector with one component for every significant term that occurs in the document. This has several limitations, for example:

1. different vector positions may be allocated to the synonyms of the same term; this way there is an information loss because the importance of a determinate *concept* is distributed among different vector components;
2. the size of a document vector have to be at least equal to the total number of words of the language used to write the document;
3. every time a new set of terms is introduced (which is a high-probability event), all document vectors must be re-constructed; the size of a repository thus grows not only as a function of the number of documents that it contains, but also of the size of the representation vectors.

To overcome these weaknesses of term-based representations, an ontology-based representation has been used [15].

An ontology-based representation has been recently proposed in [8] which exploits the hierarchical *is-a* relation among concepts, i.e., the meanings of words. For example, to describe with a term-based representation documents containing the three words: "animal", "dog", and "cat" a vector of three elements is needed; with an ontology-based representation, since "animal" subsumes both "dog" and "cat", it is possible to use a vector with only two elements, related to the "dog" and "cat" concepts, that can also implicitly contain the information given by the presence of the "animal" concept. Moreover, by defining an ontology base, which is a set of independent concepts that covers the whole ontology, an ontology-based representation allows the system to use fixed-size document vectors, consisting of one component per base concept.

Calculating term importance is a significant and fundamental aspect for representing documents in conventional information retrieval approaches. It is usually determined through term frequency-inverse document frequency (TF-IDF). When using an ontology-based representation, such usual definition of term-frequency cannot be applied because one does not operate by keywords, but by concepts. This is the reason why it has been adopted the document representation based on concepts proposed in [8], which is a concept-based adaptation of TF-IDF.

The quantity of information given by the presence of concept z in a document depends on the depth of z in the ontology graph, on how many times it appears in the document, and how many times it occurs in the whole document repository. These two frequencies also depend on the number of concepts which subsume or are subsumed by z . Let us consider a concept x which is a descendant of another concept y which has q children including x . Concept y is a descendant of a concept z which has k children including y . Concept x is a leaf of the

² See URL <http://lucene.apache.org/>.

graph representing the used ontology. For instance, considering a document containing only “ xy ”, the occurrence of x in the document is $1 + (1/q)$. In the document “ xyz ”, the occurrence of x is $1 + (1/q(1 + 1/k))$. As it is possible to see, the number of occurrences of a leaf is proportional to the number of children which all of its ancestors have. Explicit and implicit concepts are taken into account by using the following formulas:

$$N(c) = \text{occ}(c) + \sum_{c \in \text{Path}(c, \dots, \tau)} \sum_{i=2}^{\text{depth}(c)} \frac{\text{occ}(c_i)}{\prod_{j=2}^i |\text{children}(c_j)|}, \quad (1)$$

where $N(c)$ is the number of occurrences, both explicit and implicit, of concept c and $\text{occ}(c)$ is the number of lexicalizations of c occurring in the document.

Given the ontology base $I = b_1, \dots, b_n$, where the b_i s are the base concepts, the quantity of information, $\text{info}(b_i)$, pertaining to base concept b_i in a document is:

$$\text{info}(b_i) = \frac{N_{\text{doc}}(b_i)}{N_{\text{rep}}(b_i)}, \quad (2)$$

where $N_{\text{doc}}(b_i)$ is the number of explicit and implicit occurrences of b_i in the document, and $N_{\text{rep}}(b_i)$ is the total number of its explicit and implicit occurrences in the whole document repository. This way, every component of the representation vector gives a value of the importance relation between a document and the relevant base concept.

A concrete example can be explained starting from the ontology represented in Figures 1 and 2, and considering a document D_1 containing concepts “ $xyyyz$ ”.

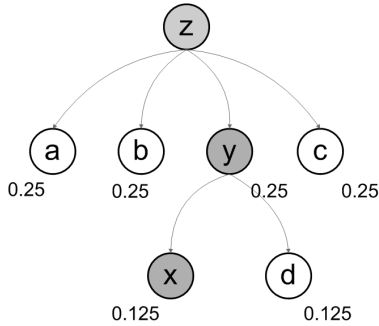


Figure 1. Ontology representation for concept ‘ z ’.

