A Conceptual Representation of Documents and Queries for Information Retrieval Systems by Using Light Ontologies

Mauro Dragoni, Célia da Costa Pereira, Andrea G. B. Tettamanzi

▶ **To cite this version:**

Mauro Dragoni, Célia da Costa Pereira, Andrea G. B. Tettamanzi. A Conceptual Representation of Documents and Queries for Information Retrieval Systems by Using Light Ontologies. Expert Systems with Applications, 2012, 39 (12), pp.10376–10388. 10.1016/j.eswa.2012.01.188 . hal-01328709

HAL Id: hal-01328709
https://hal.science/hal-01328709

Submitted on 8 Jun 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Conceptual Representation of Documents and Queries for Information Retrieval Systems by Using Light Ontologies

Mauro Dragoni[1], Célia da Costa Pereira[2], and Andrea G. B. Tettamanzi[3]

[1]) *Fondazione Bruno Kessler*
Via Sommarive 18 Povo, I-38123, Trento, Italy
dragoni@fbk.eu
[2]) *Université de Nice Sophia-Antipolis/CNRS*
UMR-6070, Laboratoire I3S, 06903, Sophia Antipolis, France
celia.pereira@unice.fr
[3]) *Università degli Studi di Milano, DTI*
Via Bramante 65, 26013 Crema (CR), Italy
andrea.tettamanzi@unimi.it

Keywords: Conceptual Representation, Lexica, Ontologies, Intelligent Information Retrieval

---

A Conceptual Representation of Documents and Queries for Information Retrieval Systems by Using Light Ontologies

Mauro Dragoni[1], Célia da Costa Pereira[2], and Andrea G. B. Tettamanzi[3]

[1]) *Fondazione Bruno Kessler*
Via Sommarive 18 Povo, I-38123, Trento, Italy
dragoni@fbk.eu
[2]) *Université de Nice Sophia-Antipolis/CNRS*
UMR-6070, Laboratoire I3S, 06903, Sophia Antipolis, France
celia.pereira@unice.fr
[3]) *Università degli Studi di Milano, DTI*
Via Bramante 65, 26013 Crema (CR), Italy
andrea.tettamanzi@unimi.it

Abstract

This article presents a vector space model approach to representing documents and queries, based on concepts instead of terms and using WordNet as a light ontology. Such representation reduces information overlap with respect to classic 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

The effectiveness of Information Retrieval Systems (IRSs) strongly depends on the way information contained in documents is represented. Commonly, documents are represented and indexed by using term-based representations; however, such representations have lost effectiveness in recent years because of the large amounts of data available on the web. Indeed, when we perform queries, an IRS simply searches documents that contain the query terms without considering, for instance, the semantic connections between them. These connections are given, for example, by the term synonymity or by the existence of different descriptions that are related to the same concept. Therefore, documents having

very different vocabularies could be similar in subject and, similarly, documents having similar vocabularies may be topically very different.

This paper presents an ontology-based approach to the conceptual representation of documents in order to solve the issues described above. Such an approach is inspired by a recently proposed idea presented in [16], and uses an adapted version of that method to standardize the representation of documents and queries. The proposed approach is somehow similar to query expansion technique [21]. However, additional considerations have been taken into account and some improvements have been applied as explained below.

Query expansion is an approach to boost the performance of Information Retrieval (IR) systems. It consists of expanding a query with the addition of terms that are semantically correlated with the original terms of the query. Several works demonstrated the improved performance of IR systems using query expansion [76, 8, 11]. However, query expansion has to be used carefully, because, as demonstrated in [14], 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 [6]. 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 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 (MRD) 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 Much-

More[1] Collection [9] 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 environments 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 evaluation of the approach are discussed; while in Section 7 we discuss about possible improvements of the presented approach. Finally, Section 8 concludes.

2. Related Works

This work is related to three different research directions that are being actively pursued in the field of information retrieval: the application of ontologies to IRSs, the adoption of expansion techniques applied to documents besides queries, and the indexing of document by using concepts instead of terms. In this section we start to present the general application of ontologies in IR. Then, we focus on the expansion task and on the conceptual indexing of documents that are the main objectives of the approach proposed in this paper.

2.1. Ontologies in Retrieval Systems

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 [19], while in [68] the author studies the application of ontologies to a large-scale IR system for web purposes. Model for the exploitation of ontology-base knowledge bases are presented in [13] and [69]. The aim of these models is to improve search over

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

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 [58]. Two other information retrieval systems based on ontologies are presented in [70] and [36]. The first describes a general architecture and a prototype application for the concise storage and presentation of the information retrieved from a wide spectrum of information sources, while the second proposes an information retrieval system which has landmark information database that has hierarchical structures and semantic meanings of the features and characteristics of the landmarks.

The implementation of ontology models has been also investigating using fuzzy models, two approaches having been presented in [77] and in [12].

Ontology-based semantic retrieval is very useful for specific-domain environments. A general IR system to facilitate specific domain search is illustrated in [42]. 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.

Other approaches implementing ontological representation for specific-domain semantic retrieval are presented in [78] and in [79] respectively for an E-Commerce information retrieval system, and a Supply Chain Management system. In both works the framework includes three parts: concepts, properties of concepts and values of properties, which can be linguistic values of fuzzy concepts. The semantic query is constructed by order relation, equivalence relation, inclusion relation, reversion relation and complement relation between fuzzy concepts defined in linguistic variable ontologies with Resource Description Framework (RDF). A system for legal and e-government information retrieval is presented in [28].

Logic-based approaches for query refinement in ontology-based information portals are presented in [64], [62] and [34]. The former two approaches are based on the model-theoretic interpretation of the refinement problem, so that the query refinement process can be considered as the process of inferring all

queries which are subsumed by a given query, while the latter implements a query expansion model to retrieve information based on knowledge base.

A natural language processing approach is presented in [46]. In this work the authors have developed ontology-based query processing to improve the performance of design information retrieval. In [18] the authors present an approach to expand queries that consists of searching an ontology for terms from the topic query 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 [63], a novel approach for determining relevance in ontology-based searching for information is presented.

2.2. Document and Query Expansion

In IR, the user's input queries usually are not detailed enough to allow fully satisfactory results to be returned. Query expansion can help to solve this problem. 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.

For example, in [75], the authors propose and implement a query expansion method which combines a domain ontology with the frequency of terms. Ontology is used to describe domain knowledge, while a logic reasoner and the frequency of terms are used to choose fitting expansion words. This way, higher recall and precision can be achieved. Another example of an ontology-like expansion approach is presented in [3]. In this case the authors exploit the link structure in Wikipedia to expand queries and they evaluate different retrieval models with the application of such an expansion method.

Recently, the document expansion direction has been also explored. The

first consideration about document expansion is that this task requires a higher computational effort due to the huge difference in size between documents and queries. Moreover, a document, in general, contains a larger number of terms than a query; therefore, if one wants to expand or to conceptually represent a document, one has to proceed cautiously, because the elements that are used to expand the document may negatively affect the final retrieval result.

In the literature, different kinds of approaches have been proposed. In [67], document expansion is applied to IR with statistical language models. The authors propose a document expansion technique to deal with the problem of insufficient sampling of documents, that is one of the main issues that affect the accuracy estimation of document models. The expansion of documents is made by clustering the repository, by computing a probabilistic neighborhood for each document, and then by using neighborhood information to expand the document.

Another well-known approach to expansion makes use of thesauri [7, 53, 49, 32]. In such approaches, concepts are extracted from one or more thesauri and queries and documents are expanded by using concepts that are connected with the terms contained in the queries. An alternative to a classic thesaurus usage is proposed in [74]. Here, the authors integrate the use of a thesaurus with the implementation of manually created metadata coming from a side collection, and with a query refinement approach based on pseudo-relevance feedbacks.

Expansion by pseudo-relevance feedback is also a well-established technique in cross-language information retrieval, and is used, for example, to enrich and disambiguate the typically terse queries entered by users. In [44], the author investigates about how the IRS effectiveness changes when document expansion techniques are applied before or after the translation of a document. The results obtained show that a post-translation expansion leads to a highly significant improvement.

