8-5-2011

Contextual Query based on Segmentation and Clustering of Selected Documents for Acquiring Web Documents for Supporting Knowledge Management

João C. Prates
UNIRIO, joao.prates@gmail.com

Sean W. M. Siqueira
UNIRIO, sean@uniriotec.br

Contextual Query based on Segmentation and Clustering of Selected Documents for Acquiring Web Documents for Supporting Knowledge Management

João C. Prates
UNIRIO
joao.prates@gmail.com

Sean S. M. Siqueira
UNIRIO
sean@uniriotec.br

ABSTRACT

The Internet is a valuable source of information whose volume of data and number of accesses grows every day. Web search engines have a key role in the discovery of relevant information, but this kind of search is usually performed using keywords and the results do not consider the context. This paper describes the use of information extraction techniques applied in previously defined resources in order to expand the queries made by users and run these expanded queries in web search engines, getting more useful search results, considering the domain context of the required information. A prototype was developed according to the proposed strategy and a case study was conducted in a Brazilian public institution. The results show that the proposed approach can be used in a corporative environment to help the execution of contextual search activities with good results.

Keywords

Adaptive search, context, query expansion.

INTRODUCTION

In search tools, the user usually reports an information need by entering keywords in search expressions. Although this approach makes it easy for users to express their information needs, it also results in contextless results. However, information retrieval is strongly dependent on the context. What a user, who has a specific knowledge and a specific experience, believes as relevant may not be relevant to another user with distinct characteristics and experiences, even if the search expressions used by both of them are the same.

Other problems of this approach are: (i) the user needs to know the terms contained in the web pages/documents in order to get the results that will meet his/her information needs, and (ii) words can have different meanings and the disambiguation of these concepts is from the responsibility of the user himself/herself, that must modify the query by adding or changing the keywords that were used.

Another important issue that maximizes these problems is the user's behavior when using Internet search tools. Users browse a few results pages, often limited to the first 5 results (Spink and Jansen, 2004), but statistics show that search expressions typically consist of few terms. 65% of searches on the Internet contain from one to three terms (Experian, 2010).

In order to make search results more relevant considering the users' behavior when building a query, it is used an information retrieval technique known as query expansion. In this technique, terms are added in the original query made by the user in an attempt to provide a greater contextualization, and retrieving more useful documents (Yates and Neto, 1999). The terms that are added to the query, taken from the context of the information need, can minimize the reported problems. In this paper, context is considered as any piece of information that can be used to characterize the situation of an entity. An entity is a person, a place or an object that is considered relevant to the interaction between a user and an application, including the user and applications themselves (Dey and Abowd, 1999).

This paper presents a proposal to make web searches adaptive to the context of the users, according to their information needs, thus improving query results.

The contextualization is provided through the expansion of queries entered by users, adding to these queries some terms extracted from selected documents that are representative of the context of the information need. An information system that supports web search engines according to the proposed approach was developed and used in a case study conducted in a Brazilian public institution. The aim of this case study was evaluating the use of proposed solution by different professional profiles in real activities of knowledge management.

CONTEXT IN INFORMATION RETRIEVAL

In general, some knowledge representation is used to model some kind of context, which can be the domain of some knowledge, the user profile, the activity developed at the time, search history or information obtained through sensors (place, temperature). This context model can be created manually, with the direct interference of experts or users, for example through the creation of ontologies, filling out preferences forms or marking relevant documents. This approach is used in (Chanana, Ginige and Murugesan, 2004), which proposes a methodology for context classification of documents, based on the type of information contained in the collection. Some researches use ontologies to improve search efficiency (Bhogal, Macfarlane and Smith, 2007).

Although performing well in making the search tool more adaptive to the context, and therefore provide more useful results, the manual context creation presents as some major problems the time and effort spent by experts or by the end users in building the context model, which can derail the use of search tools.

Another possibility is the automatic creation of context, through inferences based on user behavior (analysis of clickthrough, navigation and queries history) or information from the environment or some domain information. In (Chien, Hu and Ju, 2007), knowledge in one domain is used for the automatic construction of context. However, the construction process should be applied to a specific collection of documents available for searching, making it impossible it use in the Internet. The clickthrough analysis is used in (Wang et al., 2009), which proposed a method to try to identify when the clicked link is valid or not in addressing the need for information.

