

Experimental Assessment of the TARGET Adaptive Ontology-based Web Search Framework

Nicolas Guelfi
LASSY - University of Luxembourg
Email: nicolas.guelfi@uni.lu

Cédric Pruski
CR SANTEC
Centre de Recherche Public Henri Tudor
Luxembourg
Email: cedric.pruski@tudor.lu

Chantal Reynaud
LRI-Univ. of Paris-Sud-CNRS &
INRIA Saclay - Île-de-France
Email: chantal.reynaud@lri.fr

Abstract—Finding relevant information on the Web can be a complex task for most of the users. Although Web search applications are improving, they still need to be more intelligent to adapt to the search domain targeted by users, the evolution of this domain and users' characteristics. In this paper, we present an experimental assessment of the TARGET framework for improving the relevance of the documents when users are searching the Web by using adaptive ontologies. This is done first by introducing the TARGET approach. We will briefly present the used ontologies and their ability to adapt to domain evolution. Then, We detail the TARGET tool used in our experimentations. This includes its architecture, its ability to carry out the ontology adaptation process as well as the way it searches the Web and ranks the returned results. Finally, we discuss the results obtained using the tool through the presentation of our case study devoted to the retrieval of scientific articles.

Index Terms—Ontology evolution; Semantic Search; Information retrieval;

I. INTRODUCTION

Since the advent of the WWW, the ever-increasing number of documents available through the Web requires the development of new intelligent tools in order to assist users to find relevant information. Actually, one of the main difficulties for Web users lies in the construction of good queries. The choice of appropriate keywords for targeting the right search domain is often hard. Consider the query "publications on trees". It is difficult to decide whether the term tree refers to graph theory or to botany. In addition, by virtue of knowledge evolution, the knowledge of the search domain is constantly changing over time what makes the selection of the right keywords more complex as users must be aware of these evolutions.

Besides, the domain targeted by the query is often huge. In fact, a given domain contains information for many different kinds of users and it is hard for usual search engines to decide what information is relevant for a particular user only by considering his initial query. In consequence, the person who entered the query must skim the returned results only to keep the information relevant to his profile. Consider, for instance, the scientific research domain. An academic researcher may not be interested in the same kind of information as an industrial researcher may be even if the same initial query, which aims at targeting the same search domain, is entered.

As a result, we believe that the new generation of Web search tools should assist users to target both the search

domain and integrate both domain evolutions and users' profile. These elements (search domain and user profile) contain information for filtering unwanted documents. However, the characterisation of the evolution of a domain is a complex problem and has rarely been considered for Web search. In order to cope with this gap, we propose the TARGET framework. It implements an original approach based on adaptive ontologies for improving Web search. Such ontologies are used to represent search domain and users' characteristics, they are able to adapt to domain evolution, and serve for query enrichment and ranking purposes. In this paper, we present an experimental assessment of this framework through a case study.

The remainder of this paper is structured as follows. Section II briefly introduces the TARGET approach. Section III deals with the tool implementing the proposed approach. We present in Section IV the experimental assessment of the framework. Section V proposes some related work in the field adaptive Web search. Finally, last section wraps up with our concluding remarks and outlines future work.

II. THE TARGET APPROACH

The TARGET approach 1), aims at improving (in terms of relevance) the results of a Web search.

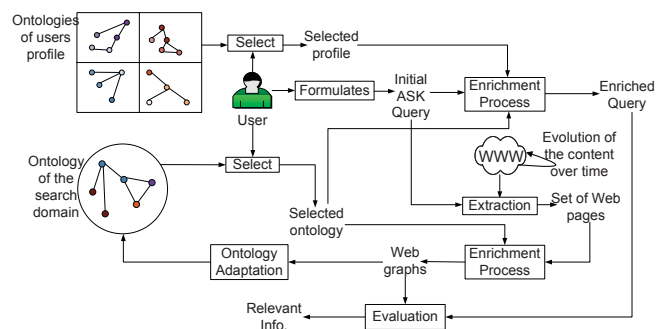


Fig. 1. The TARGET Approach

Two phases constitute the overall process. In the first one, users play the main part. As it is the case in existing Web search engine, user enters his query (here using the ASK language [1]). Then, he has to select both the adaptive

