Tuning Topical Queries through Context Vocabulary Enrichment: A Corpus-Based Approach*

Carlos M. Lorenzetti Ana G. Maguitman

Grupo de Investigación en Recuperación de Información y Gestión del Conocimiento
LIDIA - Laboratorio de Investigación y Desarrollo en Inteligencia Artificial
Departamento de Ciencias e Ingeniería de la Computación
Universidad Nacional del Sur, Av. Alem 1253, (B8000CPB) Bahía Blanca, Argentina
CONICET - Consejo Nacional de Investigaciones Científicas y Técnicas
phone: 54-291-4595135 fax: 54-291-4595136
e-mail: {cml,agm}@cs.uns.edu.ar

Abstract. Context-based Web search has become an important research area and many strategies have been proposed to reflect contextual information in search queries. Despite the success of some of these proposals they still have serious limitations due to their inability to bridge the terminology gap existing between the user context description and the relevant documents' vocabulary. This paper presents a quantitative technique to learn vocabularies useful for describing the theme of a context under analysis. The enriched vocabulary allows the formulation of search queries to identify resources with higher precision than those identified using the initial vocabulary. Rigorous experimentation leads us to conclude that the proposed technique is superior to a baseline and other well-known query reformulation techniques.

1 Introduction

Context-based search is the process of seeking information related to a user's thematic context [5,11,8,15]. Meaningful automatic context-based search can only be achieved if the semantics of the terms in the context under analysis is reflected in the search queries. From a pragmatic perspective, terms acquire meaning from the way they are used and from their co-occurrence with other terms. Therefore, mining large corpora (such as the World Wide Web) guided by the user's context can help uncover the meaning of a user's information request.

An information request is usually initiated or generated within a task. For example, if the user is editing or reading a document on a specific topic, he may be willing to explore new material related to that topic. Topical queries can be formed using small sets of terms from the user's context. The implementation of a mechanism for the automatic generation of queries from a thematic context raises several questions: (1) Which terms will be more helpful to access relevant material? (2) How many terms should be used to form each query? (3) Is the current context vocabulary good enough to access the right information?

* This research work is supported by Agencia Nacional de Promoción Científica y Tecnológica (PICT 2005 Nro. 32373) and Universidad Nacional del Sur (PGI 24/ZN13).

Attempting to find the best subsets of terms to create appropriate queries is a combinatorial problem. The situation worsens when we deal with an open search space, i.e., when other terms that are not part of the current context vocabulary can be part of the queries. Willing to use terms that are not part of the current context is not an atypical situation when attempting to tune queries based on a small context description and a large external corpus. We can think of this query tuning process as a by-product of learning a better vocabulary to characterize the topic under analysis and the user's information needs.

The contribution of this work is a method for guiding the incremental exploration of new vocabularies with the purpose of tuning queries. The goal for the queries is to reflect contextual information and to effectively retrieve semantically related material when posed to a search interface.

2 Background

Query tuning is usually achieved by replacing or extending the terms of a query, or by adjusting the weights of a query vector. Relevance feedback is a query refinement mechanism used to tune queries based on the relevance assessments of the query's results. A driving hypothesis for relevance feedback methods is that it may be difficult to formulate a good query when the collection of documents is not known in advance, but it is easy to judge particular documents, and so it makes sense to engage in an iterative query refinement process. A typical relevance feedback scenario will involve the following steps:

Step 1: A query is formulated.
Step 2: The system returns an initial set of results.
Step 3: A relevance assessment on the returned results is issued (relevance feedback).
Step 4: The system computes a better representation of the information needs based on this feedback.
Step 5: The system returns a revised set of results.

Depending on the level of automation of step 3 we can distinguish three forms of feedback:

- **Supervised Feedback**: requires explicit feedback, which is typically obtained from users who indicate the relevance of each of the retrieved documents.
- **Unsupervised Feedback**: it applies blind relevance feedback, and typically assumes that the top k documents returned by a search process are relevant.
- **Semi-supervised Feedback**: the relevance of a document is inferred by the system. A common approach is to monitor the user behavior (e.g., documents selected for viewing or time spent viewing a document). Provided that the information seeking process is performed within a thematic context, another automatic way to infer the relevance of a document is by computing the similarity of the document to the user's current context.