Document expansion has also been approached with the use of fuzzy logic. In [45], the authors have developed an approach that uses fuzzy-rough hybridization for concept-based document expansion in order to enhance the quality of

text information retrieval. The considered scenario is given by a set of text documents represented by an incomplete information system.

Finally, document expansion has been used with success in the document summarization task [22]. In [73], the authors present an approach that uses document expansion techniques in order to provide more knowledge for the single document summarization task.

2.3. Conceptual Representation and Indexing

In traditional IR systems, documents are indexed by single words. This model, however, presents some limits due to the ambiguity and the limited expressiveness of single words. As a consequence, when traditional search models, like the Vector Space Model (VSM) [56], are applied to repositories containing millions of documents, the task of measuring the similarity between documents and queries leads to unsatisfactory results. One way of improving the quality of similarity search is Latent Semantic Indexing (LSI) [20], which maps the documents from the original set of words to a concept space. Unfortunately, LSI maps the data into a domain in which it is not possible to provide effective indexing techniques. Instead, conceptual indexing permits to describe documents by using elements (i.e. concepts) that are unique and abstract human understandable notions independent from any direct material support, any language, any information representation, and that are used to organize knowledge. Several approaches, based on different techniques, have been proposed for conceptual indexing.

One of the well-known mechanism for knowledge representation is Conceptual Graph (CG). In [38] we may find the implementation of two ontologies based on CGs: the Tendered Structure and the Abstract Domain Ontology. Moreover, in that paper, the authors first survey the indexing and retrieving techniques in CG literatures, and, then, they build a slight modification of these techniques to build their indexing techniques by using these ontologies. A fuzzy alternative to CG is presented in [48]. In that work, the authors present a model for text IR that indexes documents with Fuzzy Conceptual Graphs (FCG). The proposed

approach uses natural language processing to model document content, and it automatically builds a complete and relevant conceptual structure with the use of incomplete FCG.

Ontologies have been also applied in [10], in which the authors discuss an approach where conceptual summaries are provided through a conceptualization as given by an ontology. The idea is to restrict a background ontology to the set of concepts that appears in the text to be summarized and thereby provide a structure that is specific to the domain of the text and can be used to condense to a summary not only quantitatively but also conceptually covers the subject of the text. Two other approaches are presented in [2] and [43]. In the former, the authors describe an approach to indexing texts by their conceptual content by using ontologies along with lexical-syntactic information and semantic role assignment provided by lexical resources. In the latter, the authors describe a conceptual indexing method by using the UMLS Metathesaurus. In the proposed approach the concepts are automatically mapped from text and their semantic links, given by UMLS, are used to build a Bayesian Network the is subsequently used for the retrieval process. An alternative approach to conceptual indexing of documents based on word-chains is presented in [1].

Conceptual indexing has been also performed by applying clustering techniques. In [17], the authors present an indexing method which is based on partitioning the data space. They introduce the binary counterpart of the notions of minimum volume and minimum overlap, and combine them in a global hierarchical clustering criterion. They also show how the index structure induced by the clusterization may be exploited to deal with incompleteness and imprecision expressed in terms of answer precision and recall. An alternative document clustering approach has been presented in [37], in which the author introduces a method for building a hierarchical system of concepts to represent documents.

We have introduced above the need for using language resources in order to extract the set of concepts used to represent both documents and queries. One of the most well-known and popular language resource used in IR is WordNet (see

Section 3.3). WordNet has been adopted not only for conceptual indexing, but also for improving the quality of the conceptual representation of documents by performing disambiguation operations. For instance, in [50], the authors propose a novel word sense disambiguation approach that it is applied to the set of input documents and the senses of the words are accurately determined using the senses present in the WordNet along with the contextual information present in the document. Once the senses are determined, the documents are indexed conceptually. WordNet has also been used in [5]. Here, the authors propose an approach that aims at representing the content of the document by the best semantic network called "document semantic core" in two main steps. During the first step, concepts (words and phrases) are extracted from a document, while in the second step a global disambiguation of the extracted concepts regarding the document leads to building the best semantic network.

3. Components

In this Section we provide a description of the components we used to study and implement our approach. In Section 3.1, we introduce the discourse about Ontologies, in Section 3.2 we present the use of thesauri in IR, while in Section 3.3 we present WordNet, that is the machine-readable dictionary used in this work to represent documents by using concepts.

3.1. Ontologies

A (formal) *Ontology* defines the concepts and relations used to describe, represent, and reason about an area of knowledge. Ontologies are used by people, databases, and applications that need to share domain information. A domain is just a specific subject area or area of knowledge, like medicine, tool manufacturing, real estate, automobile repair, financial management, etc. Ontologies include computer-usable definitions of basic concepts in the domain and the relationships among them. They encode knowledge in a domain and also knowledge that spans domains. This way, they make that knowledge reusable.

The word ontology has also been used to describe artifacts with different degrees of structure. These range from simple taxonomies (such as the Yahoo hierarchy), to metadata schemes (such as the Dublin Core), to logical theories.

The term "ontology" has its origin in philosophy as:

"the branch of philosophy which deals with the nature and the organization of reality" [35].

The term "ontology" has been recently adopted in several fields of computer science and information science. There have been many attempts to define what constitutes an ontology and, perhaps, the best known (in computer science) is due to Gruber [33]:

"an ontology is an explicit specification of a conceptualization".

In this definition, a *conceptualization* means an abstract model of some aspect of the world, taking the form of a definition of the properties of important concepts and relationship. An emphexplicit specification means that the model should be specified in some unambiguous language, making it amenable to processing by machines as well as by humans.

Ontologies are becoming of increasing importance in fields such as knowledge management, information integration, cooperative information systems, information retrieval, and electronic commerce.

Ontologies may be classified according to their usage.

A **Domain Ontology** is an ontology that models a specific domain. It represents the particular meanings of terms as they are applied to that domain (ex. biology, computer science, mechanics, etc.). For instance the word "card", that has different meanings, can be used in an ontology about the domain of poker to model a playing card, and used in the domain of computer hardware to model a punch card or a video card.

An **Upper Ontology**, instead, or "foundation ontology", is a model of the common objects that are generally applicable across a wide range of domain ontologies because it contains a core whose terms can be used to describe a set

of domains. An example of an Upper Ontology is DOLCE [26].

Ontologies figure prominently in the emerging Semantic Web as a way of representing the semantics of documents and enabling the semantics to be used by web applications and intelligent agents. Ontologies can prove very useful for a community as a way of structuring and defining the meaning of the metadata terms that are being collected and standardized. Using ontologies, tomorrow's applications can be "intelligent," in the sense that they can more accurately work at the human conceptual level.

The ontology role is to make semantics explicit, for instance, to constitute a community reference, to share consistent understanding of what information means, to make knowledge reuse and sharing possible, and to increase interoperability between systems. In particular, the application area which has recently seen an explosion of interests is the Semantic Web, where ontologies are poised to play a key role in establishing a common terminology between agents, thus ensuring that different agents have a shared understanding of terms using in semantic markup. The effective use of ontologies requires not only a well-designed and well-defined ontology language, but also support from reasoning tools.

In Figure 1, an example of an ontology related to the "African wildlife" is reported [61].

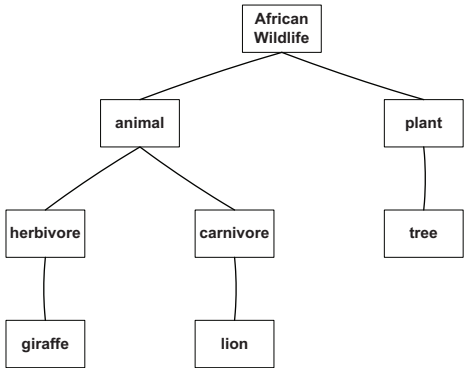

Figure 1: An example of an ontology.