ontology representing the targeted search domain and the one corresponding to the profile that characterises him best (models of user profile in 1). The second part is completely automatic and transparent for the user. The system extracts Web pages via a common Web search engine. In parallel, the system enriches the emitted query using the selected ontologies according to rigorously defined query enrichment rules [2] and transforms the pages returned by the engine into WPGraphs and W3Graph [1] (Web graphs on 1) using the domain ontology, for ranking purpose. Finally, the enriched query is evaluated on the previously built Web graphs and the most relevant Web pages are selected and ranked.

The originality of the TARGET approach is the utilisation of adaptive ontologies within the search process. This approach not only represents the knowledge of different domains but they also have the capability to adapt to the evolutions of their associated domain. Therefore, we have to zoom in the characteristics of such ontologies to understand their properties as well as their contribution in the enhancement of Web search.

A. Adaptive ontologies

TARGET is adaptive as the used ontologies consider the evolution of the knowledge of their respective domains. To make our ontologies adaptive, we had to propose a set of features for characterising knowledge evolution. These features are defined from general knowledge representation and management approaches and have been assessed via a study devoted to domain specific ontology evolution [3]. In this study, we observed the emergence and/or deletion of knowledge. It occurs when new concepts emerge in the domain or in the contrary when obsolete ones are removed from it. This phenomenon can be modelled using a date. Some concepts persist in the domain over time, different concepts do not have the same semantic weight (i.e. importance) and the semantic distance that separates concepts (i.e. concepts directly linked by a given ontological relation) can vary according to their usage in the domain. These features can be represented using integers assigned to concepts and relations of the ontology.

Lastly, we identified the ability of knowledge to resist to ongoing changes. This resistance, modelled using a real number assigned to each relation of the ontology, prevents the deletion of concepts of the domain or the modification of their semantic weight or semantic distance.

Concerning the representation of these features in existing ontology language, OWL [4] provides what is called *annotation property* that can be used to annotated ontology elements with particular information. The emergence of concept is represented using a date that corresponds to the moment when the concept integrates the domain. The persistence duration as well as the semantic weight and the semantic distance are represented using integers whereas the resistance to change is represented using a real number. Observe that the use of annotation properties requires OWL DL because we annotate only classes and object properties, which is important for reasoning purposes, since OWL DL is decidable.

These features are the foundation of our model but their associated values have to be modified over time. Thus, we define *ontology evolution* as semi-automatic incremental process. An evolution step aims at fixing the new ontology definition by applying adaptation rules on its elements. At each evolution step, we must:

- 1) Integrate every new concept of the domain resulting from the analyse of the evolution of the domain (section III-B). This implies to the definition of new concepts in the ontology, of relations to link the new concepts to the existing ones and their corresponding evolution features.
- 2) Re-evaluate persistence duration of each concept and remove those whose persistence duration is zero.
- 3) Re-evaluate the semantic distance for each relation.
- 4) Re-evaluate the semantic weight for each concept.
- 5) Reassign the resistance for existing relations.

This adaptation process only requires user intervention for linking new concepts to those of the ontology. The tool implementing our approach (section III) assists users in this task. The remaining of the process is automatic and uses a corpus of documents built as one goes along with the Web documents returned to the user after a search. We detail this aspect as well as the various implemented metrics in section III-B. We also propose an experimental assessment of the rules and discuss the results in section IV.

B. Adaptive Web search

Web search in the TARGET approach is done using adaptive ontologies. This model of ontologies combined with the adaptation rules serve to represent both the search domain targeted by users and the different views they could have over the domain (Fig. 1). These ontologies are used for two different purposes.

