Invited Paper:

Using Thematic Contexts and Previous Solutions for Maintaining and Accessing Institutional Repositories

Pablo H. Delgado Ana G. Maguitman*
Departamento de Cs. e Ing. de la Computación Universidad Nacional del Sur
Avda Alem 1253, 8000 Bahía Blanca, Argentina

and

Victor M. Ferracutti Luis A. Herrera
Biblioteca Central de la Universidad Nacional del Sur
Avda Alem 1253, 8000 Bahía Blanca, Argentina

ABSTRACT

Institutional Repositories are collections of digital resources organized to facilitate their long term access. Two fundamental issues that need to be addressed at the moment of implementing these repositories are their maintenance and search instruments. This article describes an intelligent tool whose goal is to overcome certain general limitations encountered in current mechanisms for managing Institutional Repositories. The novel aspects of the proposed tool are the use of previous examples to generate suggestions for cataloging incoming digital resources and the ability to reflect a thematic context during the search process. The proposed tool applies case-based reasoning in combination with information retrieval techniques to take advantage of a large set of previously classified resources with the purpose of supporting the cataloger's task. The tool has been evaluated using a collection consisting of 9792 training examples from the SciELO electronic library and a test set containing 100 articles coming from different disciplines.

Keywords: Institutional Repositories, Context-Based Search, Case-Based Reasoning, Information Retrieval, Information Access, Library Cataloger Support Tools.

1. INTRODUCTION

Institutional Repositories are information management systems containing digital resource collections that represent the intellectual output of an institution. A key mission of an Institutional Library is to collect, organize, preserve, disseminate and facilitate the access to these resources. Each of these tasks presents challenging research and development opportunities.

Organizing resources relevant to diverse topics is a difficult and costly task for the cataloger, who is typically unfamiliar with the heterogeneous nature of the incoming resources. In the meantime, accessing these resources by the users will be ineffective unless the repository provides appropriate search mechanisms to facilitate the identification of potentially useful material. In light of these needs a variety of solutions based on information technologies have been proposed [4, 14, 23].

This article extends and evaluates intelligent software tools originally proposed in [7], developed to facilitate the processes of cataloging incoming digital resources and searching these repositories based on a thematic context. The task of cataloging involves associating a set of data with each digital resource that arrives at the Library. For example, a thesis is associated with an author, an advisor, a title, an abstract, one or more Dewey codes, one or more thematic areas, a number of descriptors, and a publication date. Some of these data, such as the title, author, advisor, abstract and date are explicitly given in the digital resource itself, while other data, such as the relevant Dewey codes, thematic areas and descriptors have to be inferred by the cataloger. The tool presented here has the capability of supporting the cataloger in the process of assigning the missing data to the incoming resources.

Most of the work dedicated to assist the end-users of Institutional Repositories focuses on "providing services such as quality indicators, the ability to browse through subject-based collections, the inclusion of supplementary materials, the provision of links to cited material and the ability to cross-search both internal and external repository collections." [11]. Our proposal, on the other hand, is motivated by several difficulties observed in the way end-users access Institutional Repositories. Typically, users express their information needs by filling in the fields in a search form (such as author or title) or by forming queries that combine a few keywords relevant to the topic of interest. However, unless there is a perfect match between the terms used in a query and those used to index the relevant re-

*To whom correspondence should be addressed. Email: agm@cs.uns.edu.ar.

sources, these resources will not be returned to the user. A problem with this approach is that end-users may not be familiar with the specific vocabulary employed to index a resource of interest. Because of this, much of the relevant material remains unnoticed by the user. Another common approach applied by users is to browse through the Institutional Repository. This can be done by applying different strategies, such as browsing by subject area, authors, collections, communities or most popular items. While this approach can help overcome the difficulty mentioned earlier, it has the disadvantage of being very time-consuming and distracting, as users commonly lose focus on the search task. The tool described in this work attempts to overcome these limitations by augmenting the traditional mechanisms for accessing Institutional Repositories with contextualized search methods. These search methods make use of one or several paragraphs representative of the topic of interest to trigger a thematic search process.

