Association for Information Systems
AIS Electronic Library (AISeL)

ECIS 2004 Proceedings

European Conference on Information Systems (ECIS)

2004

# Enhancing Query Reformulation Performance by Combining Content and Hypertext Analyses

Anis Benammar
*University Paul Sabatier*, b_ammar_anis@yahoo.fr

# ENHANCING QUERY REFORMULATION PERFOMANCE BY COMBINING CONTENT AND HYPERTEXT ANALYSES

Benammar, Anis, Département Service et Réseaux de Communication, IUT Castres-University Paul Sabatier, Avenu George Pompidou, 81100, Castres, France.
b_ammar_anis@yahoo.fr

## Abstract

*Information retrieval techniques play a critical role in the development of the information systems. Different searches have focused on the way of improving the retrieval effectiveness. Query expansion via relevance feedback is a good technique that proved to be a good way to improve the retrieval performance. In this paper, we investigate new methods to improve the query reformulation process. A two step process is employed to reformulate query. In a preliminary step, a local set of documents is built from the retrieved result. In a second step, a co-occurrence analysis is performed on the local document set to deduce the terms to be used for the query expansion. To build the local set we use firstly a content-based analysis. It is a similarity study between the retrieved documents and the query. The second method combines content and hypertext analyses to achieve the local set construction. The TREC[1] frame is used to evaluate the proposed processes.*

*Keywords: Information retrieval, query reformulation, hypertext analysis.*

[1] Text Retrieval Conference (http://trec.nist.gov)

# 1 INTRODUCTION

As the available online information incessantly grows, retrieving the relevant information becomes a fastidious task especially for a casual user [9]. Users have always difficulties to find the keywords that express exactly their information needs. Their queries are generally composed of few keywords which favors the word mismatch problem. Query reformulation is a well-known feedback method [1, 8] that is used to overcome such a problem. During the retrieval process, a query must regularly be adapted to follow the user information need. Query reformulation involves the following goals:

- Expanding the query with new terms from relevant documents [8, 15].
- Re-weighting the terms in the expanded query [17].

Several kinds of query reformulation methods have been studied in the literature. We are interested in methods that use the feedback from top-ranked documents in the search result. The method is called pseudo-relevance feedback [11, 15] as the relevance elements are not given by the user. In fact, the top-ranked documents in the results are assumed to be relevant. These relevant documents are then used in relevance feedback process to deduce the new query formulation. This method presents the following disadvantage: if some irrelevant documents appear among the first retrieved ones, the reformulation process will degrade the retrieval performance since added terms are extracted from irrelevant documents. The improvement of the results, when applying pseudo-relevance feedback depends on the quality of the top ranked documents [7, 11].

In this paper, we present a reformulation strategy addressing this problem. The reformulation process is carried out through a two step process. In the first stage, we build a local set of documents from the initial retrieved documents. This local set contains the documents that best fit the user information need. It is a re-ranking step which goal is to improve the quality of the documents to be used for the query reformulation. The local set is then analysed in order to extract the best terms to be added to the query. We present two methods for building the local documents set. The first method is based on a similarity analysis between the retrieved documents and the initial query. It is a content-based analysis because it uses only the document contents (index). The second method combines a connectivity-based algorithm with a content analysis. In fact, the link structure between retrieved documents is exploited to enhance the documents re-ordering.

The paper is organised as follows: section 2 presents the related works on the query reformulation methods. Section 3 details the content-based reformulation process we define to automatically update query. After discussing the obtained results on a TREC database, we introduce our approach using the hypertext analysis. Section 4 details the approach we propose to combine content and hypertext-based analyses to enhance the local document set construction.

# 2 RELATED WORK

Query reformulation approaches through query expansion are generally used to create a new query expression according to the associated retrieved documents. Query expansion methods consist in adding new terms to the initial query. These approaches can be grouped in three categories:

- User relevance feedback approaches: they are based on user feedback. In these approaches, the user is asked to mark the documents he considers as relevant. An automatic analysis allows the system to select the terms that are representative of the relevant documents and add them to the new query formulation [17].
- Global techniques: these approaches are based on the analysis of the entire document collection. They are typically based on the association hypothesis [21]. One of the earliest global techniques is term clustering [19] which groups related terms into clusters based on their co-occurrence in the corpus. The most correlated terms with the query terms are then added to the query expression.