The best-known algorithm for relevance feedback has been proposed by Rocchio [17]. Given an initial query vector $\overrightarrow{q}$ a modified query $\overrightarrow{q_m}$ is computed as follows:

$$\overrightarrow{q_m} = \alpha \overrightarrow{q} + \beta \sum_{\overrightarrow{d_j} \in D_r} \overrightarrow{d_j} - \gamma \sum_{\overrightarrow{d_j} \in D_n} \overrightarrow{d_j}.$$

where D_r and D_n are the sets of relevant and non-relevant documents respectively and α, β and γ are tuning parameters. A common strategy is to set α and β to a value greater than 0 and γ to 0, which yields a positive feedback strategy. When user relevance judgments are unavailable, the set D_r is initialized with the top k retrieved documents and D_n is set to $\emptyset$. This yields an unsupervised relevance feedback method.

Several successors of the Rocchio's method have been proposed with varying success. One of them is selective query expansion [2], which monitors the evolution of the retrieved material and is disabled if query expansion appears to have a negative impact on the retrieval performance. Other successors of the Rocchio's method use an external collection different from the target collection to identify good terms for query expansion. The refined query is then used to retrieve the final set of documents from the target collection [9]. A successful generalization of the Rocchio's method is the Divergence from Randomness mechanism with Bose-Einstein statistics (Bo1-DFR) [1]. To apply this model, we first need to assign weights to terms based on their informativeness. This is estimated by the divergence of its distribution in the top-ranked documents from a random distribution as follows:

$$w(t) = tf_x.log_2 \frac{1 + P_n}{P_n} + log_2(1 + P_n)$$

where tf_x is the frequency of the query term in the top-ranked documents and P_n is the proportion of documents in the collection that contain t. Finally, the query is expanded by merging the most informative terms with the original query terms.

The main problem of the above query tuning methods is that their effectiveness is correlated with the quality of the top ranked documents returned by the first-pass retrieval. On the other hand, if a thematic context is available, the query refinement process can be guided by computing an estimation of the quality of the retrieved documents. This estimation can be used to predict which terms can help refine subsequent queries.

During the last years several techniques that formulate queries from the user context have been proposed [5,8]. Limited work, however, has been done on semi-supervised methods that simultaneously take advantage of the user context and results returned from a corpus to refine queries. Next section presents our proposal to tune topical queries based on the analysis of the terms found in the user context and in the incrementally retrieved results.

3 A Semi-Supervised Method for Query Reformulation

Much work has addressed the problem of computing the informativeness of a term across a corpus (e.g., [1,16]) and a good deal of research has focused on computing the

descriptive and discriminating power of a term in a document with respect to a corpus (e.g., [18]). All this work, however, has been done based on a predefined collection of documents and independently from a thematic context. In [12] we proposed to study the descriptive and discriminating power of a term based on its distribution across the topics of pages returned by a search engine. In that proposal the search space is the full Web and the analysis of the descriptive or discriminating power of a term is limited to a small collection of documents—incremental retrievals—that is built up over time and changes dynamically. Unlike traditional information retrieval schemes, which analyze a predefined collection of documents and search that collection, our methods use limited information to assess the importance of terms and documents as well as to manage decisions about which terms to retain for further analysis, which ones to discard, and which additional queries to generate.

To distinguish between topic descriptors and discriminators we argue that *good topic descriptors* can be found by looking for terms that occur often in documents related to the given topic. On the other hand, *good topic discriminators* can be found by looking for terms that occur only in documents related to the given topic. Both topic descriptors and discriminators are important as query terms. Because topic descriptors occur often in relevant pages, using them as query terms may improve recall. Similarly, good topic discriminators occur primarily in relevant pages, and therefore using them as query terms may improve precision.