The automatic creation of contexts, although doesn't present the problems of manual creation, may cause incorrect results if the inferred context does not reflect the current information needs of the user.

Another form of automatic model creation is to extract knowledge through text mining in document collections, using techniques such as stemming, clustering and co-occurrence of terms. Some works that use query expansion with clustering (also the focus of this work) are presented in (Bhogal, Macfarlane and Smith, 2007). As the main differences found in this study compared to previous works are: (i) the application of topic segmentation before the clustering, with the objective of grouping all the different subjects found in the collection, not only the documents, (ii) the use of terms of all clusters, not only the terms of the cluster most related to the query terms and (iii) quality assurance in the selection of documents used, consisting of educational resources selected by teachers as a reference for classes.

ADAPTIVE SEARCH BASED ON THE EXTRACTION OF CONTEXTUAL INFORMATION

For the search to become sensitive to a domain context, the strategy proposed in this work is based on the following hypothesis: the terms most often found in an information resource that is representative of a domain are more likely to also be present in other related and relevant documents available in Internet. Therefore, when using these terms to expand queries made by users, it is possible to obtain more useful search results.

However, considering that each resource that represents a context can have different topics in its content, a simple extraction of terms based on their occurrences in the resource can result in a combination of terms of different subjects in a query, possibly reducing the probability of obtaining useful results in the search.

To minimize this risk, it is proposed to perform two extractions of terms: one for acquiring the general terms of context (**context extraction**), in which the calculation of the weights of the terms is based only on their frequency on the resource; and another that is sensitive to the different subjects existing in the context (**subject extraction),** whose calculation takes into account the frequency of terms in the subjects partition and in the whole collection.

In order to demonstrate the applicability of the proposed approach, an information system was developed. The solution is divided into three modules: Knowledge Base Configuration, Information Extraction and Search.

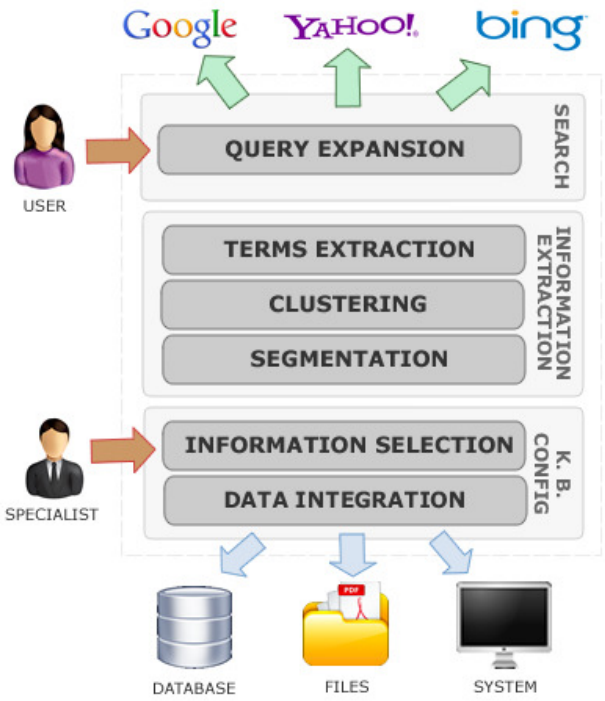

Figure 1. The proposed architecture.

Knowledge Base Configuration Module

The domain context is modeled with the use of existing resources such as databases, miscellaneous files (articles, book chapters, publications in general) or information systems (including data integrations performed through web services, connectors, etc.). Any information source that contains textual content and information, which represent the topics discussed in the domain context, can be used for this purpose.

This module has two components. The contextual information must be accessed through the system, so it is necessary to create a connector layer, called the data integration, which contains a set of specific software that can read every kind of information available.

A domain expert must, from all sources of information available and accessible through the system, select the contents that are good representatives of the subjects that compose the context.