A whole new field of knowledge engineering for building ontologies is flourishing. This paper does not cover the details of how an ontology is built or how

a suitable language for representing it is chosen, but will limit to selecting existing ontologies that may suit the needs to implement the proposed approaches. Details on the ontologies building process are exhaustively discussed in [61].

3.2. Thesaurus

When the "ontology" term is used in the Artificial Intelligence and Knowledge Representation communities, it refers to an artifact that uses a rich and formal logic-based language for specifying meaning of the entities and the relations between them. In some fields, like Information Retrieval, the use of ontologies may help to enrich information contained in documents and queries in order to improve the effectiveness of the Information Retrieval Systems. However, there does not exist an ontology that covers all possible knowledge with the formalism introduced above. For this reason, simplified tools derived from ontologies have been increasingly used; one of these tools is a *thesaurus*. A thesaurus contains terms and explicit basic relations between terms. These relations are not defined by using a formal logic-based language. Indeed, they explain a connection between terms, but the grammar containing formal constraints on how these terms may be used together is not defined. Generally, there are three kinds of relations:

- a *Hierarchical* relation describes the generalization/specialization relation between terms.
- an *Equivalence* relation describes the synonymity relation between terms.
- an *Associative* relation is used to link two related terms that are connected by a relation that is neither hierarchical or equivalence.

These relations permit to identify which terms are semantically related; therefore, they may be exploited to improve the precision of information contained in documents and queries. In Section 3.3, a more detailed description and a practical implementation of these relations are provided.

3.3. Machine Readable Dictionary

A machine-readable dictionary (MRD) is a dictionary in an electronic form that can be loaded into a database and can be queried via application software. It may be a single language explanatory dictionary or a multi-language dictionary to support translations between two or more languages or a combination of both.

For each word of the dictionary, a set of senses is associated to it. Word senses may be considered as fuzzy objects with indistinct boundaries. For instance, whether or not a person may be called "slim", is, to some degree, a subjective judgment by the user of the word. Detailed explanations about fuzziness, subjectivity and other critics to the concept of "sense" can be found in [41, 54, 39].

Regardless of exactly how one conceives of word senses, in a MRD, lexical relationships between word senses are the elements that characterize the power of a MRD [15, 23, 31]. The main kinds of lexical relations are identity of meaning, inclusion of meaning, part-whole relationships and opposite meanings.

Identity of meaning is synonymy; two or more words are synonyms if one may substitute for another in a text without changing the meaning of the text. More usually, "synonyms" are actually merely near-synonyms [30].

The primary **inclusion** relations are "hyponymy" and "hyperonymy" [15, 31]. Hyponymy is a relation such as "is-a" and "a-kind-of", while hyperonymy is a subsumption relationship. The inclusion relationship between verbs is known as troponymy, emphasizing the point that verb inclusion tends to be a matter of "manner" [31, 24]. Inclusion relationships are transitive, and thus form a semantic hierarchy among word senses. Words without hyponyms are leaves, while words without hypernyms are roots.

The **part-whole** relationships meronymy and holonymy also form hierarchies. Although they may be glossed roughly as "has-part" and "part-of", detailed analysis of part-whole relationships may be found in [15, 23, 31].

Words that are opposites, generally speaking, share most elements of their meaning, excepting for being positioned at the two extremes of one particular dimension. This kind of relationship is called **antonymy**.

The parallel between a machine readable dictionary containing all, or part of, the classical relations between words and ontologies is obvious. It even suggests that perhaps a MRD, together with the lexical relations defined on it, *is* an ontology. In this view, word senses are identified with ontological categories and lexical relations with ontological relations. The motivation for this identification should be clear from the discussion in Section 3.1. Nonetheless, a MRD, especially one that is not specific to a technical domain, is not a very good ontology.

An ontology is a set of categories of objects or ideas in the world, along with certain relationships among them; it is not a linguistic object. A lexicon, on the other hand, depends on a natural language and the word senses in it. The following example may be clarify the slight difference between ontologies and lexica. In an ontology, if the category "domesticated-mammal" subsumes the categories "dog" and "cat", then "dog" $\cap$ "cat" is empty because nothing is both a "dog" and a "cat". In a lexicon, the subsumption relation is described by the hyperonymy/hyponymy relation. Two words with a common hypernym will often overlap in sense, these words are named **near-synonyms**. Consider the English words *error* and *mistake*, and other words that denote some kinds of mistakes and errors (from WordNet): blunder, slip, faux pas, and howler. It is evident that a strict hierarchy is not possible, because a precise separation of the word senses cannot be given.

However, in technical domains, in which vocabulary and ontology are more closely tied than in a generic domain, it is possible, to some extent, to consider the hierarchical representation of the vocabulary as an ontology.

WordNet. A well-known example of a MRD is WordNet [24]. WordNet is one of the most important MRDs available to researchers in the field of text analysis, computational linguistics, and many related areas. WordNet is an electronic lexical database designed by use of psycholinguistic and computational theories of human lexical memory. It provides a list of word senses for each word, organized into synonym sets (Synsets), each representing one constitutional lex-

icalized concept. Every element of a Synset is uniquely identified by its Synset identifier (SynsetID). A synset is unambiguous and carries exactly one meaning. Furthermore, different relations link synsets to other semantically related synsets (e.g., hyperonyms, hyponyms, etc.). All related terms are also represented as synset entries. Furthermore, WordNet contains descriptions of nouns, verbs, adjectives, and adverbs.

Although WordNet was originally developed for the English language, currently versions for other languages as well as multilingual expansions, like MultiWordNet and EuroWordNet, are available.

In Figure 2 the relationships graph related to the word "memory" is presented.

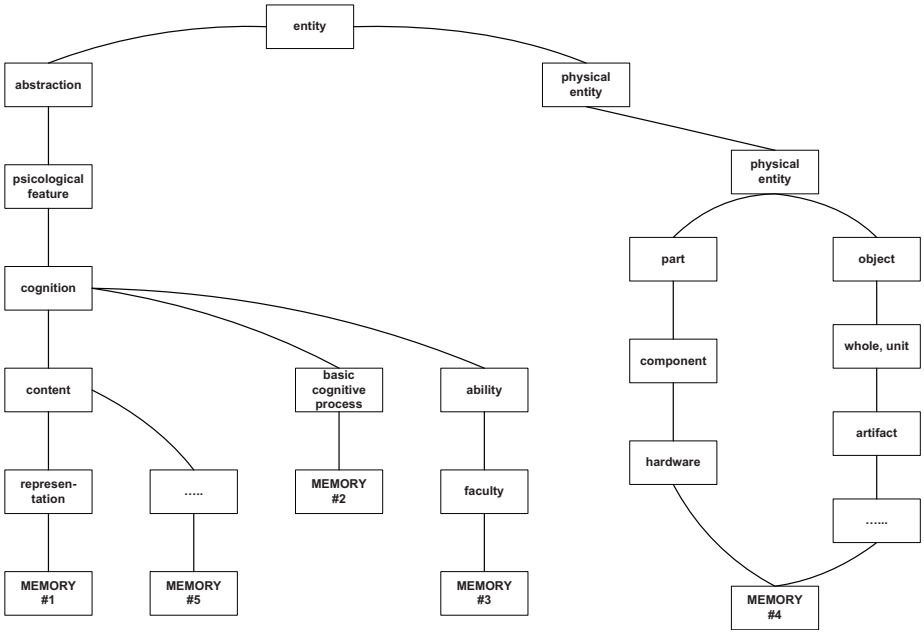

Figure 2: The tree related to the word "memory" in Wordnet.

4. Document Representation

The roadmap to prove the viability of a concept-based representation of documents and queries is composed of two main tasks:

- to choose a method that allows representing all document terms by using the same set of concepts;
- to implement an approach that allows indexing and evaluating each concept, in both documents and queries, with an "appropriate" weight.

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, including:

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 might become equal to the size of the vocabulary used in the repository;
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 recently proposed [16], which exploits the hierarchical *is-a* relation among concepts, i.e., the meanings of words. This method has been combined with the use of the WordNet MRD. From the WordNet database, the set of terms that do not have hyponyms has been extracted. We call such terms "base concepts". 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.