In this case the ontology base is:

$$I = \{a, b, c, d, x\}$$

and, for each concept in the ontology, the vectors N_{doc} are:

$$\begin{aligned} \vec{z} &= (0.25, 0.25, 0.25, 0.125, 0.125) \\ \vec{a} &= (1.0, 0.0, 0.0, 0.0, 0.0) \\ \vec{b} &= (0.0, 1.0, 0.0, 0.0, 0.0) \\ \vec{c} &= (0.0, 0.0, 1.0, 0.0, 0.0) \\ \vec{y} &= (0.0, 0.0, 0.0, 0.5, 0.5) \\ \vec{d} &= (0.0, 0.0, 0.0, 1.0, 0.0) \\ \vec{x} &= (0.0, 0.0, 0.0, 0.0, 1.0) \end{aligned}$$

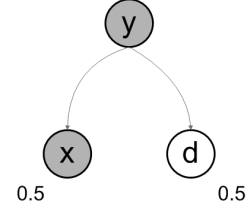


Figure 2. Ontology representation for concept ‘ y ’.

so the document vector associated to D_1 is:

$$D_1 = (2 * \vec{x}) + (3 * \vec{y}) + \vec{z} = (0.25, 0.25, 0.25, 1.625, 3.625). \quad (3)$$

In Section 5, a comparison between the proposed representation and other two classic concept-based representation is discussed.

4.1 Issues about Verbs, Adjectives, and Proper Names

The representation described above is chiefly suited to representing nouns. However, a different representation is in order to handle verbs, adjectives, and proper-names.

In WordNet, verbs and adjectives are structured in a different way than nouns. The role of the hyperonymy and hyponymy relations (that make MRD comparable to light ontologies) is different for verbs and adjectives [11, 18]. To overcome this issue, we have exploited the “derivationally related form” relation existing in WordNet. This kind of relation links each verb and adjective to the semantically closest noun sense. By such device, for each verb and adjective, the semantically correlated noun sense can be extracted. This enables us to represent the verb (or adjective) information in the same way as nouns. Examples of “derivationally related form” verb-noun relations are reported in Table 1.

Verb	Noun sense
navigate#1	voyage, voyager, sail, navigation
drink#1	drink, drinker, imbibitor
Adjective	Noun sense
high#1	highness
small#1	smallness, littleness

Table 1. Examples of “derivationally related forms” relations.

A similar approach has been followed for proper-names. These entities, which are part of the WordNet dictionary, are not linked in the WordNet hyperonymy/hyponymy light ontology. All these entities have an “instance of” relationships with nouns that describes the kind of the entity. It is then possible to represent each proper-name by using the concept base associated to the noun linked to it through the “instance of” relationship. Examples of “instance of” relationships are reported in Table 2.

Of course, the issue of proper names is much more complicated than that, and we consider this but a preliminary

Proper-name	"Instance Of" Noun
Yellowstone	river
George Bush	President of the United States

Table 2. Examples of "instance of" relations.

approximation to a satisfactory solution for handling them, whose main purpose is to enable us to run experiments on a collection of real-world documents and queries, which are highly likely to contain proper names, besides nouns, verbs, and adjectives.

5 Representation Comparison

In Section 4, the approach used to represent nouns, verbs, adjectives, and proper-names has been described. In this section, we aim at illustrating the improvements obtained, by applying the proposed approach, with respect to the information redundancy and computational time. We also make a comparison between the results obtained by using the proposed approach and those obtained by other two approaches commonly used in conceptual document representation.

Information Redundancy Approaches that apply the expansion of documents and queries use correlated concepts to expand the original terms of documents and queries. A problem with expansion is that information is redundant and there is not a real improvement of the representation of the document (or query) content. With the proposed representation this redundancy is eliminated because only independent concepts are taken into account to represent documents and queries. Another positive aspect is that the size of the vector representing the document content by using concepts is generally smaller than the size of the vector representing document content by using terms.

In [12], the authors propose an indexing technique that takes WordNet synsets into account instead of terms. For each term in documents, the synsets associated to that term are extracted and then used as a token for the indexing task. Therefore, the computational time needed to perform a query does not increase. However, there is a significant overlap of information because different synsets might be semantically correlated, as introduced in Section 4.

Computational Time When IR approaches are applied in a real-world environment, the computational time needed to evaluate the match between documents and the submitted query has to be considered. It is known that systems using the vector space model have higher efficiency. Conceptual-based approaches, such as the one presented in [2], generally implement a non-vectorial data structure which needs a higher computational time than a vector space model representation does. The approach proposed in this paper overcomes this issue, because the document content is represented by using a vector and, therefore, the computational time needed to compute the document score is comparable to the computational time needed when using the vector space model.