First, adaptive ontologies representing the search domain and the user's profile serve as basis for query enrichment. In [2], we proposed a set of ontology-based Web query enrichment rules. These rules rely on various ontological relations and aim at extracting the concepts of the ontology that best characterized the domain regarding the terms of the initial query. The extracted concepts are added, in a rigorously defined manner, to the initial query. Nevertheless, depending on the size of the ontologies, the number of extracted concepts can be high which, in turn, affects the quality of the enriched query. Thus, we refine the list of concepts using the evolution features. To this end, we first favour the semantic distance, then the semantic weight and last, the emergence date. It means that we keep the concepts that are close (from the semantic distance point of view) from those of the initial query, then if too many concepts remain, we keep those whose semantic weight is the highest and eventually filter the oldest (regarding the emergence date) ones. Then we add the most important concepts (in the sense of the semantic weight) of the selected user's profile. The so-enriched query contains information related to search domain and user's characteristics and will allow to target the right documents at query interpretation time.

The second facet concerns the ranking of the relevant pages. This is done based on the structural properties of the Web graphs that are constructed using adaptive ontologies (see section III-C).

III. IMPLEMENTING THE APPROACH: THE TARGET TOOL

This section presents the specificities of the tool that implements our approach. As the approach integrates ontology adaptation and Web search, we will base the presentation of the tool on these two perspectives.

A. Tool architecture

The architecture of the tool (<http://se2c.uni.lu/tiki-index.php?page=TargetTool>) that implements the TARGET approach is depicted in figure 2 hereafter.

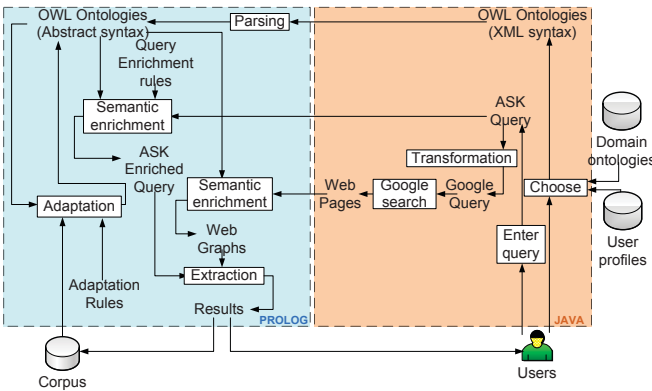


Fig. 2. Tool architecture

The tool is made up of two parts (Fig. 2). A Java front-end that acts as an interface with the user and the second part made up of the PROLOG inference engine that manages both ontology adaptation and the adaptive Web search. Actually, the OWL adaptive ontologies selected by users are parsed using THEA PROLOG (<http://semanticweb.gr/TheaOWLLib/>) and loaded as PROLOG facts in the inference engine.

B. Ontology adaptation

The adaptation process is based on a corpus (see fig. 2) incrementally built of pages returned to users at search time. We believe that the information contained in these pages is significant to the search domain so it is exploited for ontology adaptation purpose.

To explain the performed adaptation process, we follow the steps given at section II-A. To detect potential new concepts, the system analyses the corpus. We assume that the terms of the corpus that do not represent labels of concept in the ontology yet and that appear the more frequently on these pages can be considered as relevant to the domain. In consequence, they must be proposed to users as labels of new concepts at evolution time. The system then supports the definition of the new concept. It permits the definition of the relation that links the new concept to existing ones. Moreover, the evolution features' values of the new concepts can be

set up (if the user does not agree with the proposed default values) and those of the existing concepts can be adjusted. The other steps of the process implement several metrics. These are defined using statistics applied to the documents of the corpus and concern semantic distance, semantic weight and persistence duration.