In this paper, an adaptation of the approach proposed in [16] is presented. The approach presented in [16] 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. Moreover, in [16] it is assumed that all concepts are contained in the ontology used to represent information. As said in the previous section, in this paper each document is represented by exploiting the WordNet light-ontology. By applying the approach presented in [16] to the *is-a* relation of WordNet, only nouns may be represented. Therefore, verbs, adverbs, adjectives, and proper-names would not be covered. A presentation of an in-depth description of the general approach follows, while in Section 4.1 a description of how the general approach has been extended to overcome the issue explained above is presented.

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 of representing documents in conventional IR 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 on keywords, but on concepts. This is the reason why we have adopted the document representation based on concepts proposed in [16], 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 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:

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

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

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

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

where $N_{\mathrm{doc}}(b_i)$ 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 3 and 4, and by considering a document D_1 containing concepts "$xxyyyz$".

In this case, the ontology base is:

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

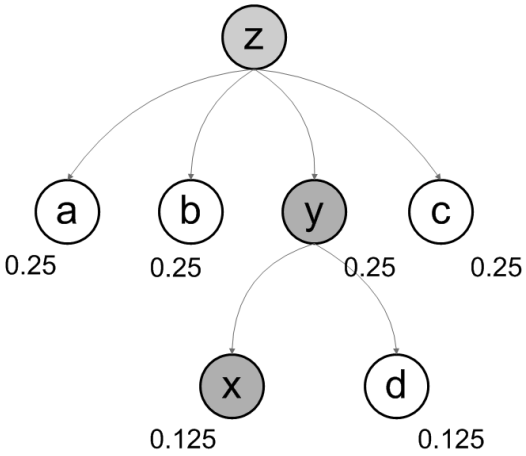

Figure 3: Ontology representation for concept 'z'.

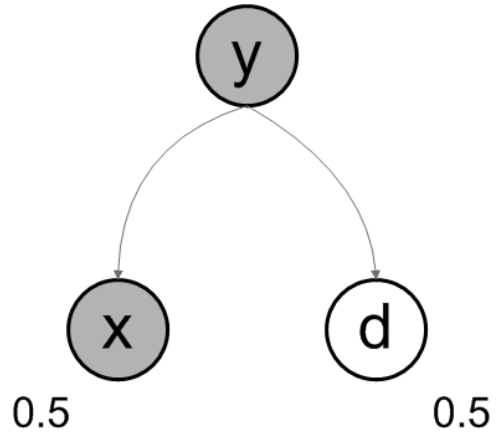

Figure 4: Ontology representation for concept 'y'.

and, for each concept in the ontology, the information vectors are

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

which yield the following document vector representation for D_1:

$$\begin{aligned}
\vec{D_1} &= 2 \cdot \text{info}(x) + 3 \cdot \text{info}(y) + \text{info}(z) \\
&= (0.25, 0.25, 0.25, 1.625, 3.625).
\end{aligned}$$

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 into 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-off value have been maintained, while the others have been discarded. The cut-off value used in these experiments is 0.01. This choice has an advantage and a drawback: the advantage is that the size of the entire index is limited due to the elimination of the less significant concepts, while the drawback is that the discarding of some minor concepts introduces an approximation of representing information. However, we have experimentally verified that this approximation does not affect the final results. This issue will be discussed in Section 6.

During the evaluation activity, queries have also been converted into the ontological representation. This way, weights are assigned to each concept in order to evaluate all concepts with the right proportion. For each element in

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

Collection	Number of Documents	Term-Based	
		Vect. Size (# of tokens)	Index Dim.
MuchMore	7823	47623	$\sim$ 3Mbyte
TREC Ad-Hoc	528155	650160	$\sim$ 2Gbyte

Collection	Number of Documents	Concept-Based	
		Vect. Size (# of tokens)	Index Dim.
MuchMore	7823	57708	$\sim$ 5Mbyte
TREC Ad-Hoc	528155	57708	$\sim$ 3.2Gbyte

Collection	Number of Documents	Difference	
		Vect. Size	Index Dim.
MuchMore	7823	+ 21.18 %	+ 66.67 %
TREC Ad-Hoc	528155	- 91.12 %	+ 60.00 %

Table 1: Comparison between the size of the term-based representation vector and the concept-based representation vector.

the concept-based representation of the query, the relevant concept weight has been used as boost value.

One of the features of Lucene is the possibility of assigning a payload to each element used both for indexing and for searching. Therefore, we exploited this feature in order to associate with each indexed concept its weight and to associate with each concept used to perform queries its boost value.

By considering the two collections used in the experiments described in Section 6, a comparison of the vector and the index size by using the classic term-based representation and by using the proposed concept-based representation is provided in Table 1. The size of the term-based representation vector is computed after the removal of the stop-words and after the application of the stemming algorithm.

With regards to the size of the vector, it is possible to notice that in the proposed approach the size of the concept-based vector remains the same for both collections. The same does not hold for the term-based vectors, this being correlated to the collection size. In fact, the TREC collection has a number of documents about ten times bigger than the number of documents contained in the MuchMore collection. The number of documents also influences the suitability of the proposed representation. By using the MuchMore collection, the vector size is 21.18% bigger than the vector size obtained by applying the term-based representation. However, the situation is dramatically different for the

TREC collection, in which the concept-based vector is 91.12% smaller than the term-based one. Therefore, the more the collection size increases, the more the proposed representation is suitable. The direct consequence is that by applying the proposed representation to large collections, the computational time needed to compare the documents and the query vectors is dramatically reduced.

A different discourse has to be done when it comes to the index size. For both collections, the size of the indexes created by applying the proposed representation is at least 60% bigger than the size of the indexes created by applying the term-based representation. This fact is mainly due to two reasons:

- Term representation: as it is presented above, each term is represented as a linear combination of concepts, therefore, each term is generally represented by using more than one concept. This way, by using the concept-based representation, each document is represented by using a number of tokens higher than the number of tokens used by applying the term-based representation.

- Token descriptor: by using the proposed approach, each concept is represented with two elements, the concept name and the concept weight. Therefore, in the proposed approach each token is stored with an overhead given by the concept weight. This overhead is not present in the term-based representation.

However, for this work, the optimization of such a representation has not been taken into account. In fact, by concentrating efforts in that direction, the discussed drawbacks would be surely limited.

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 because a relation such “x is a kind of y” is not suitable for them.

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

Table 2: Examples of "derivationally related forms" relations.

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 [25, 66]. It is out of the scope of this paper the discussion about the fact that verbs cannot be fit into the formula "x is a kind of y": more details about it may be found in [47]. It is sufficient to remark here that in WordNet, for verbs, a similar hyperonymy/hyponymy relation is called "troponymy" [25]. This relation may be expressed by the formula "To V_1 is to V_2 in some particular manner", where the term "manner" means that the troponymy relation between verbs comprises different semantic dimensions. For example, the troponymy relation for the verbs *fight, battle, war, feud, duel* expresses the occasion for or the form of the fight, while for the verbs *examine, confess, preach* it expresses the intention or the motivation of these communication verbs. For the adjectives, the only semantic relation in WordNet is "antonymy", as subsumption relations between adjectives are not considered.

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 2.

A graphical example of such a relation is shown in Figure 5.

A similar approach has been followed for proper-names. These entities,

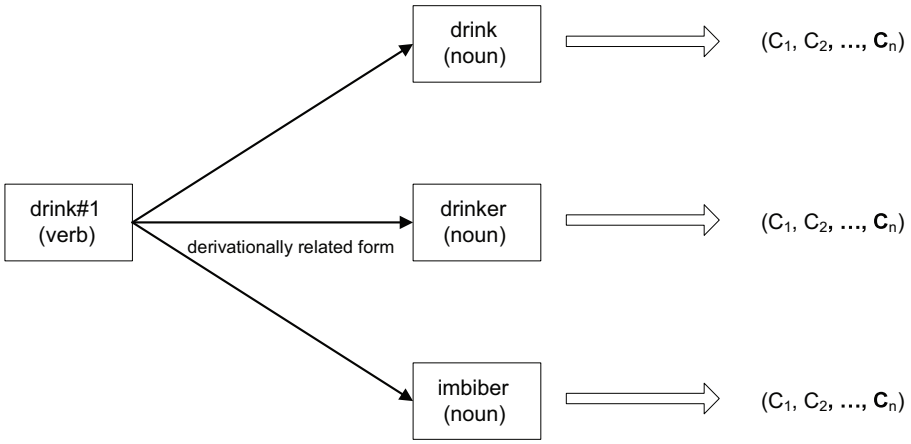

Figure 5: An example of a "derivationally related forms" relation.