The proposed tools have been put into practice to manage information related to master and doctoral theses that are being archived at the Main Library of the Universidad Nacional del Sur, in Argentina. However, the applied methods are general enough to deal with other kinds of scientific publications.

The next section provides the background material, including the vector space model, context-based search and case-based reasoning. Section 3 describes the proposed tool by briefly discussing how cases are represented and the methods applied to generate suggestions. An evaluation of the performance of the tool is presented in section 4. Finally, the conclusions of our work are presented in section 5.

2. BACKGROUND

This section introduces a number of concepts in the area of classical information retrieval and intelligent search systems, which will be used along this work.

Vector Space Model

The vector space model [21] is an algebraic model used to represent documents and queries as vectors in n dimensions, where each dimension represents a term (word, concept or stem). Non-binary weights are assigned to the terms, typically by applying the weighting scheme known as TF-IDF (term frequency inverse document frequency). A high TF-IDF value of a term t in a document d is reached by a high frequency of t in d and a low document frequency of t in the whole repository of documents. Therefore, two factors know as TF (term frequency) and IDF (inverse document frequency) are combined to obtain an overall measure of the term importance in the document. The factor $\mathrm{TF}(t, d)$ is simply computed as the number of occurrences of t in d , while $\mathrm{IDF}(t)$ is computed as follows:

$$\mathrm{IDF}(t) = \frac{|R|}{|d_i \in R : t \in d_i|},$$

where R is the repository of documents. Finally, we compute the TF-IDF weight as $\mathrm{TF\text{-}IDF}(t, d) = \mathrm{TF}(t, d) \times \mathrm{IDF}(t)$.

The vector representations containing the resultant TF-IDF values are then used to compute similarity degrees between document pairs or between a document and a query. To measure similarity in term space, the cosine measure is typically used [2]. Given two vectors $\mathbf{d}_i$ and $\mathbf{d}_j$ in an n-dimensional term space, the cosine similarity $\sigma_{\cos}$ between them is their normalized dot product:

$$\sigma_{\cos}(\mathbf{d}_i, \mathbf{d}_j) = \frac{\mathbf{d}_i \cdot \mathbf{d}_j}{\|\mathbf{d}_i\|\|\mathbf{d}_j\|}.$$

We refer the reader to [21, 2] for details on the vector space model.

Context-Based Search

Context-based search systems monitor the user, infer their information needs and search in diverse repositories for relevant resources. Thematic contexts play a key role in task-based retrieval. Unfortunately, taking advantage of contextual information is a difficult task. This is due to the fact that current search systems impose a limit on query length. Even if queries of arbitrary length were allowed they would return very few or none result if user queries are assumed to be conjunctive queries.

A number of approaches have been adopted to tackle the problem of contextualized search. For example the system Watson [5] employs the information that a user is editing or accessing to automatically generate queries by applying diverse term extraction and weighting techniques. Watson filters the recovered results, groups those that are similar and presents them as suggestions to the user. Another seminal context-based search system is the Remembrance Agent [20]. This system can be integrated to the Emacs text editor and has the purpose of continuously monitoring the task of the user to discover relevant text documents, notes and emails previously indexed. The EXTENDER system [18, 19] applies incremental search to build an enriched description of the user context. The goal of this system is to generate summaries of new topics that are relevant to a knowledge model under construction. Besides the systems mentioned above, there are several

other proposals aimed at solving the problem of context-based search [1, 15, 3, 17, 12, 16, 6].

Case-Based Reasoning

Case-Based Reasoning (CBR) [13] is a paradigm to build intelligent systems where the main sources of knowledge are not rules but cases or episodes. These systems generate solutions by recovering relevant stored cases and adapting them to new situations. The CBR paradigm is based on two premises about the nature of the world. The first premise is that the world is regular, and because of this regularity, solutions that were useful for previous problems can serve as a starting point to solve new problems. The second premise establishes that the kind of problems that an agent finds tend to be recurrent and therefore new problems may be similar to problems found in the past. CBR systems are bases on these premises to store, adapt and reuse solutions to previous problems.