Information Extraction module

The objective of the Information Extraction module is to identify the main terms of all the contextual information obtained from the Knowledge Base Configuration module, and to provide a list of terms to the search module. Two extractions of terms are executed, one for the most frequently used terms in the context (context extraction), and another for specific terms of each subject identified in context (subject extraction).

In both situations, it is necessary to apply some activities of text preparation before the extraction of terms: (i) the Tokenization, the process of breaking a stream of text into words, phrases, symbols and other meaningful elements called tokens, (ii) the removal of stop words, a list of common or general terms that have little value in the text and must not be extracted, and (iii) stemming, the process for reducing inflected (or sometimes derived) words to their stem, base or root form (Manning, Raghavan and Schütze, 2008).

The extraction of general terms of context is done by calculating the weights of the terms and extracting the n terms with highest weight, where n is the maximum number of terms that can be used in the expansion of the query. The weight calculation is performed with the formula of sublinear term frequency scaling (1) (Manning, Raghavan and Schütze, 2008).

$$wf_{t,d} = \begin{cases} 1 + \log tf_{t,d}, & \text{if } tf_{t,d} > 0 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The extraction of the specific terms of each subject that is identified in the context requires the application of more information extraction activities. The first step is the use of a routine of text segmentation to divide the contents of the contextual information into sentences. The text segmentation is an activity of natural language processing that aims to identify subtopics within a document, defining its limits. The objective in applying the segmentation is to ensure that a particular textual content contains only one subject, thus ensuring that there are no terms that deal with different subjects in such content.

Each subtopic (subject) that was identified in the extraction can be treated as an independent document. However, segments in the collection that can address the same subject or similar subjects must be grouped for modeling the context.

In order to group these subjects it was used clustering, clustering must be used. Clustering is a statistical technique that allows automatic generation of grouping data (documents). In this technique, the vectors of terms from each document are grouped according to the calculation of functions of distance and similarity applied to each pair of vectors.

After all the contextual information is grouped by subject, the weights of all terms are calculated, considering each subject that was identified as a collection of independent documents, also using the weight expression (1). After the calculation of weights for all de terms, the *n* terms with highest weight are extracted for each identified subject.

Search module

The search module receives two pieces of information from the users: an indication of the domain context in which the information need is applied and the keywords to perform the search on the web. The original query (obtained from the keywords) is expanded using the terms extracted in the Information Extraction module, and executed in the web browser.

The original query can be expanded according to two ways. The first is the **automatic expansion**, in which the original query is expanded n times, where n is equal to one (context expansion) plus the number of subjects that were identified in the selected context (subject expansion). Each expanded query is executed on the selected web browser, and the results are presented to the user in tabs.

The second mode of query expansion is the **suggestion of terms**, in which all extracted terms (generated by the context extraction and the subject extractions) are presented as a suggestion. The user has to select the terms of his/her interest that will be incorporated to the original query, performing the query expansion. When selecting the terms, the user can combine general terms with specific terms of the subjects.

PROTOTYPE

The proposed architecture was materialized in a web prototype developed in Java. In this implementation, some open source libraries were used, especially in the activities of the Information Extraction module. In the Knowledge Base Configuration module, a plugin was designed to integrate with a news database used in the case study. As the source of information used by the prototype was a database, the filter was implemented as a where clause in the SQL statement that consulted all the news whose classifications were related to the configured contexts.

The text segmentation task was implemented using the suite of tools for natural language processing called MorphAdorner[1], which implements the linear Texttiling segmentation method (Hearst, 1997). For the grouping of segments, an implementation of the clustering algorithm k-means (Manning, Raghavan and Schütze, 2008), available in the suite of tools for text mining Textgarden[2], was used.

The expanded queries are executed at Google. The search results are displayed in tabs, one for each query expansion. The results of the original query done by the user are also displayed, on the first tab.

Since the solution combines different techniques of extracting information, which in their algorithms may use different parameterizations, prior to the case study it was necessary to perform experiments to define the influence of these parameters in the query expansions.