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

Table 3: Examples of "instance of" relations.

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 3.

A graphical example of such a relation is shown in Figure 6.

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

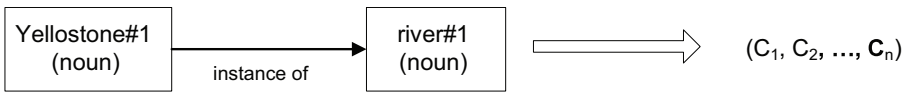

Figure 6: An example of a "instance of" relation.

proper names, besides nouns, verbs, and adjectives.

5. Representation Comparison

In Section 4, the approach used to represent information was described. This section discusses the improvements obtained by applying the proposed approach and compares the proposed approach to two other approaches commonly used in conceptual document representation. The expansion technique is generally used to enrich the information content of queries. However, in the past years some authors applied the expansion technique also to represent documents [6]. Like in [29, 6], 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 no 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 smaller than the size of the vector representing document content by using terms.

An example of a technique that shows this drawback is presented in [29]. 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 to 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, since "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 being 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 [6], 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 scores 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 into 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.

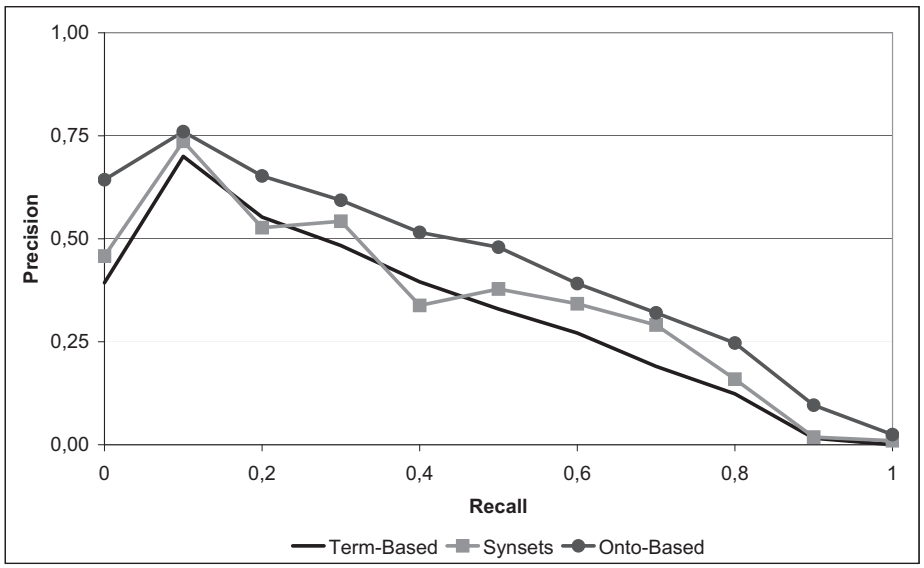

Figure 7: Precision/recall results.

The evaluation method follows the TREC protocol [72]. 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 [16] instead of the classic IDF.

The first part of the experimental evaluation has been performed by using the MuchMore collection, that consists of 7,823 abstracts of medical papers and 25 queries with their relevance judgments. One of the particular features of this collection is that there are numerous medical terms. This gives an advantage to term-based representations over the semantic representation, because specific terms present in documents (e.g., "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 all of the domain-specific terms.

Systems	Precisions				
	P5	P10	P15	P30	MAP
Baseline	0.544	0.480	0.405	0.273	0.449
Synset Indexing proposed by [29]	0.648	0.484	0.403	0.309	0.459
Conceptual Indexing proposed by [6]	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 4: Comparisons table between semantic expansion approaches.

The precision/recall graph shown in Figure 7 illustrates the comparison between the proposed approach (gray curve with circle marks), the classical term-based representation (black curve), and the synset representation method [29] (light gray curve with square marks). As expected, for all recall values, the proposed approach obtained better results than the term-based and synset-based representations. The best gain over the synset-based representation is at recall levels 0.0, 0.2, and 0.4, while, for recall values between 0.6 and 1.0, the synset-based precision curve lies within the other two curves.

A possible explanation for this scenario is that, for documents that are well related to a particular topic, the adopted ontological representation is able to improve the representation of the documents contents. However, for documents that are partially related to a topic or that contain many ambiguous terms, the proposed approach becomes less capable of maintaining a high precision. At the end of this section, some improvements that may help overcome this issue are discussed.

In Table 4, the three different representations are compared with respect to the Precision@X and MAP values. The results show that the proposed approach obtains better results for all the precision levels and also for the MAP value.

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 index contains documents of the Financial Times Ltd. (1991, 1992, 1993, 1994), the Congressional Record of the 103rd Congress (1993), the Foreign Broadcast Information Service (1996) and the Los Angeles Times (1989, 1990).

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 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 [60], 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 doc-

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 5: Precision@X and Mean Average Precision results obtained on TREC7 Topics.

uments are represented using an ontology, also each topic has been converted into the corresponding ontological representation.

The precision/recall graph shown in Figure 8 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

	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 6: Precision@X and Mean Average Precision results obtained on TREC8 topics.

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.

A more in-depth analysis of the performances for the first 20 documents retrieved is presented in Figure 9. The precision of the concept-based representation consistently outperforms the precision of the term-based representation for each rank position. In particular, the gain is very high for the first 10 positions, while it decreases a bit for positions from 11 to 20.

In Table 5, 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 10 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,

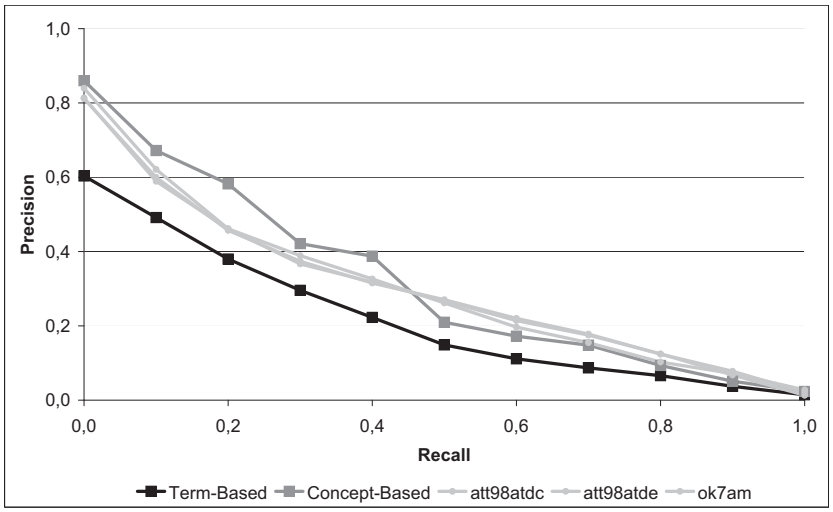

Figure 8: Precision/recall graph for TREC-7 topics.

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. Also considering TREC-8 topics, the concept-based representation overcomes the term-based representation in the performances related to the first 20 retrieved documents. This is shown in Figure 11.

The Precision@X and the MAP values shown in Table 6 confirm the impression described above.