When a new case is recovered and used at the right time, it becomes an important source of information, avoiding the time and effort necessary to develop solutions from scratch. CBR systems have been successfully applied to areas such as design, planning, diagnosis, knowledge management and legal reasoning. Some CBR systems operate autonomously, while others are integral parts of collaborative systems, where the user and the system complement and help each other with the purpose of solving problems.

3. THE PROPOSED TOOL

This section describes the data and methods used to generate suggestions in the Institutional Repository.

Data and its Representation

An important task of the proposed tool is to maintain a sufficiently large repository of cases, which will be used to generate suggestions about possible thematics, Dewey codes and descriptors. To collect such cases, we used SciELO [22], an online library containing thousands of scientific articles about different subjects. For our analysis, the language was restricted to Spanish, and therefore, only the Argentinean, Chilean and Mexican SciELO portals were used. An auxiliary tool was developed to automatically collect 9792 articles about 47 different thematics from the SciELO portals. The thematics were assigned to each article based on the journal in which the article was published.

As a starting point, the system indexes all the digital resources that are used to generate suggestions. Each indexed resource is considered a case and is stored using an XML file admitting fields such type (thesis, article, book, etc.), author, title, subtitle, advisor, abstract, content, date, thematic, and a variable number of Dewey codes and descriptors.

Methods

Consider a digital resource that arrives at an Institutional Repository and needs to be cataloged. Assuming that part of the data of the incoming resource, such as title and abstract, is available, the system will create a thematic context using the terms from these data. The cataloger will then be able to request suggestions to fill in the missing fields associated with the resource, such as thematics, Dewey codes and descriptors. These suggestions are extracted from cases, which consist of digital resources which have been categorized in the past.

The method applied to generate suggestions takes some terms from the thematic context and forms queries, which are submitted to the repository of cases. In order to select the terms to form each query we implemented a roulette selection mechanism where the probability of choosing a particular term to form a query is proportional to the number of times the term appears in the thematic context. Roulette selection is a technique typically used by Genetic Algorithms [9] to choose potentially useful solutions for recombination, where the fitness level is used to associate a probability of selection. This approach results in a non-deterministic exploration of term space that favors the fittest terms. A number of queries are then submitted and a set of documents similar to the thematic context are selected. The cosine similarity measure described earlier is used for this purpose. In order to determine which documents are selected for further analysis, a hypercone is computed with axis corresponding to the thematic context vector and angle defined in terms of the given similarity threshold. The set of documents whose vector representations fall inside this hypercone are retrieved as potentially useful cases.

The selected cases will typically contain one or more associated Dewey codes, thematics, and descriptors. These cases are integrated to generate the corresponding suggestion lists. In addition, whenever the resource content is available in the form of text, an analysis of this content is performed in order to select the terms with the highest TF-IDF values. These terms are added to the suggestion list of potentially useful descriptors.

This mechanism for generating suggestions is based on the CBR methodology described earlier, where previous cases serve as starting point to provide solutions to new situations. In this scenario, digital resources that have already been