[1] http://morphadorner.northwestern.edu/

[2] http://textgarden.org/

THE CASE STUDY

The use of case study methodology to evaluate the proposed system is justified by the absence of a paradigm for automatic evaluation of context-sensitive search tools. The case study was conducted in a Brazilian public banking institution, whose purpose is to act in the long-term financing for investments in several segments of the national economy.

The prototype was evaluated by different user profiles and was configured with two different contexts: Energy and Natural Environment. Regarding the profile, the case study intended to assess how the proposed tool could support the execution of the search activities depending on the knowledge on the context to be searched and the background of each participant. Two groups of users were selected: librarians (information specialists, who search on the Internet every day and therefore have a greater affinity with search engines and better words for querying) and employees of sectorial departments (specialists in the subject domain, the main consumers of the information about economic sectors).

Five librarians and six workers from sectorial departments (three on each context - Natural Environment and Energy) participated in the case study. The librarians did two evaluations, one for each context.

In the case study, the participants should elaborate information needs and queries that represent these needs, and use the prototype to do search in their two available modes: automatic expansion and terms suggestion. In addition, the participants should evaluate the results obtained from the searches according to their relevance. After using the tool, the participants should also answer a questionnaire aimed at identifying the qualitative criteria of the tool assessment.

Metrics used for evaluating the results

Collections of test widely used in information retrieval to evaluate tools and search algorithms can not be applied in adaptive search tools, because they use other kinds of information beyond the need for information to determine which documents must be returned in a search. According to (Voorhees, 2008), evaluation of adaptive search engines is an open research problem. Therefore, the results were evaluated according to three metrics proposed by different researchers and consolidated by (Tang and Sun, 2003) for the evaluation of web search engines: first 10 full precision, search length and rank correlation.

Precision is a metric widely used in information retrieval and represents the fraction of relevant documents among all retrieved documents (Manning, Raghavan and Schütze, 2008). However, the binary judgment of relevance adopted in traditional evaluations does not take into account the different amounts of relevant information found in each document. The full precision metric tries to consider the total amount of relevant information found in the first 10 results, through the use of a five positions scale for the relevance judgments.

The research described in this paper considered an adaptation of the full precision metric applied in the (Tang and Sun, 2003), where the evaluation was applied to the first 20 results. The change in the number of evaluated results is justified by studies that show the user behavior on searching the Internet, where most of users access only the first page of results and generally limited to the first 5 results (Spink and Jansen, 2004). Since the more visited web search engines (Google, Yahoo, and Bing) use as default setting to display 10 results per page, only this number of records were evaluated, corresponding to the first page of results.

The second metric was the search length, which reflects the number of not relevant documents that the user must evaluate until he/she finds a certain number of consecutive documents that are considered relevant. Therefore, the lower the value, lower the effort for the user to find relevant results. As in (Tang and Sun, 2003), in this study the search length was defined as the number of documents evaluated until two consecutive results were found with the value of relevance greater than or equal to three.

The last metric, which is called rank correlation, aims to compare the correlation between the priority obtained in the search with an ideal priority, where the results are sorted in descending order of relevance. The higher the correlation between the relevance of search results and the ideal prioritization, more efficient is the search tool. Since the first ten results were evaluated and the evaluation scale is five positions, the prioritization has to be considered ideal if in the first two positions the result has documents with evaluation four, next two with value of evaluation equals to three, and so on until the last two results with evaluation value equals to zero. To calculate the correlation coefficient, Pearson correlation (Kendall and Stuart, 1973) was used.

To consolidate the data obtained in the case study, the values obtained with the metrics were used to establish for each evaluation: (i) if the results were better in the original query or in the expanded queries, (ii) the impact of varying the number of terms used in the query expansions and (iii) the impact on the results according to the query expansion mode (automatic expansion or suggestion of terms).

The results are presented in percentages, representing the amount of times the result obtained in each tab was better than the other tabs.

Results

In the automatic expansion mode, in all metrics, the query expansion with general terms of the context and the query expansion with specific terms of the subjects showed better results than those obtained with the original query.