In Table 7 we present the result of the significant test obtained by analyzing the performance of our approach. These results are obtained by comparing our accuracy with the best accuracy between the ones obtained by the other systems for each precision value. On the MuchMore Collection, the improvement obtained by the proposed approach are statistically significant, especially for the values of Prec@10, Prec@15, and Prec@30, for which the significance is above the 95%. On the TREC-7 Topics, when we improve the results of the compared systems (normal font), we obtained a significant performance at Prec@10, while the significance may be considered acceptable at Prec@5 and for the MAP value. Instead, when we do not improve the results of the compared systems (italic

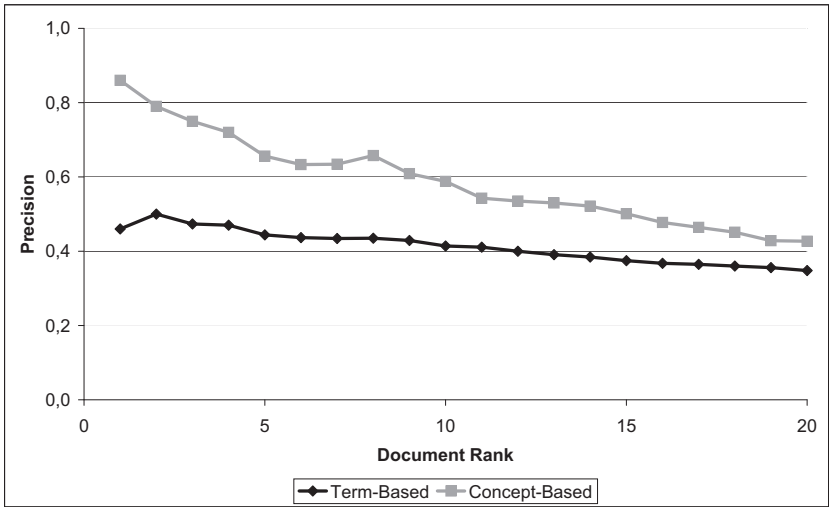

Figure 9: Precision@20 graph comparison between the proposed approach and the term-based representation for TREC-7 topics.

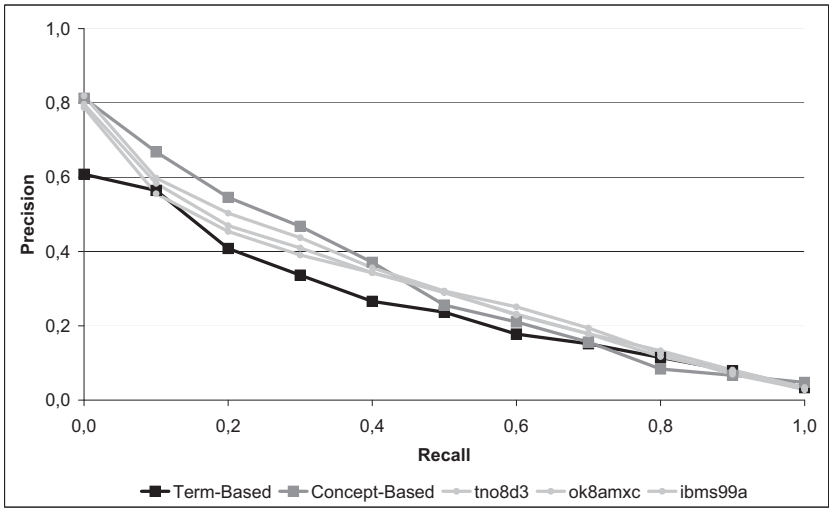

Figure 10: Precision/recall graph for TREC-8 topics.

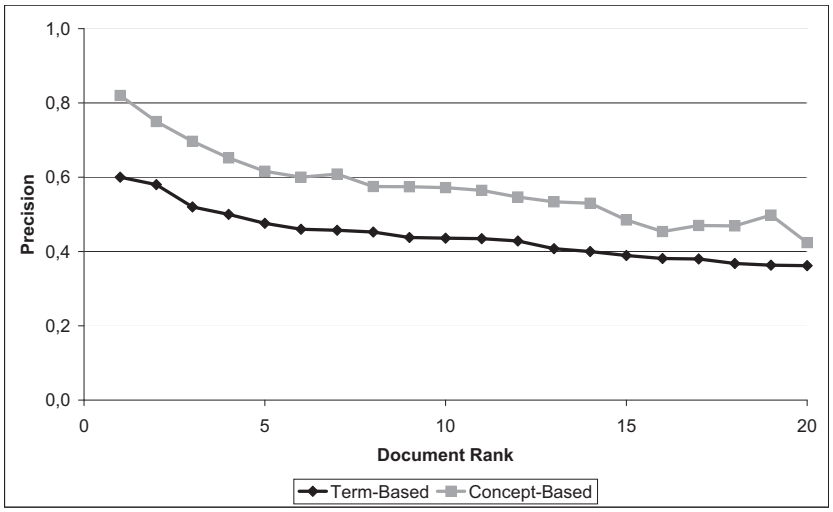

Figure 11: Precision@20 graph comparison between the proposed approach and the term-based representation for TREC-8 topics.

font), only at Prec@30 the result is statistically significant. A similar situation is present for the TREC-8 Topics. Here, the improvements obtained for Prec@5 and Prec@10 may be considered statistically significant, while for the values of Prec@30 and MAP, the significance of the results obtained by the proposed approach is below the 50%.

	P5	P10	P15
MuchMore Collection	70.72%	96.84%	99.06%
TREC-7 Topics	57.19%	94.39%	*29.56%*
TREC-8 Topics	92.79%	83.80%	*62.40%*

	P30	MAP
MuchMore Collection	99.99%	74.66%
TREC-7 Topics	*84.14%*	63.21%
TREC-8 Topics	*47.76%*	*10.81%*

Table 7: Statistical significant test of the results.

7. Future Work

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 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.

- Term ambiguity: the concept-based representation has the problem of introducing an error given by not using a word-sense disambiguation (WSD) 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 higher precision.

A more in-depth discussion about the use of a Word Sense Disambiguation (WSD) algorithm is needed because further advantages may be obtained by the use of such an algorithm for discarding uncorrected senses that are indexed by using the ontological representation introduced above. In [4], a WSD approach that uses Evolutionary Algorithms and Artificial Neural Networks is proposed. Most of the early work on the contribution of WSD to IR resulted in no performance improvement [55] [57] [40] [71]. On the contrary, encouraging evidence of the usefulness of WSD in IR has come from [59], [27], and [65]. A more detailed discussion about the impact of WSD in IR systems is presented in [51], in which the author asserts that an accurate disambiguation of the document base, together with a possible disambiguation of the query words, would allow it to eliminate from the result set of a query documents containing the same words

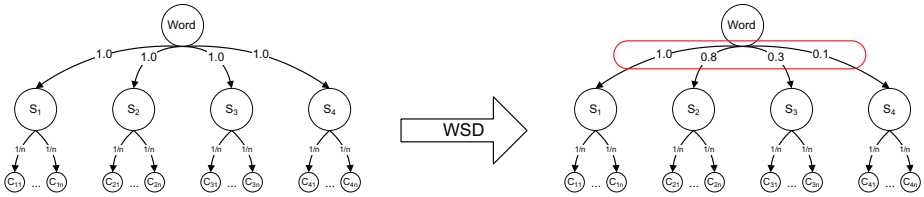

Figure 12: Example of how a WSD algorithm may be useful for a conceptual representation of documents.

used with different meanings (thus increasing precision) and to retrieve documents expressing the same meaning with different wordings (thus increasing recall).

Starting from this point of view, the thread of the approach presented in [4] is that WSD may improve IRS performances by using an effective WSD approach, in the sense that it makes possible (i) to increase the number of relevant document found; and to (ii) decrease the number of retrieved non-relevant documents. This is due to the fact that the combination WSD/IR considers documents containing only synonyms of the user query terms.

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.

A possible rationale behind this sentence may be explained with an example graphically represented in Figure 12. In the proposed approach, when we encounter ambiguous words, we consider all its senses in the same way. This way, an error is introduced in the document representation, that is given by the presence of the concepts associated with uncorrected senses. The goal of the application of a WSD algorithm is to learn which are the senses that are more correlated with the document content. Therefore, we may assign different weights to each sense in order to reduce the error that is introduced in the conceptual representation.

