

How Ontology can be used to Improve Semantic Information Retrieval: The AnimSe Finder Tool

Abdelkrim Bouramoul
Computer Science
Department, Misc Laboratory
University of Constantine
B.P. 325, Constantine, Algeria

Med-Khireddine Kholadi
Computer Science
Department, Misc Laboratory
University of Constantine
B.P. 325, Constantine, Algeria

Bich-Lien Doan
Computer Science
Department, SUPELEC
3, Rue Joliot-Curie, 91192 Gif
Sur Yvette, France

ABSTRACT

In this paper, we propose a new approach for locating and retrieving documents; the search process is guided by the 'AnimOnto' domain ontology that we have constructed for this purpose. This ontology is used at two different stages: First, for the semantic indexing of documents, in this stage the representative concepts of each document are selected by a projection of the ontology on the document by attaching their terms to the 'AnimOnto' concepts. Then, during the semantics queries reformulation; in this stage we exploit the semantic links between concepts to expand the initial query. To validate these proposals, we have implemented the 'AnimSe Finder' tool (*Animal Semantic Finder*) which materializes the different phases of the proposed approach. The obtained scores show that the semantic indexing and the queries reformulation have generated a gain of 13.06 in terms of recall and 16.13 in terms of precision, which significantly reduces the documentary noise and silence.

General Terms

Semantic Web, Information Retrieval.

Keywords

Information retrieval, ontology, semantic indexing, semantic reformulation. AnimOnto, AnimSe Finder.

1. INTRODUCTION

One of the current challenges of Information Retrieval Systems (IRS) is to develop tools able to integrate more semantics in their treatment. The aim is twofold: 'understand' the contents of documents and 'understand' the user need to be able to linking them. It is possible, according to [1], to regroup the knowledge reflected in an IRS into three main classes:

- 1) Knowledge about users (profiles or user models).
- 2) Knowledge about the documents (index).
- 3) Knowledge about the concepts of the application field.

The first point has already been explored in our work [2]. When we proposed a system for taking into account the user context via their profile for the query reformulation. The obtained results using this technique have been too convincing in the sense that the documents returned by the contextual reformulation were more relevant than those obtained with the original query. On the other hand, the satisfaction of the users who participated in the experiment was remarkably better.

A perspective of our previous contribution was to use the other two aspects described by Boughanem, namely the knowledge about documents and the concepts of the application field as a reference to reformulate the user queries. In this paper, we concretize this idea by using an external resource (ontology) to improve the relevance of IRS, the same ontology will also be used to index and describe the content of documents. Our work falls within the area of semantics consideration via ontology in SRI, particularly that of using a domain ontology to index documents and reformulate the user queries by exploiting relations between concepts. This idea is not new, but the value of our contribution lies in its completeness. This work covers, in full autonomy, the search process from the ontology creation, the establishment of the system architecture and its implementation, and finally the validation of results.

This paper is organized into three main parts: The first presents the areas on which are based our contribution, namely the use of ontology in the Information Retrieval (IR) field and the different approaches for queries reformulation. In the second part we present our contribution, to this end we first determine the choices that we have adopted for the system parameters, we then present the architecture of the tool that we propose, a description of its implementation, an evaluation of its performance and the discussion of the obtained results. The third part presents conclusion and perspectives.

2. ONTOLOGY, A CRUCIAL NEED IN IR

Several definitions of ontology have emerged in the last twenty years, but the most referenced definition and also the most synthetic is probably that given by Gruber: "An ontology is an explicit specification of a conceptualization" [3]. Starting from this definition, ontology is used in the IR field to represent shared descriptions related to a more or less formal domain in order to add a semantic aspect to IRS.

It is therefore natural that works on the integration of ontology in the IRS are developed. A first solution aimed to build ontology from corpora on which RI tasks will be carried [4], [5]. A second solution involves to reuse existing resources, in this case, ontology are usually chosen only from the knowledge domain that they address [6], [7].