- Pseudo-relevance feedback and local techniques: these approaches are based on the set of initially retrieved documents. There are mainly two local strategies: the local relevance feedback [1] and the local context analysis [12]. With regard to the local relevance feedback strategy, the main idea is to build global structures such as an association matrix that quantifies term correlation. The more correlated terms are then used to expand the query. The local context analysis is a more recent local technique; it inherits the notions of global approach (use of the passages and concepts) and applies it to a local set of documents. The concepts are selected from the top ranked documents or passages based on their co-occurrence with the query.

A problem that always raises with the pseudo-relevance feedback concerns the ‘quality’ of the documents used for the feedback [7, 11, 15]. In several cases, irrelevant documents appear among the top ranked ones. Using these documents to achieve the reformulation will probably degrade the retrieval performance. Through the reformulation process we propose, we try to address this problem by including a preliminary step which goal is to select the best documents among the first retrieved ones. These documents that constitute a local set are analysed in a second step to deduce the query reformulation scenarios.

# 3 QUERY REFORMULATION APPROACH

The approach we develop to reformulate the query deals with the document quality problem. It integrates the following aspects:

- The documents that are used for the query modification are the retrieved ones. In that sense, our approach is similar to the local relevance feedback for which the modification is based only on the retrieved documents. However, we use an additional step that consists in re-ranking the initial retrieved documents [16]. The re-ranking step is based on a similarity analysis between the top ranked documents in the search result and the current query.
- The document analysis is based on co-occurrence measures. The co-occurrences between the terms from the documents and the whole query are computed to deduce the terms to be added to the query.

In the following sections, we present in details the reformulation process.

## 3.1 A two step process

Throughout a search session the query is automatically adapted according to the search results and to the relevance feedback information.

A classical retrieval process is firstly employed to retrieve the documents in response to the initial request. The results are then used to reformulate the query expression. The major difference with the other local relevance feedback approaches is the use of a preliminary step consisting in the construction of a local set of documents from the search result. In the second step of the process, a co-occurrence analysis is performed on the local set documents. This analysis consists in measuring the correlation between the terms of the local set and the query. The most correlated terms are then used for the expansion. In the following sections we detail the adaptation process steps.

### *3.1.1 Document re-ranking based on content analysis*

This stage consists in building a local set of documents from the search result. This set gathers the documents that best fit the search context. It is computed by re-ranking the documents that have initially been retrieved (100 top ranked documents). Indeed, instead of directly using the top documents in the retrieved set as it is usually done, we re-rank the first 100 documents according to the current query and we select the new top ranked documents. The goal is to get in the top list relevant documents that were initially ranked in a lower position. The Documents reordering is achieved by computing the similarity between the documents and the query contents (index). The similarity measure used to compute the local set is different but complementary to the one used to retrieve the first result.

The documents are represented as vectors in the term space. To each term is associated a weight computed as follows:

$$w(t_d) = \log_{10}\left(\frac{N}{N_{t_d}}\right) \qquad (1)$$

Where td is a term from the document d, N is the number of documents from the collection and Ntd is the number of documents in the collection that contain the term td. The query is also represented in a vector space but a different formula (2) is used to compute a term weight:

$$w(t_q) = \frac{tf(t_q)}{tf_{max}} \qquad (2)$$

Where tf(tq) is the term frequency in the query and tfmax is the highest frequency in the query.

The documents that are most correlated to the query constitute the local document set that is used to perform the co-occurrence analysis.

The similarity value between a document vector $\vec{D}$ and a query vector $\vec{Q}$ using, for example, the Cosine measure is computed as follows:

$$\mathrm{Sim}(\vec{D},\vec{Q}) = \mathrm{Cos}(\vec{D},\vec{Q}) = \frac{\sum_{i=1}^{n} wt(td_i)wt(tq_i)}{\sqrt{\left(\sum_{i=1}^{n} w(td_i)\right)^2} \times \sqrt{\left(\sum_{i=1}^{n} w(tq_i)\right)^2}} \qquad (3)$$

Where n is the number of documents in the local set.

### *3.1.2 Local document set analysis*

The analysis is carried out on the basis of the local set of documents as built in the first step. The goal of this analysis is to automatically extract the correlations that exist between the terms from the documents and the query. Then, the most correlated terms from the local documents set are used to expand the query. The term correlations are computed using term co-occurrence as follows [4]:

$$\mathrm{Correlatio\ n}(t_i, Q) = \sum_{j \in Q} \mathrm{cooccur}(t_i, q_j) \qquad (4)$$

Where ti is a term from a document in the local set and qj is a term from the query Q.