8. Conclusion

In this paper, we have discussed an approach to indexing documents and representing queries for IR 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. Investigation on results highlights that further improvement could be obtained by integrating WSD techniques like the one discussed in [4] 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.

References

[1] Aggarwal, C., Yu, P., 2001. On effective conceptual indexing and similarity search in text data. In: Cercone, N., Lin, T., Wu, X. (Eds.), ICDM. IEEE Computer Society, pp. 3–10.

[2] Andreasen, T., Bulskov, H., Jensen, P., Lassen, T., 2009. Conceptual indexing of text using ontologies and lexical resources. In: Andreasen, T., Yager, R., Bulskov, H., Christiansen, H., Larsen, H. (Eds.), FQAS. Vol. 5822 of Lecture Notes in Computer Science. Springer, pp. 323–332.

[3] Arguello, J., Elsas, J. L., Yoo, C., Callan, J., Carbonell, J. G., 2008. Document and query expansion models for blog distillation. In: Voorhees, E. M., Buckland, L. P. (Eds.), TREC. Vol. Special Publication 500-277. National Institute of Standards and Technology (NIST).

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

[5] Baziz, M., Boughanem, M., Aussenac-Gilles, N., 2005. Conceptual indexing based on document content representation. In: Crestani, F., Ruthven, I. (Eds.), CoLIS. Vol. 3507 of Lecture Notes in Computer Science. Springer, pp. 171–186.

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

[7] Bhogal, J., MacFarlane, A., Smith, P., 2007. A review of ontology based query expansion. Inf. Process. Manage. 43 (4), 866–886.

[8] Billerbeck, B., Zobel, J., 2004. Techniques for efficient query expansion. In: Apostolico, A., Melucci, M. (Eds.), SPIRE. Vol. 3246 of Lecture Notes in Computer Science. Springer, pp. 30–42.

[9] Boughanem, M., Dkaki, T., Mothe, J., Soulé-Dupuy, C., 1998. Mercure at trec7. In: TREC. pp. 355–360.

[10] Bulskov, H., Andreasen, T., 2009. On conceptual indexing for data summarization. In: Carvalho, J., Dubois, D., Kaymak, U., da Costa Sousa, J. (Eds.), IFSA/EUSFLAT Conf. pp. 1618–1624.

[11] Cai, D., van Rijsbergen, C., Jose, J., 2001. Automatic query expansion based on divergence. In: CIKM. ACM, pp. 419–426.

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

[13] Castells, P., Fernández, M., Vallet, D., 2007. An adaptation of the vector-space model for ontology-based information retrieval. IEEE Trans. Knowl. Data Eng. 19 (2), 261–272.

[14] Cronen-Townsend, S., Zhou, Y., Croft, W., 2004. A framework for selective query expansion. In: Grossman, D., Gravano, L., Zhai, C., Herzog, O., Evans, D. (Eds.), CIKM. ACM, pp. 236–237.

[15] Cruse, A., 1986. Lexical Semantics. Cambridge University Press.

[16] da Costa Pereira, C., Tettamanzi, A., 2006. Soft computing in ontologies and semantic Web. Studies in fuzziness and soft computing. Ed. Zongmin Ma, Springer, Berlin, Ch. An ontology-based method for user model acquisition, pp. 211–227.

[17] Diamantini, C., Panti, M., 1999. A conceptual indexing method for content-based retrieval. In: DEXA Workshop. pp. 193–197.

[18] Díaz-Galiano, M., Cumbreras, M. G., Martín-Valdivia, M., Ráez, A. M., Ureña-López, L., 2007. Integrating mesh ontology to improve medical information retrieval. In: [52], pp. 601–606.

[19] Dridi, O., 2008. Ontology-based information retrieval: Overview and new proposition. In: Pastor, O., Flory, A., Cavarero, J.-L. (Eds.), RCIS. IEEE, pp. 421–426.

[20] Dumais, S., Furnas, G., Landauer, T., Deerwester, S., R. Harshman, R., 1988. Using latent semantic analysis to improve access to textual information. In: CHI '88: Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, New York, NY, USA, pp. 281–285.

[21] Efthimiadis, E., 1996. Query expansion. In: Williams, M. (Ed.), Annual review of information science and technology. Information Today Inc, Medford NJ, pp. Vol. 31, pp. 121187.

[22] Endres-Niggemeyer, B., 1998. Summarizing Information. Springer.

[23] Evens, M., 1986. Relational Models of the Lexicon. Cambridge University Press.

[24] Fellbaum, C. (Ed.), 1998. WordNet: An Electonic Lexical Database. MIT Press.

[25] Fellbaum, C., Miller, G., 1990. Folks psychology or semantic entailment? a reply to rips and conrad. The Psychology Review 97, 565–570.

[26] Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., Schneider, L., 2002. Sweetening ontologies with dolce. In: Gómez-Pérez, A., Benjamins, V. (Eds.), EKAW. Vol. 2473 of Lecture Notes in Computer Science. Springer, pp. 166–181.

[27] Gao, L., Zhang, Y., Liu, T., Liu, G., 2006. Word sense language model for information retrieval. In: Ng, H., Leong, M.-K., Kan, M.-Y., Ji, D. (Eds.), AIRS. Vol. 4182 of Lecture Notes in Computer Science. Springer, pp. 158–171.

[28] Gómez-Pérez, A., Ortiz-Rodríguez, F., Villazón-Terrazas, B., 2006. Ontology-based legal information retrieval to improve the information access in e-government. In: Carr, L., Roure, D. D., Iyengar, A., Goble, C., Dahlin, M. (Eds.), WWW. ACM, pp. 1007–1008.

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

[30] Gove, P., 1973. Webster's New Dictionary of Synonyms. G. & C. Merriam Company, Springfield, MA.

[31] Green, R., Bean, C., Myaeng, S., 2002. The Semantics of Relationships: An Interdisciplinary Perspective. Cambridge University Press.

[32] Grootjen, F., van der Weide, T., 2006. Conceptual query expansion. Data Knowl. Eng. 56 (2), 174–193.

[33] Gruber, T., 1993. A translation approach to portable ontology specifications. Knowledge Acquisition 5 (2), 199–220.

[34] Guan, J., Zhang, X., Deng, J., Qu, Y., 2005. An ontology-driven information retrieval mechanism for semantic information portals. In: SKG. IEEE Computer Society, p. 63.

[35] Guarino, N., Giaretta, P., 1995. Ontologies and knowledge bases, towards a terminological classification. In: Mars, N. (Ed.), Towards Very Large Knowledge Database Building and Knowledge Sharing. IOS Press, Amsterdam, pp. 25–32.

[36] Hattori, T., Hiramatsu, K., Okadome, T., Parsia, B., Sirin, E., 2006. Ichigen-san: An ontology-based information retrieval system. In: Zhou, X., Li, J., Shen, H., Kitsuregawa, M., Zhang, Y. (Eds.), APWeb. Vol. 3841 of Lecture Notes in Computer Science. Springer, pp. 1197–1200.

[37] Holub, M., 2003. A new approach to conceptual document indexing: building a hierarchical system of concepts based on document clusters. In: ISICT. Vol. 49 of ACM International Conference Proceeding Series. Trinity College Dublin, pp. 310–315.

[38] Kayed, A., Colomb, R., 2002. Using ontologies to index conceptual structures for tendering automation. In: Zhou, X. (Ed.), Australasian Database Conference. Vol. 5 of CRPIT. Australian Computer Society.

[39] Kilgarriff, A., 1997. "i don't believe in word senses". Computers and the Humanities 31 (2), 91–113.

[40] Krovetz, R., Croft, W., 1992. Lexical ambiguity and information retrieval. ACM Trans. Inf. Syst. 10 (2), 115–141.

[41] Lakoff, G., 1987. Women, Fire, and Dangerous Things. University of Chicago Press, Chicago, IL.