3.1 Computing Descriptive and Discriminating Power

As a first approximation to compute descriptive and discriminating power, we begin with a collection of m documents and n terms. As a starting point we build an $m \times n$ matrix $\mathbf{H}$, such that $\mathbf{H}[i,j] = k$ if k is the number of occurrences of term t_j in document d_i. In particular we can assume that one of the documents (e.g., d_0) corresponds to the initial user context.

The matrix $\mathbf{H}$ allows us to formalize the notions of good descriptors and good discriminators. We define *descriptive power of a term in a document* as a function $\lambda : \{d_0, \ldots, d_{m-1}\} \times \{t_0, \ldots, t_{n-1}\} \rightarrow [0,1]$:

$$\lambda(d_i, t_j) = \frac{\mathbf{H}[i,j]}{\sqrt{\sum_{k=0}^{n-1} (\mathbf{H}[i,k])^2}}.$$

Note that λ can be regarded as a version of matrix $\mathbf{H}$ normalized by row (i.e, by document).

If we adopt $\mathrm{s}(k) = 1$ whenever $k > 0$ and $\mathrm{s}(k) = 0$ otherwise, we can define the *discriminating power of a term in a document* as a function $\delta : \{t_0, \ldots, t_{n-1}\} \times \{d_0, \ldots, d_{m-1}\} \rightarrow [0,1]$:

$$\delta(t_i, d_j) = \frac{\mathrm{s}(\mathbf{H}[j,i])}{\sqrt{\sum_{k=0}^{m-1} \mathrm{s}(\mathbf{H}[k,i])}}.$$

In this case δ can be regarded as a transposed version of matrix $\mathbf{H}$ normalized by column (i.e, by term).

Our current goal is to learn a better characterization of the user needs. Therefore rather than extracting descriptors and discriminators directly from the user context, we want to extract them from *the topic* of the user context. This requires an incremental method to characterize the topic of the user context, which is done by identifying documents that are similar to the user current context. Assume the user context and the retrieved documents are represented as document vectors in term space. To determine how similar two documents d_i and d_j are, we adopt the IR cosine similarity [3]. This measure is defined as a function $\sigma : \{d_0, \ldots, d_{m-1}\} \times \{d_0, \ldots, d_{m-1}\} \rightarrow [0, 1]$:

$$\sigma(d_i, d_j) = \sum_{k=0}^{n-1} [\lambda(d_i, t_k) \cdot \lambda(d_j, t_k)].$$

We formally define the *term descriptive power in the topic of a document* as a function $\mathit{\Lambda} : \{d_0, \ldots, d_{m-1}\} \times \{t_0, \ldots, t_{n-1}\} \rightarrow [0, 1]$. We set $\mathit{\Lambda}(d_i, t_j) = 0$ if $\sum_{\substack{k=0 \\ k \neq i}}^{m-1} \sigma(d_i, d_k) = 0$. Otherwise we define $\mathit{\Lambda}(d_i, t_j)$ as follows:

$$\mathit{\Lambda}(d_i, t_j) = \frac{\sum_{\substack{k=0 \\ k \neq i}}^{m-1} [\sigma(d_i, d_k) \cdot [\lambda(d_k, t_j)]^2]}{\sum_{\substack{k=0 \\ k \neq i}}^{m-1} \sigma(d_i, d_k)}.$$

Thus, the descriptive power of a term t_j in the topic of a document d_i is a measure of the quality of t_j as a descriptor of documents similar to d_i.

Analogously, we define the *discriminating power of a term in the topic of a document* as a function $\mathit{\Delta} : \{t_0, \ldots, t_{n-1}\} \times \{d_0, \ldots, d_{m-1}\} \rightarrow [0, 1]$ calculated as follows:

$$\mathit{\Delta}(t_i, d_j) = \sum_{\substack{k=0 \\ k \neq j}}^{m-1} [[\delta(t_i, d_k)]^2 \cdot \sigma(d_k, d_j)].$$

Thus the discriminating power of term t_i in the topic of document d_j is an average of the similarity of d_j to other documents discriminated by t_i. For a worked example showing the results of computing topic descriptors and discriminators see [10].

3.2 An Incremental Mechanism to Tune Topical Queries

Our proposal is to approximate the terms' descriptive and discriminating power for the thematic context under analysis with the purpose of generating good queries.