In general, the knowledge represented by ontology can be used at three different levels in the IR process. It can help to index documents, and then called semantic indexing. It can also assist the formulation of the user need and access to documents.

Finally, the ontology can be used in the model itself to achieve matching between the need and the granules documentaries.

In our case, we use domain ontology in the indexing process and queries reformulation to interrogate a documentary basis using the 'AnimSe Finder' tool that we propose. The idea is to pass the user's query through the conceptual network of the 'AnimOnto' ontology, developed for this purpose, to enrich it with new words coming from the vocabulary of this ontology. The interest is twofold:

- *Increase the recall*: by the query expansion taking into account the terms that are not present in its initial form. This can be done by adding synonyms recovered from the ontology (related concepts) to the query terms that are in relation with the chosen domain (animals)

- *Increase precision*: thanks to the semantic indexing of documents using concepts which are recovered from the ontology instead of ambiguous terms. This can be done by extracting terms guided by the chosen domain ontology, then a weighting of concepts using the relationships between them.

3. QUERY REFORMULATION, APPROACHES AND LIMITATIONS

The query reformulation is to modify the user's query by adding significant terms. The idea of query refinement is not new; several approaches use different techniques for select terms to be added to the initial query. We distinguish three types of query reformulation approaches and the deference between them lies on the one hand in the source of terms used in the reformulation. On the other hand, it lies in the method used for selecting terms to be added to the initial query.

The first type of the approaches is based on a global analysis of the considered collection of documents and the most commonly among them is based on statistical analysis of document corpus [8]. The objective is to increase the frequency of words appearing together in one document and select the terms with the highest coefficient. The information thus obtained is used to reformulate query automatically by adding terms related to the terms already used in the query. The terms added from the documents give a better adequacy between the need for information and the document collection.

The second type of approaches based on the principle of relevance feedback aims to reformulate the initial query to correspond better to the content of the documents collection. The principle is as follows; the user submits his initial query and the system returns an initial set of documents that the user has to judge (relevant, irrelevant). Knowing the relevance of initially returned documents is used for selecting terms to be added to the initial query. We quote in this category the work of [9] in which the system offers, based on the first query, a set of documents and according to those viewed by the user, the system updates its terms index in concordance with automatic learning methods.

The last type of the approaches, described in the literature, uses external resources of terms such as thesauri or ontology that contain the vocabulary used in the query enrichment, such approaches use ontology with equivalence and subsumption relations [10], in order to extract the terms to be added to the initial query.

4. SYSTEM PARAMETER

The parameters of the system that we propose concern the following elements:

4.1 The concepts source

To recover terms that will be added to the initial query, we use the 'AnimOnto' ontology as a source of concept. This is a domain ontology related to the world of animals created under Protégé. The latter can produce two code formats for describing an ontology (OWL and XML), for implementation reasons we chose the XML format. The XML describing the 'AnimOnto' ontology characteristics are subsequently used as input for the reformulation process.

4.2 The concepts selection method

The concepts are recovered from the ontology as follows: if the concept (C1) is present as a term in the initial query, it will be replaced by the concept (C2) recover from the 'AnimOnto' ontology. The choice of the concept (C2) is done after parsing of the XML file describing the ontology. Finally, the enlargement of concepts is based on the presence of a semantic link between (C1) and (C2) (synonym or broader concept).

4.3 The user role

We opted for an automatic query reformulation; the user's role is therefore passive. The latter does not intervene in the reformulation process and is the system that supports the entire operation. The idea is to increase the weight of words found in documents considered as relevant and conversely decrease the weights of terms considered as irrelevant.

5. CONTRIBUTION

5.1 Construction of the 'AnimOnto' ontology

Although no general methodology has so far been able to be imposed, many principles and criteria for building ontology have been proposed. These methods may cover the whole process guiding the construction steps. This is the case of the Kactus method (modelling Knowledge About Complex Technical systems for multiple USe) proposed by Schreiber [11]. We use this method to construct the domain ontology that will serve as a terms source used in the reformulation and indexing process.