In the metric of full precision, an improvement of 76.47% of the cases in at least one of the query expansions was observed, while the original query result only showed better results in 11.76% of cases. The percentage improvement was observed in the comparison of results from the original query with the results of the expanded query with general terms of context (64.71%), and comparing results from the original query with the results of expanded queries with specific terms of subjects (64.71%).

In the metrics of search length and correlation rank, improvement in results were also observed, but in smaller proportions than the full precision.

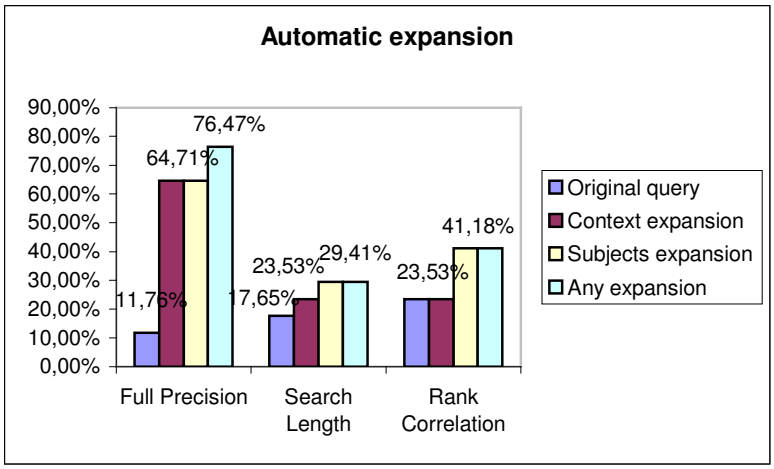

Figure 3. Graphic result of automatic expansion mode.

Unlike what happened with the automatic expansion, in the mode of suggestion of terms, all metrics showed that the query expansion presented worse results than those obtained with the original query. In the three considered metrics, the differences between the obtained percentages were big, reaching a difference of up to seven times in the case of metric search length (in 47.06% of the case the original query obtained better results against 5.88% of the expanded query).

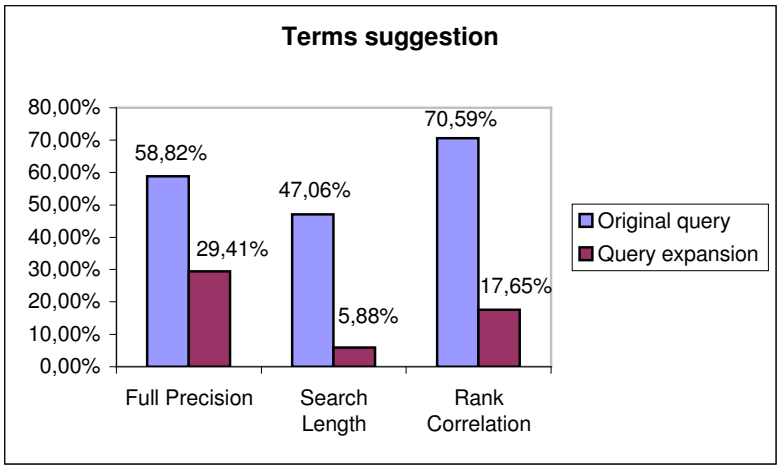

Figure 4. Graphic result of terms suggestion mode.

Justifications for the results

The hypothesis is that using the most frequent terms of the context to expand the queries done by users, would make the search results more useful. The general results obtained in the study case showed that the hypothesis is true for automatic expansion mode, but it is not so straightforward about the terms suggestion mode.

The results of the evaluations of relevance and the questionnaires were analyzed, providing some possible answers about the difference between the values obtained in the three metrics in the two modes.

Number of selected terms

In both modes the user decided how many terms would be included in the original query, but in different ways: while in the automatic expansion the user clearly influenced this amount because he was required to select the number of terms to be used; in the suggestion of terms this influence was indirect because possible terms were suggested and included by the user according to his judgment if the term would be relevant or not to meet his information need, without bothering to limit the number of terms that would be included in the query.