6 Experiments

In this section, the impact of the ontology-based document and query representation is evaluated. The experiments have been divided in two different phases:

1. in the first phase, the proposed approach has been compared to the most well-known state of the art kinds of semantic expansion techniques: document representation by synsets and document representation by semantic trees;
2. in the second phase, the proposed approach has been validated with systems that use semantic expansion presented at the TREC7 and TREC8 conferences.

The evaluation method follows the TREC protocol [20]. For each query, the first 1,000 documents have been retrieved and the precision of the system has been calculated at different points: 5, 10, 15, and 30 documents retrieved. Moreover, the Mean Average Precision of the system has been calculated. The document assessment has been computed by adopting the Vector Space Model with the slightly variance of using the Conceptual-IDF proposed in [8] instead of the classic IDF.

The first part of the experiments has been performed by using the MuchMore³ collection [4], which consists of 7,823 abstracts of medical papers and 25 queries with their relevance judgments. Table 3 illustrates the results obtained by applying the synset indexing approach, the semantic approach proposed by [2], and the approach proposed in this paper. We have considered the classic term-based representation as baseline.

Systems	Precisions				
	P5	P10	P15	P30	MAP
Baseline	0.544	0.480	0.405	0.273	0.449
Synset Indexing proposed by [12]	0.648	0.484	0.403	0.309	0.459
Conceptual Indexing proposed by [2]	0.770	0.735	0.690	0.523	0.449
Proposed Ontology Indexing approach	0.784	0.765	0.728	0.594	0.477

Table 3. Comparisons table between semantic expansion approaches.

The results show that the proposed method obtains better performances than the two approaches it is compared to. We can notice that the improvement increases with the increment of the retrieved documents considered.

The second part of these experiments has been performed by using the TREC collections. In particular, the TREC Ad-Hoc Collection Volumes 4 and 5 (containing over 500,000 documents) has been used. The approach has been evaluated on topics from 351 to 450. These topics correspond to two editions of the TREC conference, namely TREC-7 and TREC-8.

The approach is also compared to the approaches presented in the TREC-7 and TREC-8 conferences.

For each conference, dozens of runs have been submitted; therefore we have chosen the three systems implementing a semantic expansion that obtained higher precision values at

³ <http://muchmore.dfki.de>

lower recall levels. The rationale behind this decision is the fact that the majority of search result click activity (89.8%) happens on the first page of search results [16], that is, generally, users only consider the first 10 to 20 documents.

Another aspect that we have taken into account is the way queries are composed by each system and which kind of information has been used to do that. Two possible query composition methods are used in the TREC conferences: manual and automatic. Queries are formed completely automatically if the used software already exists at the time of query evaluation; in all other cases, the queries are considered to be manual. Automatic queries provide a reasonably well controlled basis for cross-system comparison, although they are typically representative of only the first query in an interactive search process. On the contrary, manual queries are used to demonstrate the retrieval effectiveness that can be obtained after interactive optimization of the query. Examples of manual queries are queries in which stop words or stop structure are manually removed.

Each topic (query) is composed of three main fields: title, description, and narrative. A query might consist of one or more of these fields. The proposed approach builds queries using only the title and the description fields; therefore, it has been compared only to systems that used the same fields. Because documents are represented using an ontology, also each topic has been converted into the corresponding ontological representation.

Systems	Precisions				
	P5	P10	P15	P30	MAP
Term-Based Representation	0.444	0.414	0.375	0.348	0.199
AT&T Labs Research (att98atdc)	0.644	0.558	0.499	0.419	0.296
AT&T Labs Research (att98atde)	0.644	0.558	0.497	0.413	0.294
City University, Univ. of Sheffield, Microsoft (ok7am)	0.572	0.542	0.507	0.412	0.288
Proposed Approach	0.656	0.588	0.501	0.397	0.309

Table 4. Precision@X and Mean Average Precision results obtained on TREC7 Topics.

Systems	Precisions				
	P5	P10	P15	P30	MAP
Term-Based Representation	0.476	0.436	0.389	0.362	0.243
IBM T.J. Watson Research Center (ibms99a)	0.588	0.504	0.472	0.410	0.301
Microsoft Research Ltd (ok8amxc)	0.580	0.550	0.499	0.425	0.317
TwentyOne (tno8d3)	0.500	0.454	0.433	0.368	0.292
Proposed Approach	0.616	0.572	0.485	0.415	0.315

Table 5. Precision@X and Mean Average Precision results obtained on TREC8 topics.

The precision/recall graph shown in Figure 3 illustrates the comparison between the proposed approach (heavy gray curve), the classical term-based representation (black curve), and the three systems presented at the TREC-7 Ad-Hoc Track (light gray curves). As expected, for all recall values, the proposed approach obtained better results than the term-based representation.