Basing on the idea that our goal is not to build the ontology in itself, but rather its use to guide our approach and thus validate our proposal. The choice of this method is justified by two main reasons: first, by the fact that its use is not expensive, and unlike other methods, it does not require collaboration between domain experts, engineers and future ontology users. Secondly, it is adapted to the development of small ontology that are generally used during test and validation steps. The Kactus method is based on three main stages, namely:

- 1) Specification of the ontology basing on the chosen field, it consists in particular to determine the terms to be collected and tasks to be performed using this ontology.
- 2) Organization of terms using the meta-categories such as: concepts, terms, relationships, attributes, etc.
- 3) Refinement of the ontology by its structuration according to principles of modularization and hierarchical organization.

5.1.1. Step 1: Domain specification

The 'AminOnto' ontology will be constructed to provide a conceptual vocabulary, which allows the annotation of documents relating to the animal world. The choice of the animals' world is justified by the availability of data to construct the ontology on the one hand, and the possibility of building a scalable documentary base whose terms are semantically related to the concepts of ontology on the other hand. We note in this context that our system is open in the sense that it can be used for information search in another documentary base whose contents are related to another domain.

2.1.2. Step 2 and 3: From the terms collection to the ontology refinement

We can practically not separate the ontology construction steps, because it is a nonlinear process. Several roundtrips were made during the construction of the 'AminOnto' ontology for the following reasons:

- It was not possible to know from the outset, that the collected terms are sufficient to meet the purpose for which the ontology was constructed, we have added new terms when it was necessary, all the same we have removed terms that we judged unnecessary.
- It was not always easy to predict that a term will play the role of a class or an attribute; several changes were made in this direction.

Figure 01, presents a part of the 'AminOnto' ontology that we've created.

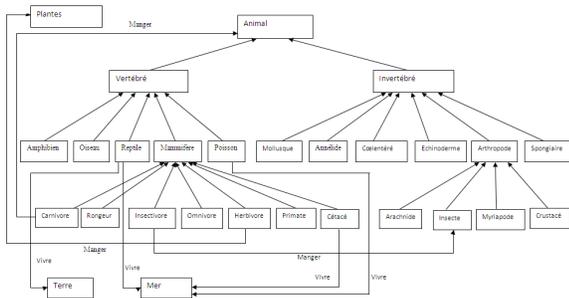


Fig 1: A part of the 'AminOnto' ontology

5.2 Architecture of the 'AnimSe Finder' tool

Animal Semantic Finder is a tool for semantic information searching in a documentary basis whose contents is related to the animal world.

The semantics is taken into account by the 'AnimOnto' domain ontology used to select terms that are used in the documents indexing and the queries reformulation. In addition to search guided by ontology, our system allows for a classical search using the query as that was formulated by the user. The aim is to compare the results given in both cases, and thus measures the contribution of the semantics consideration in the information retrieval process.

The different modules composing our system communicate by message sending. A message can be a user query, a query reformulated by the system, a concept of ontology or a document. In this context, and in order to ensure a better structuration of its module, the 'AnimSe Finder' tool is based on three complementary processes: indexing process, reformulation process and search process. We present in what follows each of

these processes, we give the different elements that compose them and the manner in which they operate.

5.2.1 Indexing process

The direct use of the documentary base during a search is a tedious operation, which takes a considerable time according to the document number and the volume of each document. For this reason a special treatment for identifying the relevant elements to be used by the search process is necessary, it is the indexing operation.

Two types of indexing have been considered: the semantic indexing guided by ontology, and the classical indexing which consists to construct a set of terms to characterize the content of a document. In this second type of indexing, weighting is done before creating the index by calculating the occurrence frequency of each term in the concerned document. A step common to both types of indexing consists to clean the document; this can be done by eliminating words with blank informational content using an anti-dictionary. These blank words appear in most documents and are not discriminating. They can be articles, prepositions, conjunctions, or even verbs.