Our approach adapts the typical relevance feedback mechanism to account for a thematic context $\mathcal{C}$ as follows:

Step 1: A query is formulated based on $\mathcal{C}$.
Step 2: The system returns an initial set of results.
Step 3: Repeat for at least v iterations or until no improvements are registered
- **Step 3.1:** A relevance assessment on the returned results is issued based on $\mathcal{C}$.
- **Step 3.2:** After a certain number of trials and depending on the relevance assessments, the system computes a better representation of the thematic context (phase change).
- **Step 3.3:** The system formulates new queries and returns a revised set of results.

In order to learn better characterizations of the thematic context, the system undergoes a series of phases. At the end of each phase, the context characterization is updated with new learned material. Each phase evolves through a sequence of trials, where each trial consists in the formulation of a set of queries, the analysis of the retrieved results, the adjustment of the terms' weights, and the discovery of new potentially useful terms. For a given phase $\mathcal{P}_i$, the context is represented by a set of weighted terms. Let $w^{\mathcal{P}_i}(t, \mathcal{C})$ be an estimation of the importance of term t in context $\mathcal{C}$ during phase i. If t occurs in the initial context, then the value $w^{\mathcal{P}_0}(t, \mathcal{C})$ is initialized as the normalized frequency of term t in $\mathcal{C}$, while the weight of those terms that are not part of $\mathcal{C}$ are assumed to be 0.

Let $w_{\Lambda}^{(i,j)}(t, \mathcal{C})$ and $w_{\Delta}^{(i,j)}(t, \mathcal{C})$ be an estimation of the descriptive and discriminating power of term t for context $\mathcal{C}$ at trial j of phase i. These values are incrementally computed as follows:

$$w_{\Lambda}^{(i,j+1)}(t, \mathcal{C}) = \alpha.w_{\Lambda}^{(i,j)}(t, \mathcal{C}) + \beta.\Lambda^{(i,j)}(t, \mathcal{C}).$$

$$w_{\Delta}^{(i,j+1)}(t, \mathcal{C}) = \alpha.w_{\Delta}^{(i,j)}(t, \mathcal{C}) + \beta.\Delta^{(i,j)}(t, \mathcal{C}).$$

We assume $w_{\Lambda}^{(i,0)}(t, \mathcal{C}) = w_{\Delta}^{(i,0)}(t, \mathcal{C}) = 0$ and use the results returned during each trial j to compute $\Lambda^{(i,j)}(t, \mathcal{C})$ and $\Delta^{(i,j)}(t, \mathcal{C})$, the descriptive and discriminating power of term t for the topic of $\mathcal{C}$. To form queries during phase i we implemented a roulette selection mechanisms where the probability of choosing a particular term t to form a query is proportional to $w^{\mathcal{P}_i}(t, \mathcal{C})$. Roulette selection is a technique typically used by Genetic Algorithms [7] to choose potentially useful solutions for recombination, where the fitness level is used to associate a probability of selection. This approach resulted in a non-deterministic exploration of term space that favored the fittest terms.

The system monitors the effectiveness achieved at each iteration. In our approach we use *novelty-driven similarity* introduced in section 4 as an estimation of the retrieval effectiveness. If after a window of u trials the retrieval effectiveness has not crossed a given threshold μ (i.e., no significant improvements are observed after certain number of trials), the system forces a phase change to explore new potentially useful regions of the vocabulary landscape. A phase change can be regarded as a vocabulary leap, which can be thought of as a significant transformation (typically an improvement) of the context characterization. If a phase change takes effect during trial j, the value of $w_{\Lambda}^{\mathcal{P}_i}(t, \mathcal{C})$ is set to $w_{\Lambda}^{(i,j)}(t, \mathcal{C})$ and $w_{\Delta}^{\mathcal{P}_i}(t, \mathcal{C})$ is set to $w_{\Delta}^{(i,j)}(t, \mathcal{C})$. To reflect the phase change in the new characterization of the thematic context, the weight of each term t is updated as follows:

$$w^{\mathcal{P}_{i+1}}(t, \mathcal{C}) = \gamma.w^{\mathcal{P}_i}(t, \mathcal{C}) + \zeta.w_{\Lambda}^{\mathcal{P}_i}(t, \mathcal{C}) + \xi.w_{\Delta}^{\mathcal{P}_i}(t, \mathcal{C}).$$

These weights are then used to generate new queries during the sequence of trials at phase $i + 1$.