By comparing the proposed approach with the three TREC-7 systems, we can notice that the results obtained by our approach are better than the results obtained by the other approaches. Indeed, we obtained better results for the recall levels between 0.0 and 0.4, the best results being at recall levels 0.0 and 0.2. At recall levels 0.5 up to 1, the proposed approach is slightly worst, but substantially in line with the other concept-based approaches. A possible explanation for this scenario is that, for documents that are well related to a particular topic, the adopted ontology-based representation is able to improve the representation of the document contents. However, for documents that are partially related to a topic or that contain many ambiguous terms, the proposed approach is not able to maintain a high precision of the results. At the end of this section, a couple of improvements that may overcome this issue are discussed.

In Table 4, all systems are compared for the Precision@X and MAP values. The results confirm that the proposed approach obtains better results for the top 10 retrieved documents. Indeed, the values for Prec@5 and Prec@10 are the best results. The same consideration holds for the MAP value. However, the Prec@15 value is in line with the other systems, while the Prec@30 value does not outperform the values obtained by the three TREC-7 systems.

The same evaluations have been carried out for the topics of the TREC-8 Ad-Hoc Track. The precision/recall graph in Figure 4 shows how the concept-based representation curve approaches and overtakes the curves of the three TREC-8 systems for recall levels between 0.0 and 0.4. The behavior of the proposed approach is similar to the one shown by using the TREC-7 topics, however, in this case the gain is reduced. It is also interesting to observe that, with the TREC-8 topics, the results of all presented systems are closer to the ones obtained on the TREC-7 topics.

The Precision@X and the MAP values shown in Table 5 confirm the impression described above.

Inspecting the precision/recall curve obtained by the system with both TREC-7 and TREC-8 topics, we can notice that the performance of the system decreases in both cases. We think that this situation can mainly be due to two reasons:

- Absence of some terms in the ontology: some terms, in particular terms related to specific domains (biomedical, mechanical, business, etc.), are not defined in the MRD because WordNet does not cover all possible terms of the English language. Therefore, in some cases there is a loss of information that can affect the final retrieval result.
- Term ambiguity: the concept-based representation can introduce an error given by not using a word sense disambiguation algorithm. This could cause concepts associated to incorrect senses being discarded or weighted less.

We are convinced that improving the actual model with the above considerations would yield significantly better results in forthcoming experiments. This positive view is motivated by

the fact that, by expanding semantically each term, the ambiguity plays a significant role in the representation of document content.

7 Conclusion

This paper has discussed an approach to indexing documents and to representing queries for information retrieval purposes which exploits a conceptual representation based on ontologies. Experiments have been performed on the MuchMore Collection and on TREC Ad-Hoc Collection to validate the approach with respect to problems like term-synonymity in documents.

Preliminary experimental results show that the proposed representation improves the ranking of the documents in the first positions. Investigation on results highlights that further improvement could be obtained by integrating WSD techniques to avoid the error introduced by considering incorrect word senses, and a better usage of linguistic resources to overcome the absence of domain-specific terms.