The semantic indexing process is to construct a set of terms representing the informational content of each document; these keywords are derived from the ontology concepts. The indexing process is composed of two stages: The documents indexing then their weighting. Once the weighting is complete, the indexing module transmits the set of triple (concept, weight, document location) associated to each document to the storage module. The latter updated the indexes base for an eventual use in future search sessions. Figure 2, shows the operating principle of the indexing process and illustrates the interaction between its different elements.

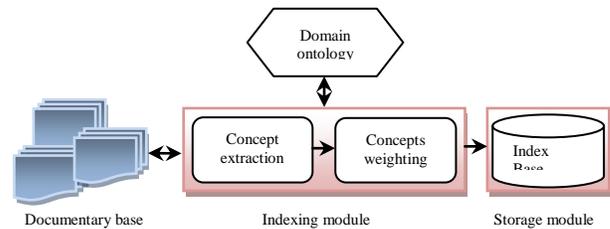


Fig 2: Architecture of the indexing process

The indexing module is the heart of this process. It retrieves the concepts from the ontology and calculates their frequency in the documents, it proceeds as follows:

A) *Concept extraction*: The aim of this step is to extract all the terms of the document that may represent concepts in the ontology. These terms correspond to different entries (or nodes) in the ontology. For this purpose, we use a technique that consists to project the ontology on the document. This is done by parsing the ontology using to identify the concepts that occur as terms in the document.

B) *Concepts Weighting*: The weight of a word reflects its importance in the document. The weighting phase has a major impact on the quality of the search process itself, it allows to associate with each concept found in the document a weight (appearance frequency). We used a statistical weighting method to calculate the frequency of each term extracted according the number of its occurrences.

5.2.2 Reformulation process

The user often finds difficulties to translate his exact information needs. Therefore, among the documents that are returned by the search tool, some of them less interested him than others. To this end, an improvement in how the user expresses his need is a supplement that can enhance the quality of the returned documents. In this context, the reformulation process supports the generation of a new query using the 'AnimOnto' ontology and the initial query in order to return documents which are more relevant than those provided by a non-reformulated query.

The structure of the reformulation process, the different modules that compose it and the interactions between them are illustrated in Figure 3.

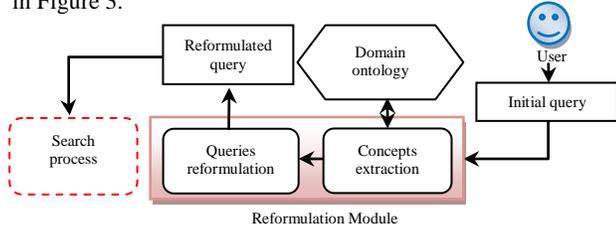


Fig 3: Architecture of the reformulation process

This process is based essentially on the reformulation module that supports the expansion of the user query to better reflect his information need. This module is conducted in two steps acting on different aspects of the initial query, these steps are:

A) *Concepts extraction*: In this step the reformulation module retrieves the query terms present in the ontology, then it goes through the ontology using these words as an entry point to extract the concepts directly related to each term in different hierarchies' levels of ontology.

B) *Query reformulation*: In this step, the reformulation module takes as input the concepts retrieved in the previous step; they use them in generating the new query that will be forwarded later to the research process.

5.2.3 Search process

This process is mainly responsible for the fundamental decision that allows associating to a query, all relevant documents to be returned. It is based on the boolean search model, in which a document (d) is represented by its set of terms (t_i), and a query (q) as a logical terms expression. A document does correspond to a query if the implication (d⇒q) is valid. This correspondence C(d, q) is determined as follows:

- C(d, t_i) = 1 if t_i ∈ d ; 0 otherwise
- C(d, q1 ∧ q2) = 1 if C(d, q1) = 1 et C(d, q2) = 1 ; 0 otherwise
- C(d, q1 ∨ q2) = 1 if C(d, q1) = 1 ou C(d, q2) = 1 ; 0 otherwise
- C(d, ¬q) = 1 if C(d, q) = 0 ; 0 otherwise

To clarify the operating principle of the search process, we present two search scenarios according to the two search modes provided by our system:

A) *Scenario 1 'a classical search'*: Suppose the user wishes to have information about the cats' world. He then formulates a query containing the word 'cat'. The search process takes as input this query and interrogates all the documents in the documentary basis through the indexes base and retrieves the relevant ones. These are documents {d₁, d₂, d₃, ..., d_n} in which will find the word 'cat'. The search module then ranks the documents from the highest weight to the lower {P₁, P₂, P₃, ..., P_n}, the result is finally displayed to the user.