The semantic distance between two concepts c and d at time τ , $SD_{(c,d)\tau}$, is defined as a function of its former value, $SD_{(c,d)\tau-1}$, the resistance to change, $R_{(c,d)}$, and a value computed on the documents of the corpus that have been added between $\tau - 1$ and τ namely $\Sigma_{(c,d)\tau}$. This latter value represents the average of words separating consecutive occurrences of c and d in the newly added documents. For instance, if in a document the first occurrence of c and d are separated by 10 words and if there are 20 words between the second occurrences of both terms the corresponding value of $\Sigma_{(c,d)\tau}$ is $(10+20)/2 = 15$. As a result,

$$SD_{(c,d)\tau} = SD_{(c,d)\tau-1} + \frac{\Delta SD_{(c,d)}}{R_{(c,d)}}$$

with

$$\Delta SD_{(c,d)} = \Sigma_{(c,d)\tau} - SD_{(c,d)\tau-1}$$

We assume that the more frequently two concepts are jointly cited and the less words separate these concepts in the relevant documents of the domain, the closer they are (from the semantics point of view). Therefore, we take into account the variation of the semantic distance engendered by the evolution weighted by the resistance to change. The semantic weight of a concept c ($SW_{(c)\tau}$) is updated in the same manner but here the value computed on the corpus ($I_{(c)\tau}$) is the frequency of concept c in the newly added documents. Hence:

$$SW_{(c)\tau} = SW_{(c)\tau-1} + \frac{\Delta SW_{(c)}}{R_{(c)}}$$

with

$$\Delta SW_{(c)} = I_{(c)\tau} - SW_{(c)\tau-1}$$

We assume that the semantic weight of a concept is linked to the frequency its associated label is cited in the documents of the domain. The resistance plays an important part in the evolution process since it has an influence on the impact of the evolution of the domain. A high resistance slows down the modifications of the semantic distance and semantic weight. In the contrary, a low resistance speeds up the evolution process. Concerning persistence duration, resistance is applied in a different manner. In fact, the value of the resistance tells how many steps it will take to reduce the persistence duration. For instance, a resistance of 2 means that the persistence will decrease every 2 evolution steps while a resistance of 0.5 means a decrease of 2 units in 1 evolution step. A value set to 0 for the resistance associated to a concept implies a direct deletion of the concept from the ontology since its persistence duration will be set to 0.

The ontology adaptation process we proposed is consistent in the sense that the resulting ontology remains coherent (i.e. reasoning can still be performed) with respect to the

modifications that have been done mainly the addition and/or deletion of new concepts and relations that engender structural modifications on the ontology. This property can be checked easily using the predicates offered by the THEA library.

C. Web search

The tool uses Google as an entry point to the Web. Actually, the initial query is sent to Google, then it is enriched using the selected domain ontology and user profile according to the aforementioned rules. The pages returned by Google are transformed using the domain ontology into WPGraphs and W^3 Graphs and loaded into the PROLOG inference engine as PROLOG facts. Finally, the enriched query is processed as a PROLOG formula over the PROLOG Web graphs. The system re-ranks pages, which fully match the enriched query, and displays it on user's screen. The ranking we propose relies on the structure of the Web graphs. In fact, to each Web page corresponds a WPGraph [1]. The vertices of the graph represent the label of concepts (or terms) of the page and edges represent the semantic link between concepts. The set of edges is built using the domain ontology. Consider two terms of a page, if these terms are part of the domain ontology and if there is a path between these concepts in the ontology, then there is an edge between these two concepts in the WPGraph. As a result, the more edges the WPGraph will have, the most relevant (regarding the domain ontology) the associated Web page will be. Hence, we rank the results with respect to the number of edges of the WPGraphs that fully match the enriched query.

IV. EXPERIMENTAL ASSESSMENT

In previous sections, we have presented the different components of the TARGET approach. As explained, the adaptation is made with respect to domain evolution and user's profile and using ontology-based query enrichment rules. In order to strengthen these aspects, we made several experimentations using the tool detailed in section III.

A. The case study

Our case study deals with the retrieval of scientific articles published from 1996 to 2006 at the World Wide Web series of conference. This set of documents forms our "micro" Web. From the call for papers of the conference we have built the (initial) domain ontology [3], made this ontology evolved and used it to build Web graphs. Then, we use the obtained graphs and ontologies for query enrichment purpose.