This difference significantly impacted the number of terms used in queries. While the users expanded the query with about 3.5 terms on average in the automatic expansion mode, in the terms suggestion mode this number was in average approximately 7.1 terms, more than double. The average size of the original queries did not change much in the different modalities. The automatic expansion mode presented in the average a size of approximately 2.7 terms and the terms suggestion mode, about 2.4.

Choosing keywords

The results showed that the selection of terms with the use of automatic rules based on the frequency of these terms in documents of context shows better results than the manual selection of terms, done by users, even if the list of terms suggestions also have been selected based on frequency. Works in interactive query expansion (IQE) show that users often find it difficult to select terms for query expansion from a suggested list (Joho, Sanderson and Beaulieu, 2004).

An indication of the difficulty of selecting good keywords is related to the poor results obtained with the terms suggestion mode can be proved by analyzing the different results obtained in the full precision metric with different profiles of users. The staff of sectorial departments, who know the domain information but are not experts in information cataloguing and search, considered 75% of the original queries better results while 25% of the results were better with the expanded queries.

However, the results were quite different considering only the librarians, information specialists, whose daily work involves capturing and registering information of the evaluated contexts. While the original queries performed better in 55.56% of cases, the expanded queries were better in 44.44% of cases.

Selecting general terms of context and specific terms of the different subjects

The third and final reason identified for the poor results in the terms suggestion mode was the possibility to combine terms from different subjects in a single query expansion. In the automatic expansion mode, different expansions were performed, one for each subject and one for the general terms of context.

Analyzing the results of the questionnaire, it was observed that the possibility of mixing the terms has been applied in practice. 51.72% of participants revealed that in the majority of queries, they used terms of the context and subjects at the same time to prepare the query. 34.48% of the participants did not mix the terms.

As in the search tools only documents containing all terms presented in the query are returned, the use of terms of different subjects, but all referring to the same context, could have caused the worse results.

Impact of the number of terms in the evaluation

Besides the general evaluation of the results, the influence of some variables in the relevance judgments made by participants was also evaluated. Some examples of these variables are the number of terms used in the expansion, the different profiles of users and the different contexts that were configured (Energy and Natural Environment).

The biggest noticed impact was the number of terms used to expand queries. In the automatic expansion mode, considering only the cases where the number of terms used for expansion was less than or equal to the size of the query made by the user, a great improvement in full precision metric was noticed. In none case with this configuration it was observed better results

in the original query. Considering only the expansion with general terms from context, in 75% of cases the obtained results were better than the original queries. Considering the query expansion with subject terms, in 87.50% of cases at least one of expansions obtained better results than the original queries. And finally, in all cases (100%) at least one of the expanded queries presented best results than the original ones (Fig. 5).

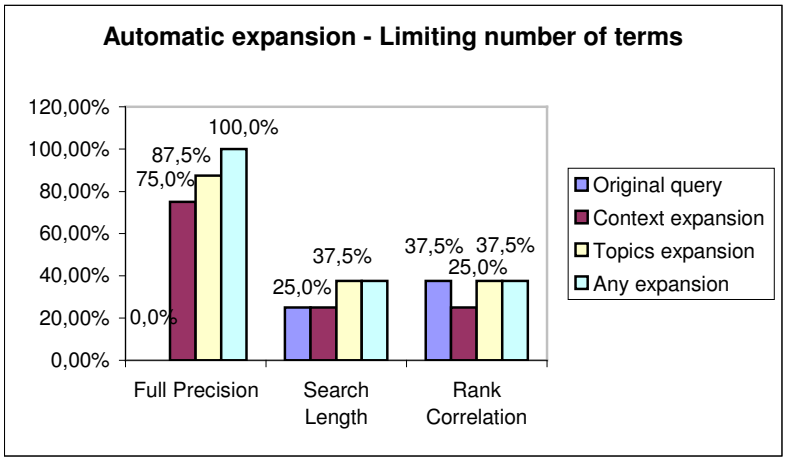

Fig. 5. Graphic result of automatic expansion mode with the number of terms of expansion limited to the number of query terms.