The co-occurrence degree between a term from a document and a term from the query is expressed as follows:

$$\mathrm{cooccur}(t_i, q_j) = \log_{10}(\mathrm{Co-occ}(t_i, q_j) + 1) \times \mathrm{idf}(t_i) / \log_{10}(n) \qquad (5)$$

$$\mathrm{Co\text{-}occ}(t_i, q_j) = \sum_{d\ in\ S} tf(t_i, d) \times tf(q_j, d) \qquad (6)$$

Where:

- tf (ti,d) and tf (qj,d) are the frequencies of the terms ti and qj in the document d
- n is the number of documents in the local set
- N is the number of documents in the collection
- Nti is the number of documents in the collection that contain the term ti.
- $\mathrm{idf}(t_i) = \log_{10}(N/N_{t_i})$

Several parameters in these formulas such as the number of documents in the local set are evaluated in the experimentation section. The goal is to determine the optimal values of the different parameters.

## 3.2 Experiments and results

The experiments were conducted to evaluate the importance of the re-ranking step: we compare the search result when applying the entire reformulation process and the result when directly using the top ranked documents as it is usually done in the literature (without re-ranking step). The experiments were carried out on the medical collection OHSUMED 1987. Preliminary experiments have been done to choose the optimal parameter values. Several tests were performed to deduce the optimal local set size. Finally, we set it at 5 documents. The inner product proved to be the best measure when searching similar documents to the query. We also used a threshold value when building the local set of documents. The number of terms (n) to add to the query expression depends on the number of documents in the local set. We did several evaluations to set this parameter (number of terms to add). The best results were obtained using the formula (6) defined in [4]. Let i be the number of documents in the local set, n is computed as follows:

$$\mathrm{n} = 5 \times i + 5 \qquad (7)$$

The following table summarises the results. The results are evaluated according to the precision (at 10, 15 and 20 documents), the average precision and the exact precision. The first row represents the first retrieved results. The second row presents the results when applying the co-occurrence analysis on the top ranked documents. The third row describes the obtained results when applying the full process including the re-ranking and the co-occurrence analysis.

| | P10 | P15 | P20 | AvgPr | ExactPr |
|---|---|---|---|---|---|
| Initial result | 0.387 | 0.326 | 0.276 | 0.394 | 0.389 |
| Reformulation results: No re-ranking | 0.427 | 0.342 | 0.290 | 0.417 | 0.418 |
| Reformulation results: Full process | 0.427 | 0.361 | 0.305 | 0.418 | 0.423 |

*Table 1. Results*

The experiments show that the adaptation process using the re-ranking step improves the results better than the adaptation process using directly the top ranked documents. In 63% of the cases, the reformulation process is more effective (or similar) when using the re-ranking. In fact the re-ranking improves the quality of the local set used in the analysis by including more relevant documents. In some cases, even if the re-ranking step does not include more relevant documents in the local set, the retrieval performance increases. In this latter case, the documents are not relevant, but they are close enough to the search context so the results are also improved. The re-ranking step shows the importance of the quality of the documents used to achieve the relevance feedback. In these cases, the analysis performs better because the added terms are relevant since they are extracted from relevant documents.

The result table shown above reveals the importance of the preliminary step of constructing the local document set. The experiments show that using the re-ranking step improves the results more than the adaptation process using directly the top ranked documents in the initial retrieved set. Additional experiments confirm the last conclusion: when the local set is manually constructed by choosing the relevant documents[2], the retrieval results are always improved (user relevance feedback). however, we can investigate other methods to enhance these results. In fact, the enhancement of retrieval performance is insufficient especially for the elementary precision. We extend our study to deal with

[2] A relevance file indicating the relevant documents for each query is provided with TREC collections.

the network structure of a hyperlinked environment as a special document type. Indeed, the link structure is a rich source of information that can be exploited to enhance the quality of the documents in the local set.

## 4 HYPERTEXT ANALYSIS IN INFORMATION RETRIEVAL

Actually, the majority of the information retrieval systems uses only the content information to independently index, retrieve and organize documents. Firstly, user has to formulate query that specifies their information need. Then, the retrieval system does keywords searches against documents corpus to find relevant documents. Many factors influencing the retrieval result are ignored. Mainly, the link structure is a rich source of information that can be exploited to enhance the quality of the retrieval results.