Our micro Web has the advantage to be finite and of suitable size (i.e. about 600 documents). Thus we were able to measure the recall and the precision of the returned results, which is of utmost importance for information retrieval approaches. Moreover, because of this micro Web, we have shortcut the use of Google as the entry point to the Web for our experimentations.

Concerning the user profiles, we defined two various profiles. The first one reflects the basic knowledge of a fundamental researcher (i.e. someone who is interested in fundamental

research). In consequence, the ontology representing this domain is made up of concepts like formal, mathematics, algebra, proof, theorem and so on. The second profile describes the knowledge of an applied researcher, which corresponds, in some words, to an industrial researcher. Therefore, concepts of this ontology are rather: tool, prototype, application, industry, etc.

B. Scenarios

Our objective is to show that our adaptation process and query enrichment rules improve the precision and the recall of the results of a Web search. Thus, we propose four scenarios based on our case study.

The first one (scen1 on fig. 3 and 4) is the basic case. It corresponds to the Web search process offered by usual Web search engines. Therefore, we neither perform any query enrichment nor use any user profiles or domain ontologies. Observe that we restrain the search space to our micro web. This consists in a first filter, which is not the case when we search the Web using a Web search engine. In consequence, the precision of the returned results is already enhanced.

The second set of experimentations (scen2) highlights the benefit of using ontologies for improving Web search. It consists in implementing classic OWL ontology for representing the search domain and uses it as input for query enrichment rules in order to enrich the initial query. The reader can refer to [2] for a detailed example of query enrichment.

The third one (scen3) put the stress on the benefit of adaptive ontologies. We use adaptive ontologies for representing the search domain but we do not use any user profile in the query enrichment process.

The last one (scen4) implements the full TARGET approach, which includes the use of adaptive ontologies obtained using the adaptation process presented section III-B, user's profile and advanced query enrichment rules.

For all scenarios, we have tested a hundred of queries implementing all the constructors of the ASK query language and we measure manually the precision and the recall of the documents returned by our tool.

C. Experimental results

In order to build the set of initial queries, we consider ourselves as Web users. Thus, we had to define the focus of the search and the corresponding profile (i.e fundamental or industrial researchers). Figures 3, 4 and 5 contain the results of our experiences¹. The graphics represent the precision, recall and F-Score respectively.

As illustrated the TARGET approach gives better results as usual search applications or basic ontology-based ones. This is all the more so true concerning the precision of the returned documents. Actually, the TARGET approach (Scen. 4 on the graphics) reaches an average of 80% in precision. The high precision underlines the quality of the adaptation process defined in section III-B and the comparison of the precision

¹On the graphics, the x axis represents the query number and the y axis represents the ratio