B) *Scenario 2, 'a semantic search'*: For the same information need, the user query is sent this time to the reformulation process that automatically generates a new query using the 'AnimOnto' ontology. The new query will be transmitted to the search process to be used later in order to return the documents reflecting the user need. In the chosen example, the result of the query reformulating is: "carnivorous + cat", the word 'carnivore' was recovered from the ontology using a parser developed for this purpose. Finally, the search process retrieves the set of documents {d₁, d₂, d₃, ..., d_n} containing the word 'cat' and/or the word 'carnivore', it calculates thereafter the sum of the weights {P₁+P₁'₁, P₂+P₂'₂, P₃+P₃'₃, ..., P_n+P_n'_n} in order to classify the resulting documents according to their relevance. The result is then displayed to the user.

Figure 4 shows the manner in which the search process operates and the interaction between its different elements.

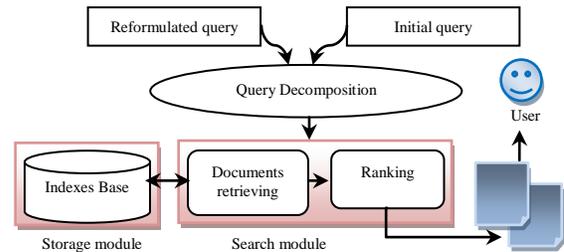


Fig 4: Search process architecture

Finally, the combination of the three processes will allow us to define the general architecture of our system; it is presented in Figure 5.