After presenting some existing works using the hypertext structures in information retrieval, we present the approach we propose to enhance the effectiveness of the local set construction. Our approach combines the hypertext and the content-based analyses to re-rank the first retrieved documents.

### 4.1 Hypertext in information retrieval: several approaches

Recently, many searches confirm the importance of the hyperlink structure as an efficient tool for indexing [13], searching [6, 13, 18] and organizing [5, 10, 14, 18, 20] information. In his work, Gery [13] presents an information retrieval system exploiting the link structure. The main idea in these works is to combine content-based analysis with hypertext analysis to enhance the indexing and the retrieval effectiveness. Accessibility information is the main notion introduced by Gery [13]. It traduces the information that user can join from a given document. This information is taken into account when indexing and when searching information. A document is modeled using at the same time its content and its accessibility information. Therefore, when searching relevant documents for a given query the accessibility information is also included in the RSV (Retrieval Status Value) computation. Contrary, PageRank [6] is a link-based method that integrates the link analysis only in the retrieval process. It is a method of rating documents objectively and automatically. It is implemented in Google. The importance of the document depends on its incoming links (documents that point to it). Others work focus on the use of the link-based analysis to organize the result returned by a content-based retrieval system. In fact, the indexing and the retrieval processes are achieved using only the document and the query contents. Then, the hypertext analysis is used to reorganize retrieved documents essentially to enhance precision at the top-ranked documents. The HITS algorithm [14] computes for each document in the initial retrieved set a hub and an authority score. An authority document contains relevant information. On the other hand, hub document contains links towards relevant documents. The HITS algorithm is built around the mutually reinforcing principle [14]. Many following searches had worked on the HITS algorithm to improve its effectiveness [5, 10]. Bharat [5] reminds mainly two problems. Firstly, if the vicinity graph contains irrelevant documents that are well interconnected, then the search topic can be drifted. The second problem concerns the automatically generated links. The solution proposed by Bharat is based on link weight evaluation. A link weight value is equivalent to the relevance of the document containing the link (source). Documents having small relevance values are removed from the vicinity graph. A modified version of the HITS algorithm is then applied on the obtained weighted graph.Chakrabarti in [10] presents another solution to overcome the problems associated with the HITS algorithm. In fact, this latter is enhanced by a content-based analysis. A weight is associated to each link in the vicinity graph depending on the text anchor. Indeed, the text around the link source for a document d1 to another document d2 describes the content of d2 [10]. The link weight is computed as a correspondence function between the anchor text and the initial query. The link weights are then included in the authority and the hub scores computing.

The approach we develop is an extension of HITS algorithm that integrates the links weights computation. Many criteria are taken into account when evaluating the importance of a link in the vicinity graph. The following section details our method for analyzing the hypertext and the content information.

## 4.2 Document re-ranking based on hypertext analysis

The algorithm we propose to re-rank the retrieved document set is composed of the following steps.

### *4.2.1 Vicinity graph building*

The vicinity graph is a directed graph where nodes correspond to the first retrieved documents and the edges correspond to the hyperlinks [20]. When building the link graph, two alternatives arise. Firstly, the link graph can be limited to the retrieved documents. It means that only the initially retrieved documents are included in the graph. Indeed, in the first test series, we have noticed that many relevant documents appear in the first retrieved result but in lower positions. Our goal is then to get these documents in the top list. This alternative presents the following disadvantages:

- Some relevant documents do not appear in the first retrieved result.
- The graph analysis will be less effective if the link structure between retrieved documents is poor (not enough links between documents).

On the other hand, this alternative is less expensive in term of time and space processing because we consider only the documents in the retrieved result.

The second alternative expands the retrieved document set by new documents that do not appear in the first result. These documents point to, or are pointed to by, any documents in the first retrieved result. In spite of the cost of the latter alternative, we choose to apply it because when we expand the initially retrieved set, we guarantee a rich link structure.