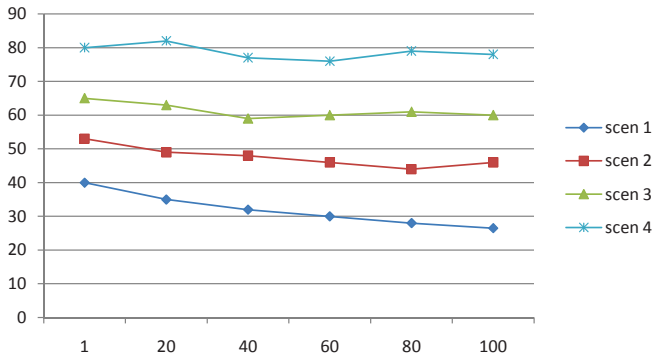


Fig. 3. Precision of the search results

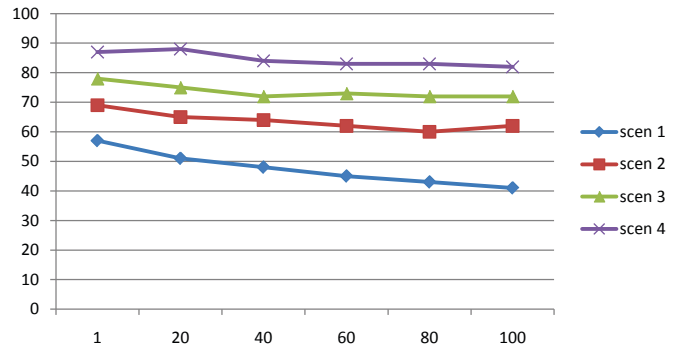


Fig. 5. F-Measure of the search results

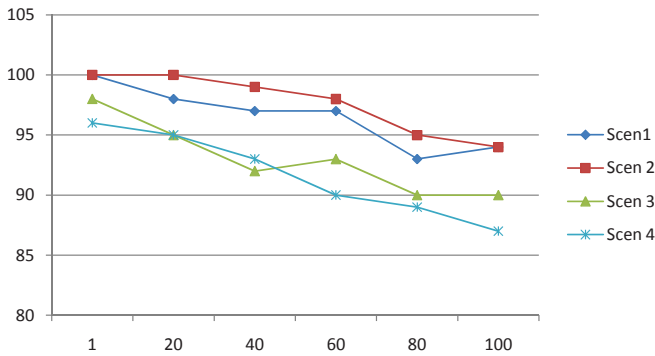


Fig. 4. Recall of the search results

of scen2 and scen3 shows the interest to consider adaptive ontologies. Furthermore, the difference of the precision between scen3 and scen4 highlights the benefit of using user's profile for searching the Web. In fact, the adaptive ontology representing the search domain allows to target the right search space while the profile of the user permits to filter among the relevant pages of the domain the remaining pages that will not interest the user. This explains the high precision.

Recall is less affected. Actually, as aforementioned in the paper, our main objective is to increase the relevance of the returned results and our measures of the precision reinforce our position but the precision increases to the detriment of the recall. Because we filter many unwanted pages (by virtue of the query enrichment rules where we favour the use of the conjunction operator of the ASK query language), it happens that some pages that could be considered as relevant, slip among hence affecting the recall. The other scenarios have a better recall because they return a huge number of documents and among them of course there are the most relevant ones. This, in turn, reduces drastically the precision of the returned results.

The measurement of the F-Score confirms what has been said. The concepts developed in the framework of the TARGET approach contribute in enhancing the result of a search and offers the best compromise recall/precision. Our approach is particularly efficient when the keywords of the initial query are imprecise, ambiguous or few, which is the case for

most of Web queries [5]. This is due to adaptive ontologies representing the search domain that allows the disambiguation of the query by adding the most relevant keywords to it. Furthermore, information extracted from user's profile and added to the query reinforces the filtering effect of the enriched query.

The complexity in time of the TARGET approach could be improve because of the construction of the Web graphs, which is time consuming. However, as a guide, during our experimentations the building of Web graphs, representing about six hundred of huge (i.e. 8000 words per page) Web pages, took less than five minutes, which means less than one minute for hundred pages (this is important as we apply this approach on the hundred first pages returned by Google). This construction cost is largely compensated by the quality of the search results compared to the very low precision of usual search application (scen1 on figure 3). We believe that it is acceptable for a user to wait few seconds to have the relevant information he is really interested in instead of wasting his time to skim the return results or to build several queries until reaching the relevant information.

V. RELATED WORK

In the field adaptive Web search, various interesting approaches were proposed. They differentiate mainly in the way the adaptation is performed.

First of all, many approaches claim that Web search systems should directly adapt to the characteristics of the user who initiates the search. Such an approach has been proposed in [6] where the authors build a profile of the user based on information related to his Web navigation history. The approach proposed by Liu et al. [7] advocates the study of user's previous queries to build his associated profile. Although having shown great results, this kind of approach raises several problems. The first one concerns the quality of the so build profile. Actually, the behaviour of Web users can influence the quality of the profile in particular when users are not focusing on a given domain but have numerous and completely different focuses. The profile will be vague and can hardly be reuse for Web search. The second major problem deals with data privacy issues because personal information are stored and treated without users' agreement.