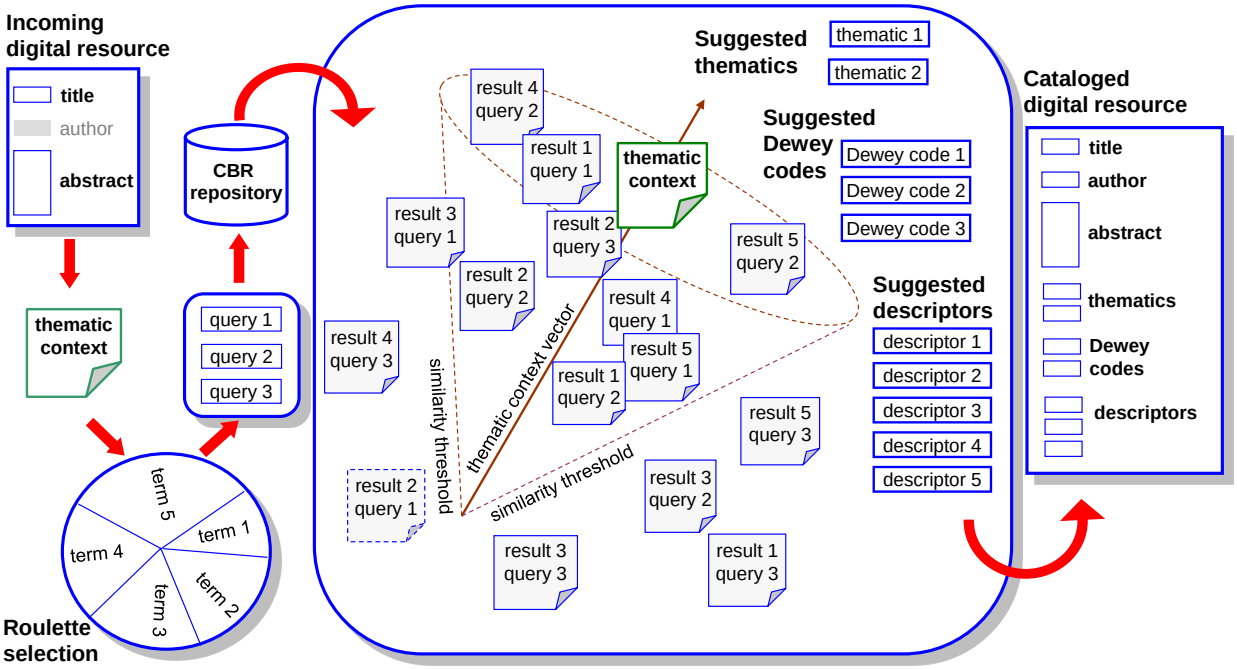


Figure 1: The process of generating suggestions.

cataloged offer tentative solutions to the problem of determining the thematics, Dewey codes and descriptors of new resources. Figure 1 shows the process of generating suggestions.

It is worth mentioning that the greater the number of submitted queries, the higher is the probability of generating significant queries and therefore the better are the chances of obtaining cases that are relevant to the thematic context at hand. The number of queries automatically generated by the system, the number of terms used to form each query, the similarity threshold used in the search process, and the number of suggestions that the system will return to the user have an important impact on the precision, coverage and speed of the suggestion generation process, and are configurable by the user.

A similar approach is adopted to support the user that is seeking for material relevant to a thematic context. In this case, the user will enter a description of his or her information needs consisting of a text document or a relevant paragraph. Differently from traditional mechanisms where information needs are typically specified using a small number of terms, our proposal allows using an extended description to specify the user's information needs. Queries are generated from the user's thematic context through the roulette selection mechanism as described earlier. Again, results are selected based on a hypercone with axis and angle determined by the thematic context vector and the similarity threshold, respectively. Finally, the results, consisting of pointers to potentially relevant resources, are presented to the user. These results are ranked in terms of their similarity to the user's thematic context. Classical search by author, title, or a few keywords is also supported by the tool.

4. EVALUATION

The performance evaluation of the proposed tool was conducted using 9792 articles from the SciELO repository, covering 47 different thematics, and 100 test articles from SciELO that were not indexed in our repository. The test articles were collected from journals covering 10 different thematics, being 10 the number of test articles for each thematic. The selected thematics used for the test set as well as the total number of indexed articles associated with each of these thematics are shown in the following table.