4 Evaluation

The goal of this section is to compare the proposed method against two other methods. The first is a baseline that submits queries directly from the thematic context and does

not apply any refinement mechanism. The second method used for comparison is the Bo1-DFR described in section 2.

To perform our tests we used 448 topics from the Open Directory Project (ODP)[1]. The topics were selected from the third level of the ODP hierarchy. A number of constraints were imposed on this selection with the purpose of ensuring the quality of our test set. The minimum size for each selected topic was 100 URLs and the language was restricted to English. For each topic we collected all of its URLs as well as those in its subtopics. The total number of collected pages was more than 350K. The Terrier framework [14] was used to index these pages and to run our experiments.

In our tests we used the ODP description of each selected topic to create an initial context description $\mathcal{C}$. The proposed algorithm was run for each topic for at least $v = 100$ iterations, with 10 queries per iteration and retrieving 10 results per queries.

The descriptor and discriminator lists at each iteration were limited to up to 100 terms each. The other parameters in our algorithm were set as follows: $u = 10$, α=0.5, β=0.5, γ=0.33, ζ=0.33, ξ=0.33, μ=0.2. In addition, we used the stopword list provided by Terrier, Porter stemming was performed on all terms and none of the query expansion methods offered by Terrier was applied.

In order to compare the implemented methods we used three measures of query performance:

- Novelty-driven similarity: this measure of similarity is based on σ but disregards the terms that form the query, overcoming the bias introduced by those terms and favoring the exploration of new material. Given a query q and documents d_i and d_j, the novelty-driven similarity measure is defined as $\sigma^N(\mathbf{q}, d_i, d_j) = \sigma(d_i - \mathbf{q}, d_j - \mathbf{q})$. The notation $d_i - \mathbf{q}$ stands for the representation of the document d_i with all the values corresponding to the terms from query $\mathbf{q}$ set to zero. The same applies to $d_j - \mathbf{q}$.
- Precision: this performance evaluation measures the fraction of retrieved documents which are known to be relevant, i.e., Precision$= |A \cap R|/|A|$, where R and A are the relevant and answer set respectively. The relevant set for each analyzed topic was set as the collection of its URLs as well as those in its subtopics.
- Semantic Precision: other topics in the ontology could be semantically similar (and therefore partially relevant) to the topic of the given context. Therefore, we propose a measure of semantic precision defined as Precision$^S = \sum_{p \in A} \sigma^S(t(\mathcal{C}), t(p))/|A|$, where $t(\mathcal{C})$ is the ODP topic associated with the description used as the initial context, $t(p)$ is the topic of page p and $\sigma^S(t(\mathcal{C}), t(p))$ is the semantic similarity between these two topics. To compute σ^S we used a semantic similarity measure for generalized ontologies proposed by Maguitman et al. [13].

The charts in figure 1 compare the performance of queries for each tested method using novelty-driven similarity and precision. Each of the 448 topics corresponds to a trial and is represented by a point. The point's vertical coordinate (z) corresponds to the performance of the incremental method, while the point's other two coordinates (x and y) correspond to the baseline and the Bo1-DFR methods. In addition we can observe

