An Ontological Representation of Documents and Queries for Information Retrieval Systems

Mauro Dragoni
Università degli Studi di Milano
Dipartimento di Tecnologie dell'Informazione
Via Bramante 65, I-26013
Crema (CR), Italy
mauro.dragoni@unimi.it

Célia da Costa Pereira
Università degli Studi di Milano
Dipartimento di Tecnologie dell'Informazione
Via Bramante 65, I-26013
Crema (CR), Italy
celia.pereira@unimi.it

Andrea G.B. Tettamanzi
Università degli Studi di Milano
Dipartimento di Tecnologie dell'Informazione
Via Bramante 65, I-26013
Crema (CR), Italy
andrea.tettamanzi@unimi.it

ABSTRACT

This paper presents a vector space model approach, for representing documents and queries, using concepts instead of terms and WordNet as a light ontology. This way, information overlap is reduced with respect to the classic semantic expansion techniques. Experiments undertaken on MuchMore benchmark showed the effectiveness of the approach.

1. INTRODUCTION

This paper presents an ontology-based approach for a conceptual representation of documents. Such an approach is inspired by a recently proposed idea presented in [9], and uses an adapted version of that method to standardize the representation of documents and queries. The proposed approach is somehow similar 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 [12]. Several works demonstrated the enhanced performance of IR systems that implement query expansion approaches [19] [3] [5]. However, the query expansion approach has to be used carefully because, as demonstrated in [8], 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.

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. The smart idea behind the representation of documents by using concepts is that documents and queries may be represented in the same way. This way, the risk of omitting some related terms (as it may happen in the classical query expansion technique), is reduced. However, it is necessary to use a language resource that permits to cover a higher number of terms in order to avoid information loss.

This paper presents a new representation for documents and queries. The proposed approach exploits the structure of the well-known machine readable dictionary WordNet 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 by using concepts. This representation has been applied to the ad-hoc retrieval problem. The approach has been evaluated on the MuchMore[1] Collection [4] and the results demonstrate its viability.

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 results obtained from the test campaign 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 [11], while in [16] the author studies the application of ontologies to a large-scale IR system for web purposes. A model for the exploitation of ontology-based knowledge bases is presented in [7]. The aim of this model is to improve search over large document repositories. The model includes an ontology-based scheme for the annotation of documents, and a retrieval model based on an adaptation of the classic vector-space model [15]. Another information retrieval system based on ontologies is presented in [14]. The authors propose an information retrieval system which has landmark information database that has hierarchical structures and semantic meanings of the features and

Appears in the Proceedings of the 1st Italian Information Retrieval Workshop (IIR'10), January 27–28, 2010, Padova, Italy.
`http://ims.dei.unipd.it/websites/iir10/index.html`

[1] http://muchmore.dfki.de

characteristics of the landmarks.

The implementation of ontology models has been also investigated by using fuzzy models [6].

In IR, the user's input queries usually are not detailed enough, so the satisfactory query results can not be brought 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 [18] the authors propose and implement query expansion method which combines domain ontology with the frequency 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 precision can be gotten as user's query results.

In [10] 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.

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 by using the same set of concepts;
- to implement an approach that permits to index and to evaluate each concept, in both documents and queries, with the appropriate weight.

To represent documents, the method described in Section 4 has been used, combined with the use of the WordNet machine-readable dictionary. 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 by 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 ontological 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 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 payload to each term of the query. Therefore, to each element present in the concept-based representation of the query, its concept weight has been used as boost value.

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 reconstructed; 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 [9].

An ontology-based representation has been recently proposed in [9] 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 [9], which is a concept-based adaptation of TF-IDF.

In this paper, an adaptation of the approach proposed in [9] is presented. The original approach was proposed for domain specific ontologies and does not always consider all the possible concepts in the considered ontology, in the sense that it assumes a cut at a given specificity level. Instead, the proposed approach has been adapted for more general purpose ontologies and it takes into account all independent concepts contained in the considered ontology. This way, information associated to each concept is more precise and the problem of choosing the suitable level to apply the cut is overcome.

[2] See URL http://lucene.apache.org/.

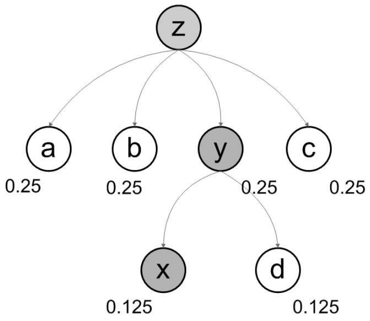

Figure 1: Ontology representation for concept 'z'.

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 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:

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

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. The value $N(c)$ is the weight associated with the concept c.

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

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

where $N_{\text{doc}}(b_i)$ is the number of explicit and implicit occurrences of b_i in the document, and $N_{rep}(b_i)$ 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 light ontology represented in Figures 1 and 2, and by considering a document D_1 containing concepts "$xxyyyz$".