Thematic	Number of articles
Anthropology & Archeology	279
Biotechnology	434
Botanics	381
Chemistry	337
Economy	164
Engineering	275
Geology	573
Medicine	2015
Philosophy	224
Political Science	173

We used the abstract of each test article to initiate the search for suggestions. Each automatically generated query was formed by 3 terms selected using the roulette mechanism described earlier. Four different settings consisting of 1, 10, 50 and 100 queries generated by the system were analyzed. The similarity threshold used for document selection was set to 0.2. Finally, for each setting we looked at whether the system correctly predicted the thematic of each test article by looking at the top 3 and top 5 suggestions.

Because the roulette technique used to generate suggestions is nondeterministic, we run the experiment five times for each of the 100 test articles and computed the average performance for each article. The next table summarizes the statistics (means and 95% confidence intervals) resulting from analyzing the performance of the tool.

	Mean	**95% C.I.**
1 query		
among top 3	0.40	[0.34 , 0.46]
among top 5	0.48	[0.42 , 0.54]
10 queries		
among top 3	0.57	[0.50 , 0.64]
among top 5	0.63	[0.56 , 0.70]
50 queries		
among top 3	0.63	[0.55 , 0.70]
among top 5	0.67	[0.60 , 0.75]
100 queries		
among top 3	0.60	[0.52 , 0.68]
among top 5	0.65	[0.57 , 0.72]

In order to dig deeper into the behavior of the tool, we computed the average performance by thematic. This is reported in the chart and table of figure 2, where the analyzed thematics are sorted based on their overall performance for better visualization (based on 100 automatically generated queries).

We observe that for most of the analyzed thematics the performance of the tool improves as we increase the number of submitted queries. However, 10 queries seem to be sufficient to attain results virtually as good as those obtained with a larger number of queries. This indicates that it is unnecessary to submit an extremely large number of queries since it is possible to achieve a close to the peak performance reasonably fast.

These results also reveal a great variation in performance among thematics. While the Biotechnology thematic seems to be very hard to identify, other thematics such as Medicine or Geology are easily recognized. This is probably due the fact that the associated vocabulary has more discriminating power for some thematics than for other.

5. CONCLUSIONS

This article described new tools that can be used to overcome certain limitations found in mechanisms currently employed for maintaining and accessing Institutional Repositories. The proposed tools are based on information retrieval and artificial intelligence techniques. In particular, context-based search is combined with case-based reasoning to define powerful methods for information access and suggestion generation based on the description of a topic of interest and a collection of previous solutions.

The suggestion generation tool was evaluated using material from the SciELO repository. We have observed that, although the methods showed a poor performance for a few cases, the general performance is very good for most of the analyzed thematics.

As part of our future work we expect to improve the suggestion generation mechanism by using semantic information sources such as subject classification labels, ontologies and thesaurus containing controlled vocabulary. Another possible future direction consists in using other techniques for reflecting thematic contexts, such as those described in [10, 19, 16, 6]. We also plan to conduct user studies to monitor the way catalogers and end-users interact with the tool and to identify areas of improvement.

Although the tool is being used and has been evaluated for the Spanish language, it is possible to adapt it to other languages by simply using a collection of articles in that language to provide suggestions and by applying the appropriate stopword lists.