[1] http://dmoz.org

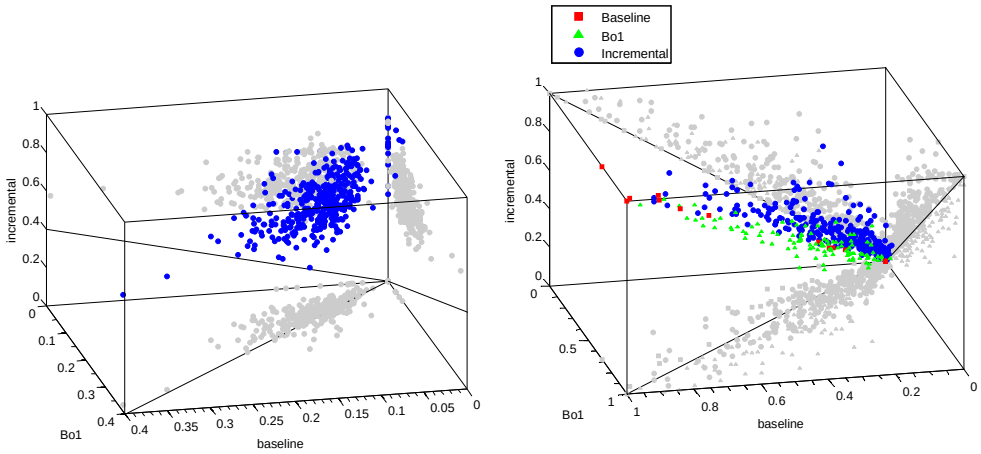

Fig. 1. A comparison of the three tested methods based on novelty-driven similarity (left) and precision (right).

the projection of each point on the x-y, x-z and y-z planes. For the x-z plane, the points above the diagonal correspond to cases in which the incremental method is superior to the baseline. Similarly, for the y-z plane, the points above the diagonal correspond to cases in which the incremental method is superior to Bo1-DFR. The x-y plane compares the performance of the baseline against Bo1-DFR.

It is interesting to note that for all the tested cases the incremental method was superior to the baseline and the Bo1-DFM method in terms of novelty-driven similarity. This highlights the usefulness of evolving the context vocabularies to discover good query terms. For the precision metric, the incremental method was strictly superior to the other two methods for 66.96% of the evaluated topics. Bo1-DFR was the best method for 24.33% of the topics and the baseline performed as well as one of the other two other methods for 8.70% of the topics. For the semantic precision metric (not shown for space limitations) the incremental method was strictly superior to the other methods for 65.18% of the topics, Bo1-DFR was superior for 27.90% of the topics and the baseline performed as well as one of the other two methods for 6.92% of the topics.

The next table presents the means and confidence intervals of the methods' performance based on σ^N, Precision and Precision^S. This comparison table shows that the improvements achieved by the incremental method with respect to the other methods are statistically significant.

		σ^N		Precision		Precision^S	
	N	mean	95% C.I.	mean	95% C.I.	mean	95% C.I.
Baseline	448	0.087	[0.0822;0.0924]	0.266	[0.2461;0.2863]	0.553	[0.5383;0.5679]
Bo1-DFR	448	0.075	[0.0710;0.0803]	0.307	[0.2859;0.3298]	0.590	[0.5750;0.6066]
Incremental	448	0.597	[0.5866;0.6073]	0.354	[0.3325;0.3764]	0.622	[0.6068;0.6372]

5 Conclusions

The vocabulary problem is a main challenge in human-system communication. In this paper we propose a solution to the semantic sensitivity problem, that is the limitation that arises when documents with similar context but different term vocabulary won't be associated, resulting in a false negative match. Our method operates by incrementally learning better vocabularies from a large external corpus such as the Web.

Other corpus-based approaches have been proposed to address the semantic sensitivity problem. For example, latent semantic analysis [6] applies singular value decomposition to reduce the dimensions of the term-document space, harvesting the latent relations existing between documents and between terms in large text corpora. Another corpus-based technique that has been applied to estimate semantic similarity is PMI-IR [20]. This information retrieval method is based on pointwise mutual information, which measures the strength of association between two elements (e.g., terms) by contrasting their observed frequency against their expected frequency. Differently from our approaches, these techniques are not based on an incrementally refined query submission process. Instead, they use a predefined collection of document to identify latent semantic relations. In addition, these techniques do not distinguish between the notions of topic descriptors and topic discriminators. The techniques for query term selection proposed in this paper share insights and motivations with other methods for query expansion and refinement [19,4]. However, systems applying these methods differ from our framework in that they support this process through query or browsing interfaces requiring explicit user intervention, rather than formulating queries automatically.