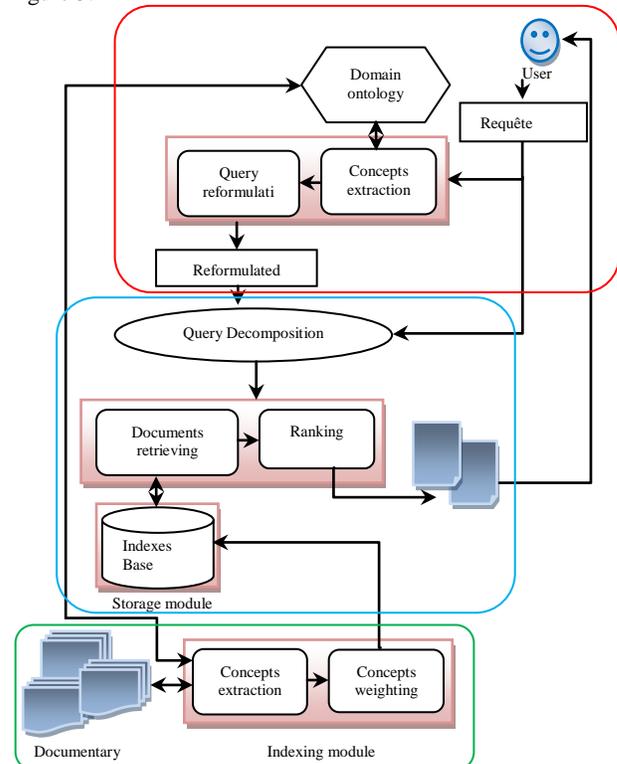


Fig 5: Architecture of the semantic information search system

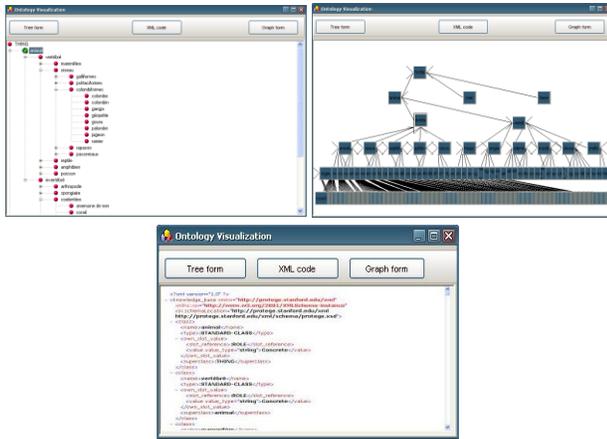
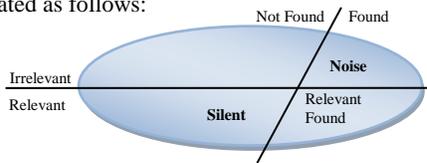


Fig 8: Visualization of the 'AnimOnto' ontology

6. EVALUATION OF THE 'ANIMSE FINDER' PERFORMANCE

To assess the 'AnimSe Finder' tool performance, we conducted an experiment that aims to understand and measure the contribution of the semantics consideration in the indexing and the reformulating process. The evaluation was made based on the metrics generally used in the information retrieval field. These metrics are precision, recall and F-measure, they are calculated as follows:



$$\text{Recall} = \frac{\text{Number of Relevant Found Documents}}{\text{Total Number of Relevant Documents}}$$

$$\text{Precision} = \frac{\text{Number of Relevant Found Documents}}{\text{Total Number of Found Documents}}$$

$$\text{F-Measure} = \frac{2 \times (\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}$$

6.1 Test collection characteristics and evaluation method

The evaluation protocol that we used is based on four elements: a corpus of documents (documentary basis), a corpus of queries, relevance judgments (referential) indicating that such document is relevant for such query, and evaluation metrics (recall, precision and F-measure).

A) The documentary basis: The used test collection has been created from the web and encyclopedias by choosing arbitrarily documents that are related to the animals' world. These documents are selected according to two criteria: their accessibility and diversity (writing styles, vocabulary, text length, etc.). The documentary database includes 13000 text documents. The advantage of this documents corpus lies in the fact that it consists of reliable resources.

B) The referential: referential of the correct answers are made up as follows: for each query, all potentially relevant documents are gathered in a 'relevance group'. We have, to this end,

controlled the relevance of each document according to the query during the construction of referential.

Table 1, summarizes the characteristics of the test collection used for the performance evaluation of the 'AnimSe Finder' tool.

Table 1. Characteristics of the test collection

Category	Number
Total document number	13000
Queries number	25
Average number of relevant documents per query	220

C) Operation principle: Using the documentary basis previously described, the system performs the semantic indexing of all documents. It takes as input a set of 25 queries posed by users in natural language. These users have the role of judging whether the documents found in both search (semantics and classical) are relevant or not according to the query. The documents returned in both cases search has been listed by their relevance order according to each query. They were finally subjected to evaluation using the previously described protocol.

6.2 Results and discussion

The obtained results are presented in Table 2.

Table 2. Evaluation results

Search type	Recall (%)	Precision (%)	F-Measure
Semantic	91.18	88.37	0.792
Classical	78.12	72.24	0.687

The obtained scores show that the recall rate is higher in the case of search guided by the ontology. The addition of the semantic dimension during indexing and reformulation produced a gain of 13.06 compared to the classical search. This rate is considered a 'silent' corresponding to the relevant documents that have not been returned in the case of a classic research. This gap in terms of recall is also explained by the strong match between the words that were added to the query and the information content of the returned documents.

With respect to the precision, scores show that the semantic search presents an improvement of 16.13 compared to classical search. This high rate of 88.37 means that little unnecessary documents are provided by 'AnimSe Finder' and that the latter may be considered as "precise". In contrast the loss of 16.13 in the case of classical search represents documents that are returned although it is unnecessary or irrelevant; they constitute the 'noise'.

Finally, these results are confirmed by the F-measure that combines precision and recall and their weighting.