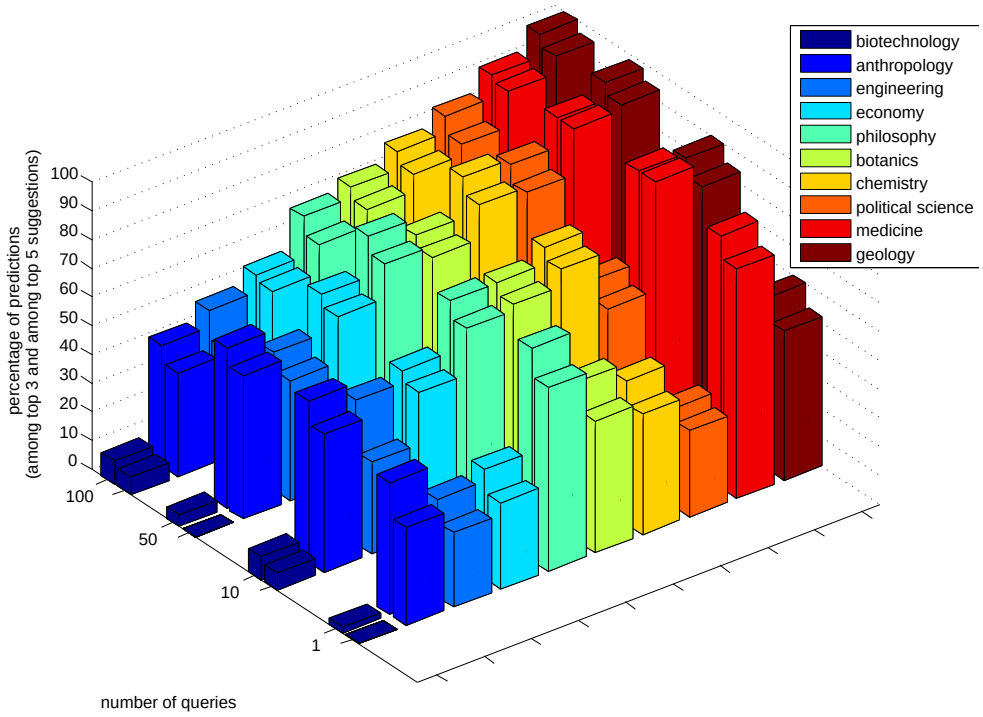
Acknowledgement

We thank Mariana Anahí Varela, María Alejandra Dini and Verónica Antúnez for developing the first prototype of the system described in this article. We also thank the Library support Staff at Universidad Nacional del Sur, in particular Paola Cruciani, Fernando A. Martínez, Natalia Mitzig, Ricardo A. Piriz and Jerónimo Spadaccioli from the System Area and Nélida Benavente, Guillermina Castellano and Marta Ibarlucea whose useful insights helped improve the quality of this research work.

This research work is partially supported by CONICET (PIP 11220090100863) and Universidad Nacional del Sur (PGI 24/ZN13).

References

[1] Armstrong, R., Freitag, D., Joachims, T., & Mitchell, T. (1995). WebWatcher: A learning apprentice for the World Wide Web. In



	Biotechnology	Anthropology & Archeology	Engineering	Economy	Philosophy	Botanics	Chemistry	Political Science	Medicine	Geology
1 query										
among top 3	0	0.34	0.26	0.30	0.64	0.46	0.42	0.30	0.80	0.52
among top 5	0.02	0.46	0.34	0.38	0.74	0.56	0.50	0.34	0.88	0.60
10 queries										
among top 3	0.06	0.48	0.32	0.50	0.66	0.68	0.74	0.54	0.92	0.84
among top 5	0.08	0.56	0.50	0.54	0.72	0.72	0.78	0.60	0.92	0.90
50 queries										
among top 3	0	0.50	0.42	0.58	0.70	0.66	0.78	0.76	0.92	0.94
among top 5	0.04	0.56	0.48	0.62	0.76	0.70	0.84	0.82	0.92	0.98
100 queries										
among top 3	0.06	0.36	0.38	0.52	0.62	0.68	0.74	0.78	0.90	0.96
among top 5	0.08	0.42	0.48	0.54	0.68	0.72	0.78	0.84	0.92	1

Figure 2: Performance by thematic measured as the average percentage of correct predictions.

AAAI spring symposium on information gathering (pp. 6-12).

[2] Baeza-Yates R. & Ribeiro-Neto B. (2010). Modern Information Retrieval. Second Edition. Addison-Wesley.

[3] Baldonado, M. Q. W., & Winograd, T. (1997). SenseMaker: an information-exploration interface supporting the contextual evolution of a user's interests. In Proceedings of the SIGCHI conference on human factors in computing systems (pp. 11-18). ACM Press.