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:

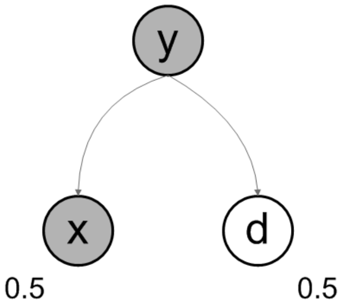

Figure 2: Ontology representation for concept 'y'.

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

so the document vector associated to D_1 is:

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

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

5. REPRESENTATION COMPARISON

In Section 4 the approach used to represent information was described. This section shows the improvements obtained by applying the proposed approach and it illustrates a comparison between the proposed approach and other two approaches commonly used in conceptual document representation. The expansion technique is generally used to enrich information content of queries. However, in the last years some authors applied the expansion technique also to represent documents [2]. Like in [13] [2], we propose an approach that uses WordNet to extract concepts from terms.

The two main improvements obtained by the application of the ontology-based approach are illustrated below.

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 document content by using concepts is generally lower than the size of the vector representing document content by using terms.

An example of technique that shows this drawback is presented in [13]. In this work the authors propose an indexing technique that takes into account WordNet synsets instead of terms. For each term in documents, the synsets associated to that terms are extracted and then used as token for the indexing task. This way, the computational time needed to perform a query is not increased, however, there is a significant overlap of information because different synsets might be semantically correlated. An example is given by the terms “animal” and “pet”, these terms have two different synsets, however, observing the WordNet lattice, the term “pet” is linked with an “is-a” relation with the term “animal”. Therefore, in a scenario in which a document contains both terms, the same conceptual information is repeated. This is clear because, even if the terms “animal” and “pet” are not represented by using the same synset, they are semantically correlated because “pet” is a sub-concept of “animal”. This way, when a document contains both terms, the presence of the term “animal” has to contribute to the importance of the concept “pet” instead of to be represented with a different token.

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 with respect to a vector space model representation. 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 document score is comparable to the computational time needed by using the vector space model.

6. EXPERIMENTS

In this section, the impact of the ontology document and query representation is evaluated. The evaluation method follows the TREC protocol [17]. For each query the first 1000 retrieved documents have been considered and the precision of the system has been calculated at different points: 5, 10, 15, and 30 documents retrieved. Moreover, the precision/recall graph has been calculated

The experimental campaign has been performed by using the MuchMore collection that consists of 7823 abstracts of medical papers and 25 queries with their relevance judgments. One of the particular features of this collection is that there are a lot of medical terms. This way, a term-based representation is more advantaged with respect to semantic representation, because specific terms present in documents (for example “Arthroscopic”) are very discriminant. Indeed, by using a semantic expansion some problems may occur because, generally, the MRD and thesaurus used to expand terms do not contain some domain-specific terms.

The precision/recall graph showed in Figure 3 illustrates the comparison between the proposed approach (gray curve with circle marks), the classical term-based representation (black curve), and the synset representation method [13] (light gray curve with square marks). As expected, for all recall values, the proposed approach obtained better results than the term-based representation. The best gain of the concept-based representation is at recall levels 0.0, 0.2, and 0.4. While for recall values between 0.6 and 1.0, the concept-based precision curve lies with the other two curves.

A possible explanation for this scenario is that for documents that are well related to a particular topic the adopted ontology representation is able to improve the representation of the documents contents. However, for documents that are partially related to a topic or that contains many ambiguous terms, the proposed approach is not able to maintain an high precision of the results. At the end of this section some improvements that may be responsible of this fact are discussed.

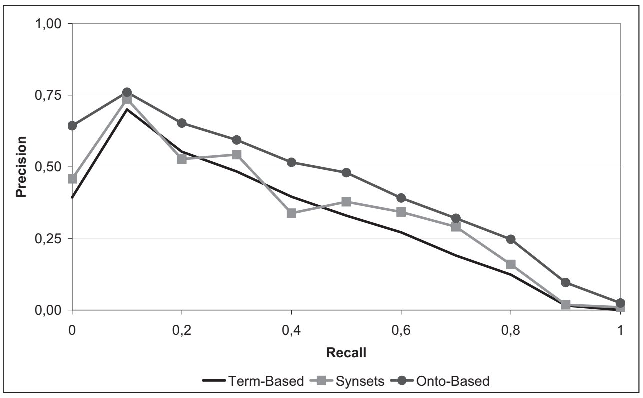

Figure 3: Precision/recall results.

In Table 1 the three different representations are compared for the Precision@X and MAP values. The results show that the proposed approach obtains better results for the all precision levels and also for the MAP value.

Systems	Precisions				
	P5	P10	P15	P30	MAP
Term-Based	0.544	0.480	0.405	0.273	0.449
Synset-Indexing [13]	0.648	0.484	0.403	0.309	0.459
Concept-Based	0.744	0.544	0.478	0.394	0.507

Table 1: Comparisons table between semantic expansion approaches.