7. CONCLUSION

In this paper we presented two complementary proposals reflecting two viewpoints of ontology using in information retrieval:

The first proposal, semantic indexing, is based on the hypothesis that a document can be viewed as a set of concepts, where the importance of a concept depend in the number of links with other concepts that share the same document. In this proposal, the document concepts are selected by a projection of the 'AnimOnto' ontology on the different documents.

The second proposal concerns the queries reformulation by exploiting the semantic relationships between their terms in order to improve the performance of the search tool.

To validate these proposals, we implemented the 'AnimSe Finder' tool. In this application the different phases of document representation has been implemented, namely the term extraction, the new concepts extraction and weighting. 'AnimSe Finder' materialized also the queries reformulation phase by expansion using the 'AnimOnto' ontology. This is done by taking into account the semantic relationships between concepts presented in the queries and those of the ontology.

Finally, the realized tool has the advantage of being generic and adaptable to other search types. It just needs to use a different ontology and another documentary basis corresponding to the desired domain in order to exploit the functionality offered by the 'AnimSe Finder' tool. The comparison of scores obtained by a classical search and those of a search guided by the 'AnimOnto' ontology, in the case of present documents collection, has shown that the use of terms generated by the ontology and the query reformulation provides a remarkable improvement in terms of the returned results relevance.

As perspective, it would be interesting to deepen the realized work with the idea of using a query combined of several words, and also to make a query lexical analysis to correct the user spelling errors.

It would also be interesting to widen the application domain by extending the 'AnimOnto' ontology by adding new concepts and semantic relationships, or to extend the search into other fields than that of animals.

8. REFERENCES

- [1] Boughanem, M. Berrut, C. Mothe, J. Dupuy, C. 2009. Advances in Information Retrieval, 31th European Conference on IR Research, ECIR 2009, Toulouse, France. Proceedings Springer
- [2] Bouramoul, A. Kholadi, M.K. Doan, B.L. 2010. PRESY : A Context based query reformulation tool for information retrieval on the Web, In JCS : Journal of Computer Science, Vol 6, Issue 4, pp. 470-477, 2010., ISSN 1549-3636, New York, USA.
- [3] Gruber, T.R. 1993. A translation approach to portable ontology specifications, Knowledge Acquisition, 5 (2), pp 199-220.
- [4] Saias, J. Quaresma, P. 2003. A Methodology to Create Ontology-Based Information Retrieval Systems, In Proceedings of the EPIA Conference, pp 424-434.
- [5] Koo, S. Lim, S.Y. Lee, S.J. 2003. Building an Ontology based on Hub Words for Informational Retrieval, In Proceedings of the IEEE/WIC International Conference on Web Intelligence.
- [6] Vallet, D. Fernández, N. Castells, P. 2005. An Ontology-Based Information Retrieval Model, In Proceedings of the 2nd European Semantic Web Conference, pp 455-470.
- [7] Baziz, B. Boughanem, M. Chriment, C. 2005. Semantic Cores for Representing Documents in IR, In Proceedings of the 20th ACM Symposium on Applied Computing, pp. 1020-1026, ACM Press ISBN: 1-58113-964-0.
- [8] Cui, H. Wen, J.R. Nie, J.Y. 2002. Probabilistic query expansion using query logs. Proceeding of the 11th International Conference on World Wide Web, May 07-11, ACM New York, NY, USA. pp: 325-332. DOI: 10.1145/511446.511489.
- [9] Lin, H.C. Wang, L.H. 2006. Query expansion for document retrieval based on fuzzy rules and user relevance feedback techniques. Expert Systems with Appli., 31: 97-405. DOI: 10.1016/j.eswa.2005.09.078
- [10] Navigli, R. Velardi, P. 2003. An analysis of ontology-based query expansion strategies. Proceeding of the Workshop on Adaptive Text Extraction and Mining, Sept. 2003, Dubrovnik-Croatia. pp: 42-49.
- [11] Schreiber, G. Wielinga, B. Jansweijer, W. 1995. The tactus view of the 'o' word. IJCAI'1995, Workshop on Basic Ontological Issues in Knowledge.