### *4.2.2 Link weight computation*

In this step, a weight is assigned to each link. When computing the link weight, many criteria are taken into account:

- Link typology: A link between documents in the same host (organizational link) must have less importance then links between documents from different hosts (navigational link). In fact, many links are created for organizational purposes. Some works, such as the link weighting scheme of Bart and Henzeinger [5], assume that organizational links must have a null weight value. In our case, we will not eliminate such links because even in the same host, we can find many relevant documents. Though, organizational links will have reduced weight to avoid that a same host dominate the analysis result.
- Link relevance: In [5], the link weight depends only on the document representing the link source. In fact, if the document representing the link source is not relevant then there is less chance that the referenced document will be relevant. Inversely, when the link source document is relevant, the referenced document will probably be as well [18]. Let examine the following example to show the limit of this assumption: suppose we examine a query dealing with a scientific topic, in the first result we can get the home page of a searcher. From this page we have many links as the searcher publications that are also relevant, but we can also have links to the searcher hobbies. These latter links are not relevant. If we consider the assumption that the link weight depends on the document containing the link, such pages will also be considered as relevant. To overcome such a problem, the link weight in our model depends on the relevance of the documents representing the source and the destination of the link. Document relevance is computed as the similarity with the initial query as done in our first approach. Here we combine two analyses: a link-based analysis to search related documents and content-based analysis to evaluate the link weight relevance. The link weight between

two document d1 and d2 depends simultaneously on Sim(d1,Q) and Sim(d2,Q) that represent the similarity between the query and each document. The similarity values are compute as in (3).

Let wld1→d2 be the weight value of the link between d1 and d2 when d1 points to d2 (the same value is attributed to wld2→d1). The link weight is computed as follows:

$$wl_{d1\rightarrow d2} = \beta \times \text{Relevance}\ (d_1 \rightarrow d_2) \qquad (8)$$

$$\text{Relevance}\ (d_1 \rightarrow d_2) = \text{Sim}\ (d_1, Q) \times \text{Sim}\ (d_2, Q) \qquad (9)$$

The parameter β depends on the link typology. It is used to moderate the organisational link effect.

### *4.2.3 Hub and Authority scores computation*

In this step, an extension of the HITS algorithm using the link weights is performed to compute the hub and the authority scores. Let $x^{<di>}$ and $y^{<di>}$ be the authority and the hub scores of the document $d_i$. The score values are alternatively computed as follows:

$$x^{<di>} \leftarrow \sum_{dj:(dj,\ di)\in E} wl_{dj\rightarrow di} \times y^{<dj>} \qquad (10)$$

$$y^{<di>} \leftarrow \sum_{dj:(di,\ dj)\in E} wl_{di\rightarrow dj} \times x^{<dj>} \qquad (11)$$

E is a set of couples $(d_i,d_j)$ where $d_i$ points to $d_j$.

These operations (10 and 11) are successively performed until we reach a stability point.

### *4.2.4 Document re-ordering*

Finally, the documents having the best authority scores are included in the local document set.

## 4.3 Experiments and results on hypertext analysis

The following table describes the WebTrack10 collection used during the hypertext analysis experiments:

| | |
|---|---|
| Number of documents | 1679521 documents |
| Collection size | 10 Giga Bytes |
| Number of terms | 2372149 terms |
| Document average size | 267 terms |
| Number of distinct terms per document | 122 terms |

*Table 2. Experiment collection characteristics*

The obtained results are summarised as follows [2]:

- The vicinity graph analysis can be limited to the initially retrieved documents. The graph extension does not improve the final re-ranking results.
- The contribution of the organizational link in the hypertext analysis is limited. However, the organizational links can be considered only when one domain (host) contains many relevant documents.
- The combination of the hypertext and the content analyses proved to be effective. Indeed, the assignment of the link weight makes it possible to improve the values of high precision more than the use of only hypertext analysis.
- The hypertext analysis approach contribution is not significant. Indeed, the proposed process did not improve the precision values when re-ranking the first retrieved documents.

With an aim of validating our hypertext analysis approach, we also participated in the Web Distillation task in TREC'2002 conference [2, 3]. This enabled us to compare our results to those of the other participants being interested in hypertext analysis. The following graph describes our submitted results in comparison with the others participants results.

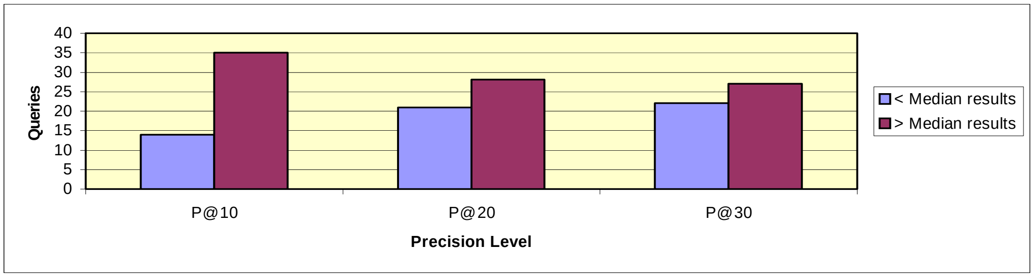

*Figure 1.* *Comparison to TREC'2002 participant results*

The graphics shows that we overtake the median values for the majority of the queries especially for the precision at 10. However, theses submitted results to TREC were not our best runs. In fact, best results were obtained when we take into account the first document ranking: for the documents having a null authority value we keep the same ranking order it obtain in the first retrieved result.