References

- [1] A. Abdelali, J. Cowie, and H. Soliman. Improving query precision using semantic expansion. *Inf. Process. Manage.*, 43(3):705–716, 2007.
- [2] M. Baziz, M. Boughanem, G. Pasi, and H. Prade. An information retrieval driven by ontology: from query to document expansion. In *RIAIO. CID*, 2007.
- [3] B. Billerbeck and J. Zobel. Techniques for efficient query expansion. In *SPIRE*, volume 3246 of *Lecture Notes in Computer Science*, pages 30–42. Springer, 2004.
- [4] M. Boughanem, T. Dkaki, J. Mothe, and C. Soulé-Dupuy. Mercure at trec7. In *TREC*, pages 355–360, 1998.
- [5] S. Calegari and E. Sanchez. A fuzzy ontology-approach to improve semantic information retrieval. In *URSW*, volume 327 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2007.
- [6] P. Castells, M. Fernández, and D. Vallet. An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Trans. Knowl. Data Eng.*, 19(2):261–272, 2007.
- [7] S. Cronen-Townsend, Y. Zhou, and W. Croft. A framework for selective query expansion. In *CIKM*, pages 236–237. ACM, 2004.
- [8] C. da Costa Pereira and A. G. B. Tettamanzi. *Soft computing in ontologies and semantic Web*, chapter An ontology-based method for user model acquisition, pages 211–227. Studies in fuzziness and soft computing. Ed. Zongmin Ma, Springer, Berlin, 2006.
- [9] M. Díaz-Galiano, M. G. Cumberras, M. Martín-Valdivia, A. M. Ráez, and L. Ureña-López. Integrating mesh ontology to improve medical information retrieval. In *CLEF*, volume 5152 of *Lecture Notes in Computer Science*, pages 601–606. Springer, 2007.
- [10] O. Dridi. Ontology-based information retrieval: Overview and new proposition. In O. Pastor, A. Flory, and J.-L. Cavarero, editors, *RCIS*, pages 421–426. IEEE, 2008.
- [11] C. Fellbaum and G. Miller. Folks psychology or semantic entailment? a reply to rips and conrad. *The Psychology Review*, 97:565–570, 1990.
- [12] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarrán. Indexing with wordnet synsets can improve text retrieval. *CoRR*, cmp-lg/9808002, 1998.
- [13] R. Lau, C. Lai, and Y. Li. Mining fuzzy ontology for a web-based granular information retrieval system. In *RSKT*, volume 5589 of *Lecture Notes in Computer Science*, pages 239–246. Springer, 2009.
- [14] Z. Li and K. Ramani. Ontology-based design information extraction and retrieval. *AI EDAM*, 21(2):137–154, 2007.
- [15] L. Shen, Y. Lim, and H. Loh. Domain-specific concept-based information retrieval system. In *Proceedings of the IEEE International Engineering Management Conference*, volume 2, pages 525–529, 2004.
- [16] A. Spink, B. Jansen, C. Blakely, and S. Koshman. A study of results overlap and uniqueness among major web search engines. *Inf. Process. Manage.*, 42(5):1379–1391, 2006.
- [17] N. Stojanovic. An approach for defining relevance in the ontology-based information retrieval. In *Web Intelligence*, pages 359–365. IEEE Computer Society, 2005.
- [18] L. Talmy. Lexicalization patterns: Semantic structure in lexical forms. *Language Typology and Syntactic Description*, 3:57–149, 1985.
- [19] D. Vallet, M. Fernández, and P. Castells. An ontology-based information retrieval model. In *ESWC*, volume 3532 of *Lecture Notes in Computer Science*, pages 455–470. Springer, 2005.
- [20] E. Voorhees and D. Harman. Overview of the sixth text retrieval conference (trec-6). In *TREC*, pages 1–24, 1997.
- [21] F. Wu, G. Wu, and X. Fu. Design and implementation of ontology-based query expansion for information retrieval. In *CONFENIS (1)*, volume 254 of *IFIP*, pages 293–298. Springer, 2007.
- [22] J. Xu and W. Croft. Query expansion using local and global document analysis. In *SIGIR*, pages 4–11. ACM, 1996.

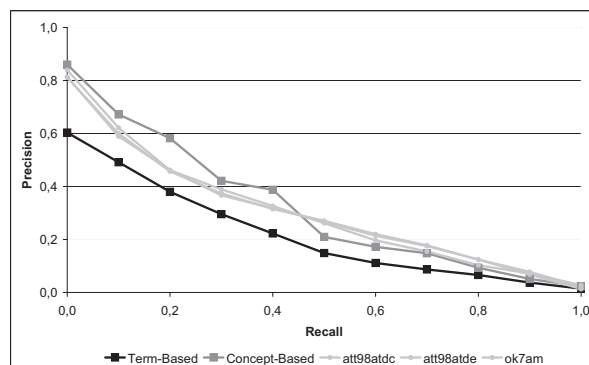


Figure 3. Precision/recall graph for TREC-7 topics.

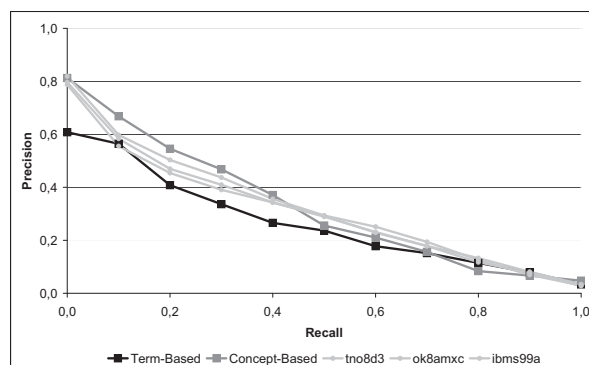


Figure 4. Precision/recall graph for TREC-8 topics.