With these results it is possible to notice a greater accuracy when terms entered by the user have greater weight (or equal) to the automatically expanded terms. The context terms and the subject terms make the query to obtain more useful results, but these terms should not have in the final query greater importance than the terms defined by the users in the original query. This can be confirmed by analyzing only those cases where the size of the expansion, i.e., the number of terms used in the expansions, is greater than the size of the original query made by users. There was a worsening of the results, increasing the number of situations in which the original query's results are more accurate.

CONCLUSIONS AND FUTURE WORK

The use of information extraction activities in existing resources in databases, archives and information systems can be considered in order to make search results more contextualized and therefore more useful to users. The contextualization is done through the expansion of queries entered by users, adding the terms extracted from selected resources that are representative of the domain context of the information need.

A case study showed that the proposed architecture, in the automatic expansion mode, can be used in a corporate environment to improve the process of seeking information. The results obtained with the terms suggestion mode were worse than those obtained with queries made by users. The reasons for the differences in results obtained with the two modes were presented, but a future work is to do further experiments to confirm these justifications.

Other future works are: (i) analysis of the terms of the queries done by the users as a criterion for the terms expansion, for example, using as a criterion for calculating the weights of the terms the distance of each term with the terms informed in the query, in addition to frequency in the context information; (ii) identification of the subject more related with the search expression, allowing that a single expansion be made with terms of the identified subject, making it easier to view results; and (iii) application of the tool in different types of context, so it would be possible to make the search tools adaptable to the user experience context or professional activities to be performed, besides the domain context.

ACKNOWLEDGMENTS

The authors would like to thank the Knowledge Representation and Reasoning research group RCR/PPGI/UNIRIO. This work was partially supported by FAPERJ (through grants E-26/170028/2008 INC&T Program - Project: Brazilian Institute of Research on Web Science, and E-26/ 101.509/2010 - BBP/Bursary Representation and contextualized retrieval of learning content) and CNPQ (project: 557.128/2009-9, INCT on Web Science).

REFERENCES

1. Bhogal, J., Macfarlane, A., Smith, P. (2007) A review of ontology based query expansion, *Information Processing and Management*, 43 (4), July, 2007
2. Chanana, V., Ginige, A., Murugesan, S.: (2004) Improving information retrieval effectiveness by assigning context to documents. *International symposium on Information and communication technologies*
3. Chien, B.C., Hu, C.H., Ju, M.Y. (2007) Intelligent Information Retrieval Applying Automatic Constructed Fuzzy Ontology. *International Conference on Machine Learning and Cybernetics*
4. Dey, A. K., Abowd, G. D. (1999) Towards a better understanding of context and contextawareness. *International symposium on Handheld and Ubiquitous Computing*, pp. 304—307
5. Experian Hitwise Searches statistics (2010) http://www.hitwise.com/us/press-center/press-releases/google-searches-feb-10/
6. Hearst, M.A. (1997) TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, pp 33—64
7. Joho, H., Sanderson, M., Beaulieu, M. (2004) A study of user interaction with a concept-based interactive query expansion support tool, *European Conference on IR Research*, April 2004, Springer , pp. 42-56
8. Kendall, M.G., Stuart, A. (1973) The Advanced Theory of Statistics, v. 2: Inference and Relationship, *Griffin*
9. Manning, C.D., Raghavan, P., Schütze, H. (2008) Introduction to Information Retrieval, *Cambridge University Press*
10. Spink, A., Jansen, B. J. (2004) A study of Web search trends. Webology, 1(2), Article 4. Available at: http://www.webology.ir/2004/v1n2/a4.html
11. Tang, M.C., Sun Y. (2003) Evaluation of Web-Based Search Engines Using User-Effort Measures. *LIBRES Research Electronic Journal*, 13(2)
12. Voorhees, E.M. (2008) On test collections for adaptative information retrieval. *Information Processing and Management*. 44(6)
13. Yates, R. B., Neto, B. R. (1999) Modern Information Retrieval, *Addison Wesley*, 1a ed
14. Wang, C., Chang, G., Wang X., Ma, Y., Ma, H. (2009) A User Motivation Model for Web Search Engine. *Internation conference on Hybrid Intelligent Systems*