In this paper we have shown that by implementing an incremental context refinement method we can perform better than a baseline method, which submit queries directly from the initial context, and to the Bo1-DFR method, which does not refine queries based on context. This points to the usefulness of simultaneously taking advantage of the terms in the current thematic context and an external corpus to learn better vocabularies and to automatically tune queries.

References

1. Giambattista Amanti. *Probabilistics Models for Information Retrieval based on Divergence from Randomness*. PhD thesis, Department of Computing Science, University of Glasgow, UK, 2003.
2. Giambattista Amati, Claudio Carpineto, and Giovanni Romano. Query difficulty, robustness and selective application of query expansion. In *Advances in Information Retrieval, 26th European Conference on IR research*, pages 127–137. Springer Berlin / Heidelberg, 2004.
3. Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.
4. Bodo Billerbeck, Falk Scholer, Hugh E. Williams, and Justin Zobel. Query expansion using associated queries. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 2–9. ACM Press, 2003.
5. Jay Budzik, Kristian J. Hammond, and Larry Birnbaum. Information access in context. *Knowledge based systems*, 14(1–2):37–53, 2001.

6. Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.
7. John H. Holland. *Adaptation in natural and artificial systems*. Ann Arbor: The University of Michigan Press, 1975.
8. Reiner Kraft, Chi Chao Chang, Farzin Maghoul, and Ravi Kumar. Searching with context. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 477–486, New York, NY, USA, 2006. ACM.
9. K. L. Kwok and M. Chan. Improving two-stage ad-hoc retrieval for short queries. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 250–256, New York, NY, USA, 1998. ACM.
10. Carlos M. Lorenzetti, Rocio L. Cecchini, and Ana G. Maguitman. Intelligent methods for information access in context: The role of topic descriptors and discriminators. In *VIII Workshop de Agentes y Sistemas Inteligentes - CACIC 2007: XIII Congreso Argentino de Ciencias de la Computación*, Corrientes, Argentina, October 2007.
11. Ana Maguitman, David Leake, and Thomas Reichherzer. Suggesting novel but related topics: towards context-based support for knowledge model extension. In *IUI '05: Proceedings of the 10th international conference on Intelligent user interfaces*, pages 207–214, New York, NY, USA, 2005. ACM Press.
12. Ana Maguitman, David Leake, Thomas Reichherzer, and Filippo Menczer. Dynamic extraction of topic descriptors and discriminators: Towards automatic context-based topic search. In *Proceedings of the Thirteenth Conference on Information and Knowledge Management (CIKM)*, Washington, DC, November 2004. ACM Press.
13. Ana G. Maguitman, Filippo Menczer, Heather Roinestad, and Alessandro Vespignani. Algorithmic detection of semantic similarity. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 107–116, New York, NY, USA, 2005. ACM.
14. Iadh Ounis, Christina Lioma, Craig Macdonald, and Vassilis Plachouras. Research directions in Terrier: a search engine for advanced retrieval on the web. *Novatica/UPGRADE Special Issue on Web Information Access, Ricardo Baeza-Yates et al. (Eds), Invited Paper*, VIII(1):49–56, February 2007.
15. Eduardo H. Ramirez and Ramon F. Brena. Semantic contexts in the internet. In *LA-WEB '06: Proceedings of the Fourth Latin American Web Congress*, pages 74–81, Washington, DC, USA, 2006. IEEE Computer Society.
16. Jason D. M. Rennie and Tommi Jaakkola. Using term informativeness for named entity detection. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 353–360, New York, NY, USA, 2005. ACM.
17. J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The Smart retrieval system - experiments in automatic document processing*, pages 313–323. Englewood Cliffs, NJ: Prentice-Hall, 1971.
18. G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.
19. Falk Scholer and Hugh E. Williams. Query association for effective retrieval. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 324–331. ACM Press, 2002.
20. Peter D. Turney. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *EMCL '01: Proceedings of the 12th European Conference on Machine Learning*, pages 491–502, London, UK, 2001. Springer-Verlag.