[42] Lau, R., Lai, C., Li, Y., 2009. Mining fuzzy ontology for a web-based granular information retrieval system. In: Wen, P., Li, Y., Polkowski, L., Yao, Y., Tsumoto, S., Wang, G. (Eds.), RSKT. Vol. 5589 of Lecture Notes in Computer Science. Springer, pp. 239–246.

[43] Le, D., Chevallet, J.-P., Lim, J.-H., 2007. Using bayesian network for conceptual indexing: Application to medical document indexing with umls metathesaurus. In: [52], pp. 631–636.

[44] Levow, G.-A., 2003. Issues in pre- and post-translation document expansion: untranslatable cognates and missegmented words. In: Adachi, J. (Ed.), IRAL. ACL, pp. 77–83.

[45] Li, Y., Shiu, S. C.-K., Pal, S. K., Liu, J. N.-K., 2004. A fuzzy-rough method for concept-based document expansion. In: Tsumoto, S., Slowinski, R., Komorowski, H. J., Grzymala-Busse, J. W. (Eds.), Rough Sets and Current Trends in Computing. Vol. 3066 of Lecture Notes in Computer Science. Springer, pp. 699–707.

[46] Li, Z., Ramani, K., 2007. Ontology-based design information extraction and retrieval. AI EDAM 21 (2), 137–154.

[47] Lyons, J., 1977. Semantics. Cambridge University Press, New York.

[48] Maisonnasse, L., Chevallet, J.-P., Berrut, C., 2007. Incomplete and fuzzy conceptual graphs to automatically index medical reports. In: Kedad, Z., Lammari, N., Métais, E., Meziane, F., Rezgui, Y. (Eds.), NLDB. Vol. 4592 of Lecture Notes in Computer Science. Springer, pp. 240–251.

[49] Mandala, R., Tokunaga, T., Tanaka, H., 2000. Query expansion using heterogeneous thesauri. Inf. Process. Manage. 36 (3), 361–378.

[50] Manjula, D., Kulandaiyan, S., Sudarshan, S., Francis, A., Geetha, T., 2003. Semantics based information retrieval using conceptual indexing of documents. In: Liu, J., Cheung, Y.-M., Yin, H. (Eds.), IDEAL. Vol. 2690 of Lecture Notes in Computer Science. Springer, pp. 685–692.

[51] Navigli, R., 2009. Word sense disambiguation: A survey. ACM Comput. Surv. 41 (2).

[52] Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D., Peñas, A., Petras, V., Santos, D. (Eds.), 2008. Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers. Vol. 5152 of Lecture Notes in Computer Science. Springer.

[53] Qiu, Y., Frei, H.-P., 1993. Concept based query expansion. In: Korfhage, R., Rasmussen, E., Willett, P. (Eds.), SIGIR. ACM, pp. 160–169.

[54] Ruhl, C., 1989. On Monosemy: A study in linguistic semantics. State University of New York Press, Albany, NY.

[55] Salton, G., 1968. Automatic Information Organization and Retrieval. McGraw-Hill, New York, NY.

[56] Salton, G., 1975. Dynamic information and library processing. Prentice Hall.

[57] Salton, G., McGill, M., 1983. Introduction to Modern Information Retrieval. McGraw-Hill, New York, NY.

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

[59] Schütze, H., Pedersen, J., 1995. Information retrieval based on word senses. In: In Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval. pp. 161–175.

[60] Spink, A., Jansen, B., Blakely, C., Koshman, S., 2006. A study of results overlap and uniqueness among major web search engines. Inf. Process. Manage. 42 (5), 1379–1391.

[61] Staab, S., Studer, R. (Eds.), 2004. Handbook on Ontologies. International Handbooks on Information Systems. Springer.

[62] Stojanovic, N., 2004. An approach for the efficient retrieval in ontology-enhanced information portals. In: Karagiannis, D., Reimer, U. (Eds.),

PAKM. Vol. 3336 of Lecture Notes in Computer Science. Springer, pp. 414–424.

[63] Stojanovic, N., 2005. An approach for defining relevance in the ontology-based information retrieval. In: Skowron, A., Agrawal, R., Luck, M., Yamaguchi, T., Morizet-Mahoudeaux, P., Liu, J., Zhong, N. (Eds.), Web Intelligence. IEEE Computer Society, pp. 359–365.

[64] Stojanovic, N., Stojanovic, L., 2004. A logic-based approach for query refinement in ontology-based information retrieval s. In: ICTAI. IEEE Computer Society, pp. 450–457.

[65] Stokoe, C., Oakes, M., Tait, J., 2003. Word sense disambiguation in information retrieval revisited. In: SIGIR. ACM, pp. 159–166.

[66] Talmy, L., 1985. Lexicalization patters: Semantic structure in lexical forms. Language Typology and Syntactic Description 3, 57–149.

[67] Tao, T., Wang, X., Mei, Q., Zhai, C., 2006. Language model information retrieval with document expansion. In: Moore, R. C., Bilmes, J. A., Chu-Carroll, J., Sanderson, M. (Eds.), HLT-NAACL. The Association for Computational Linguistics.

[68] Tomassen, S., 2006. Research on ontology-driven information retrieval. In: Meersman, R., Tari, Z., Herrero, P. (Eds.), OTM Workshops (2). Vol. 4278 of Lecture Notes in Computer Science. Springer, pp. 1460–1468.

[69] Vallet, D., Fernández, M., Castells, P., 2005. An ontology-based information retrieval model. In: Gómez-Pérez, A., Euzenat, J. (Eds.), ESWC. Vol. 3532 of Lecture Notes in Computer Science. Springer, pp. 455–470.

[70] Varga, P., Mészáros, T., Dezsényi, C., Dobrowiecki, T., 2003. An ontology-based information retrieval system. In: Chung, P., Hinde, C. J., Ali, M. (Eds.), IEA/AIE. Vol. 2718 of Lecture Notes in Computer Science. Springer, pp. 359–368.

[71] Voorhees, E., 1993. Using wordnet to disambiguate word senses for text retrieval. In: Korfhage, R., Rasmussen, E., Willett, P. (Eds.), SIGIR. ACM, pp. 171–180.

[72] Voorhees, E., Harman, D., 1997. Overview of the sixth text retrieval conference (trec-6). In: TREC. pp. 1–24.

[73] Wan, X., Yang, J., 2007. Single document summarization with document expansion. In: AAAI. AAAI Press, pp. 931–936.

[74] Wang, J., Oard, D., 2005. Clef-2005 cl-sr at maryland: Document and query expansion using side collections and thesauri. In: Peters, C., Gey, F., Gonzalo, J., Müller, H., Jones, G., Kluck, M., Magnini, B., de Rijke, M. (Eds.), CLEF. Vol. 4022 of Lecture Notes in Computer Science. Springer, pp. 800–809.

[75] Wu, F., Wu, G., Fu, X., 2007. Design and implementation of ontology-based query expansion for information retrieval. In: Xu, L., Tjoa, A., Chaudhry, S. (Eds.), CONFENIS (1). Vol. 254 of IFIP. Springer, pp. 293–298.

[76] Xu, J., Croft, W., 1996. Query expansion using local and global document analysis. In: Frei, H.-P., Harman, D., Schäuble, P., Wilkinson, R. (Eds.), SIGIR. ACM, pp. 4–11.

[77] Zhai, J., Liang, Y., Jiang, J., Yu, Y., 2008. Fuzzy ontology models based on fuzzy linguistic variable for knowledge management and information retrieval. In: Shi, Z., Mercier-Laurent, E., Leake, D. (Eds.), Intelligent Information Processing. Vol. 288 of IFIP. Springer, pp. 58–67.

[78] Zhai, J., Liang, Y., Yu, Y., Jiang, J., 2008. Semantic information retrieval based on fuzzy ontology for electronic commerce. JSW 3 (9), 20–27.

[79] Zhai, J., Wang, Q., Lv, M., 2008. Application of fuzzy ontology framework to information retrieval for scm. In: Yu, F., Luo, Q. (Eds.), ISIP. IEEE Computer Society, pp. 173–177.