Second, existing approaches put the stress on the context of the search [8]. Most of them apply query expansion techniques using various artefacts like linguistic tools such as ontologies [9], statistics heuristics or machine learning techniques based on text collections [10], or feedback [11]. This consists roughly in extracting domain specific keywords and appending them to the initial query. This leads to the disambiguation of the query. Moreover, Chirita et al. [12] combine personalisation and query reformulation by analysing the so-called Personal Information Repository (i.e. the personal collection of text documents, emails, cached web pages, etc). However, not any of the evoked approaches take into account the domain evolution problem. This can be problematic mainly for Web search approaches that rely on the use of linguistic tools.

To wrap up, we try to make the most from both families of approaches. Hence, our work is original from existing one in the sense that we use ontologies for representing both the search domain and user's profile that have the ability to adapt to the evolution of their respective domains for enhancing Web search. Moreover, we use these adaptive ontologies as input of query enrichment rules and ranking technique that have shown great performances in Web search.

VI. CONCLUSION

In this paper we have presented the TARGET framework for optimizing the relevance of the results when searching the Web. Our approach, which integrates domain evolution and user profile in order to make the search adaptive, has shown interesting results as highlighted by the serious experimental validation we have carried out. Nevertheless, some aspects of the approach need improvements. Actually, the construction of the Web graphs is time consuming which could be considered as a drawback. This is why our future work will be devoted to enhance the construction of these data structures. We also plan to improve the adaptation process by making it as automatic as possible. The aspect of the user interface dealing with the query capture and the display of the results is also under improvement.

REFERENCES

- [1] N. Guelfi and C. Pruski, "On the use of ontologies for an optimal representation and exploration of the web," *Journal of Digital Information Management (JDIM)*, vol. 4, no. 3, pp. 159–168, September 2006.
- [2] N. Guelfi, C. Pruski, and C. Reynaud, "Les ontologies pour la recherche cible d'information sur le Web: une utilisation et extension d'OWL pour l'expansion de requetes," in *Proceedings of the Ingenierie des Connaissances 2007 (IC07) french conference*, Grenoble, July 2007.
- [3] —, "Understanding and Supporting Ontology Evolution by Observing the WWW Conference," in *Proceedings of the International Workshop on Emergent Semantic and Ontology Evolution (ESOE 07)*, Busan, South-Korea, November 2007.
- [4] D. McGuinness and F. van Harmelen, "OWL Web Ontology Language Overview," W3C Recommendation, February 2004.
- [5] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz, "Analysis of a very large web search engine query log," *SIGIR Forum*, vol. 33, no. 1, pp. 6–12, 1999.
- [6] K. Sugiyama, K. Hatano, and M. Yoshikawa, "Adaptive web search based on user profile constructed without any effort from users," in *WWW '04: Proceedings of the 13th international conference on World Wide Web*. ACM, 2004, pp. 675–684.
- [7] F. Liu, C. Yu, and W. Meng, "Personalized web search by mapping user queries to categories," in *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management*. ACM, 2002, pp. 558–565.
- [8] S. Lawrence, "Context in web search," *IEEE Data Engineering Bulletin*, vol. 23, no. 3, pp. 25–32, 2000.
- [9] J. Bhogal, A. Macfarlane, and P. Smith, "A review of ontology based query expansion," *Information Processing and Management*, vol. 43, pp. 866–886, 2007.
- [10] S. Oyama, T. Kokubo, and T. Ishida, "Domain-specific web search with keyword spices," *IEEE Trans. on Knowl. and Data Eng.*, vol. 16, pp. 17–27, 2004.
- [11] Z. Chen, X. Meng, B. Zhu, and R. H. Fowler, "Websail: From on-line learning to web search," *Knowledge and Information Systems*, vol. 4, no. 2, pp. 219–227, 2002.
- [12] P. A. Chirita, C. S. Firan, and W. Nejdl, "Personalized query expansion for the web," in *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. Amsterdam, The Netherlands: ACM, 2007, pp. 7–14.