An in-depth study of this first experiments campaign has been performed, and we have noticed that for some queries the concept-based representation obtained results that are below our expectations. By inspecting the implemented model, some issues have been noticed and are at now under analysis:

- 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 machine readable dictionary used to define the concept-based version of the documents. This way there is, in some cases, a loss of information that affects the final retrieval result.

- Proper names have not been considered: proper names of persons, geographical locations, industries, etc., are not present in the concept-based index. Observing the content of some documents and topics, proper names turn out to be a discriminant feature in some cases.

- Verbs and adjective are not present as well in the ontology: the concept representation of terms, described in Section 4, does not take into account verbs and adjectives.

This happens because verbs and adjectives are structured in a different way than nouns. The hyperonymy and hyponymy relations (that make MRD comparable with ontologies) are not defined for verbs and adjectives, therefore another approach will be studied and implemented to overcome this drawback.

- Term ambiguity: the concept-based representation has the problem of introducing an error given by not using a word sense disambiguation algorithm. Using such a method, concepts associated to incorrect senses would be discarded or weighted less. Therefore, the concept-based representation of each word would be finer, with the consequence of representing the information contained in a document with more precision.

Improving the actual model with the above features, would certainly yield significantly better results in the next experiments campaign. This positive view is motivated by the fact that, in spite of these issues, the preliminary goal of outperforming the precision of the term-based representation has been accomplished.

7. CONCLUSION

In this paper we have discussed an approach to index documents and to represent queries for information retrieval purposes which exploits a conceptual representation based on ontologies.

Experiments have been performed on the MuchMore 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. Investigation on results highlights that further improvement could be obtained by integrating WSD techniques like the one discussed in [1] to avoid the error introduced by considering incorrect word senses, and with a better usage and interpretation of WordNet to overcome the loss of information caused by the absence of proper nouns, verbs, and adjectives.

8. REFERENCES

[1] A. Azzini, M. Dragoni, C. da Costa Pereira, and A. Tettamanzi. Evolving neural networks for word sense disambiguation. In *Proc. of HIS '08, Barcelona, Spain, September 10-12*, pages 332–337, 2008.

[2] M. Baziz, M. Boughanem, G. Pasi, and H. Prade. An information retrieval driven by ontology: from query to document expansion. In D. Evans, S. Furui, and C. Soulé-Dupuy, editors, *RIAO*. CID, 2007.

[3] B. Billerbeck and J. Zobel. Techniques for efficient query expansion. In A. Apostolico and M. Melucci, editors, *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] D. Cai, C. van Rijsbergen, and J. Jose. Automatic query expansion based on divergence. In *CIKM*, pages 419–426. ACM, 2001.

[6] S. Calegari and E. Sanchez. A fuzzy ontology-approach to improve semantic information retrieval. In F. Bobillo, P. da Costa, C. d'Amato, N. Fanizzi, F. Fung, T. Lukasiewicz, T. Martin, M. Nickles, Y. Peng, M. Pool, P. Smrz, and P. Vojtás, editors, *URSW*, volume 327 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2007.

[7] 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.

[8] S. Cronen-Townsend, Y. Zhou, and W. Croft. A framework for selective query expansion. In D. Grossman, L. Gravano, C. Zhai, O. Herzog, and D. Evans, editors, *CIKM*, pages 236–237. ACM, 2004.

[9] 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.

[10] M. Díaz-Galiano, M. G. Cumbreras, 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.

[11] 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.

[12] E. Efthimiadis. Query expansion. In M. Williams, editor, *Annual review of information science and technology*, pages Vol. 31, pp. 121Ű187. Information Today Inc, Medford NJ, 1996.

[13] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarrán. Indexing with wordnet synsets can improve text retrieval. *CoRR*, cmp-lg/9808002, 1998.

[14] T. Hattori, K. Hiramatsu, T. Okadome, B. Parsia, and E. Sirin. Ichigen-san: An ontology-based information retrieval system. In X. Zhou, J. Li, H. Shen, M. Kitsuregawa, and Y. Zhang, editors, *APWeb*, volume 3841 of *Lecture Notes in Computer Science*, pages 1197–1200. Springer, 2006.

[15] G. Salton, A. Wong, and C. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[16] S. Tomassen. Research on ontology-driven information retrieval. In R. Meersman, Z. Tari, and P. Herrero, editors, *OTM Workshops (2)*, volume 4278 of *Lecture Notes in Computer Science*, pages 1460–1468. Springer, 2006.

[17] E. Voorhees and D. Harman. Overview of the sixth text retrieval conference (trec-6). In *TREC*, pages 1–24, 1997.

[18] F. Wu, G. Wu, and X. Fu. Design and implementation of ontology-based query expansion for information retrieval. In L. Xu, A. Tjoa, and S. Chaudhry, editors, *CONFENIS (1)*, volume 254 of *IFIP*, pages 293–298. Springer, 2007.

[19] J. Xu and W. Croft. Query expansion using local and global document analysis. In H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson, editors, *SIGIR*, pages 4–11. ACM, 1996.