[4] Buckland, M. (1992). Redesigning Library Services, A Manifesto. American Library Association, Chicago.

[5] Budzik, J., Hammond K. J., & Birnbaum L.(2001). Information Access in Context. Knowledge Based Systems, 14(1-2):37-53, Elsevier.

[6] Cecchini, R., Lorenzetti, C., Maguitman, A. & Brignole, N. B. (2010). Multiobjective Evolutionary Algorithms for Context-based Search. Journal of the American Society for Information Science and Technology, 61(6):1258-1274, John Wiley & Sons, Inc.

[7] Dini, M. A., Varela, M. A., Antúnez, V., Maguitman, A. & Herrera, L. (2010). Soporte Inteligente para el Mantenimiento y Acceso Contextualizado a Repositorios Institucionales. 8^a Jornada sobre la Biblioteca Digital Universitaria.

[8] Gospodnetic, O., Hatcher, E. & McCandless M. (2009). Lucene in Action (2nd ed.). Manning Publications.

[9] Holland, J. H. (1975). Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor.

[10] Hulth, A., Karlgren, J., Jonsson, A., Boström, H. & Asker, L. (2001). Automatic Keyword Extraction Using Domain Knowledge. Computational Linguistics and Intelligent Text Processing, 2004/2001:472-482. Springer Verlag, Berlin.

[11] Jean, B. S., Rieh, S. Y., Yakel E., & Markey K. (2011). Unheard Voices: Institutional Repository End-Users. College & Research Libraries.

[12] Kraft, R., Chang, C. C., Maghoul, F., & Kumar, R. (2006). Searching with Context. In Proceedings of the 15th International Conference on the World Wide Web, pages 477-486, New York, NY, USA. ACM Press.

[13] Leake, D. (1996). CBR in Context: The Present and Future, In Leake, D., editor, Case-Based Reasoning: Experiences, Lessons, and Future Directions. 1-30, AAAI Press/MIT Press.

[14] Levy, D. M. & Marshall, C. C. (1995). Going digital: a look at assumptions underlying digital libraries. Communications of the ACM. 38(4): 77-84. New York, NY, USA. ACM Press.

[15] Lieberman, H. (1995). Letizia: An agent that assists Web browsing. In C. S. Mellish (Ed.), Proceedings IJCAI (pp. 924-929). Montreal, Quebec, Canada : Morgan Kaufmann publishers Inc.: San Mateo, CA, USA.

[16] Lorenzetti, C. M., & Maguitman, A. G. (2009). A Semi-Supervised Incremental Algorithm to Automatically Formulate Topical Queries. Information Sciences, 179(12):1881-1892. Special Issue on Web Search.

[17] Maglio, P. P., Barrett, R., Campbell, C. S., & Selker, T. (2000). SUITOR: an attentive information system. In Proceedings of the 5th international conference on intelligent user interfaces (pp. 169-176). ACM Press.

[18] Maguitman, A., Leake, D., Reichherzer, T., & Menczer, F. (2004). Dynamic extraction of topic descriptors and discriminators: Towards automatic context-based topic search. In Proceedings of CIKM (pp. 463-472). Washington, DC : ACM Press.

[19] Maguitman, A., Leake, D., & Reichherzer, T. (2005). Suggesting novel but related topics: towards context-based support for knowledge model extension. In Proceedings of IUI (pp. 207-214). New York, NY, USA : ACM Press.

[20] Rhodes, B. & Starner, T. (1996). The remembrance agent: A continuously running automated information retrieval system. In Proceedings of PAAM (pp. 487-495).

[21] Salton, G., Wong, A & Yang C. S. (1975). A Vector Space Model for Automatic Indexing, Communications of the ACM, 18 (11): 613-620, ACM Press.

[22] http://www.scielo.org.ar/. Accessed on August 7 2011.

[23] Sølvberg, I. T. (2001). Digital libraries and information retrieval, Lectures on information retrieval, Springer-Verlag New York, Inc., New York, NY.