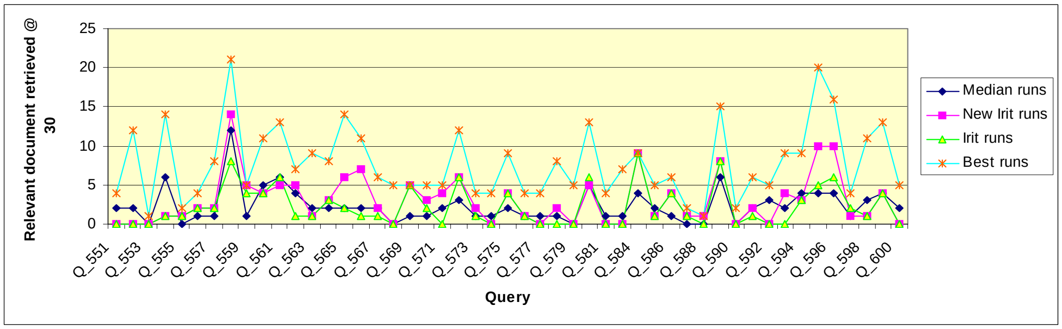

*Figure 2.* *New runs (precision at 30 documents)*

In many case, the new runs considering the document ranking in the initially retrieved result are best than the submitted results to TREC. The precision values at 10 in both submitted runs and new runs are mostly comparable to the median results. The submitted runs are slightly below the median runs for the precision at 20 and 30.

## 5 GENERAL CONCLUSION

In this paper, we are interested on query reformulation techniques. The process we propose is based on a co-occurrence analysis performed over a subset of the initially retrieved documents. It is a similar approach to the local relevance feedback where the first retrieved documents are directly analysed to deduce the query modification. The reformulation results can be degraded if some irrelevant documents appear among the first retrieved ones. To overcome this problem, we induce in our process a preliminary re-ranking step which goal is to construct a local set that will contain the documents that best fit the user information need. The local documents set is firstly constructed using a content-based analysis. The test evaluations performed on a TREC collection shows that elementary as well as average and exact precision are enhanced when applying the re-ranking step. Indeed, it improves the quality of the documents used to achieve the query reformulation. However, the enhancement percentages are not very important. Our study is then completed to deal with the network structure of a hyperlinked environment. The second solution we propose to enhance the quality of the local document set is a combined analysis. To re-rank the retrieved documents, we combine a hypertext and a content-based analyses. Through the experiments performed on WebTrack10 collection, we noticed that the hypertext analysis approach for document re-ranking is not always effective. Nevertheless, we have proved the importance of the association of content information when analysing the hypertext structure.

Actually, we continue to study the hypertext approach for document re-ranking. We focus particularly on a more effective combination of hypertext and content analyses. In fact, in the proposed approach except the link weight assignment, the content and the hypertext information are alternatively exploited. Others test collections will be also employed to validate the hypertext analysis approach.

## References

1. Attar, R. and Frankel, A. S. (1977). Local feedback in full-text retrieval systems, Journal of associations for Computing Machinery, 24 (3), 397-417.
2. Benammar, A. (2003). Profils en recherche d'information : définition, exploitation et adaptation, Thèse de Doctorat, Université Paul Sabatier, Toulouse France.
3. Benammar, A., Hubert, G., Laffaire, C., Mothe, J. and Boughanem, M. (2002). IRIT at trec'2002: Web Track, http://trec.nist.gov/pubs/trec11/t11_proceedings.html.
4. Belkin, N. J. (1999). Relevance feedback versus local context analysis as term suggestion devices, Rutger's TREC-8 Interactive Track Experience, Proceedings of TREC-8, 565- 574.
5. Bharat, K. and Henzeinger, M. (1998). Improved algorithms for topic distillation in a hyperlinked environment, Proceedings of ACM-SIGIR.
6. Brin, S. and Page L. (1998). The PageRank citation ranking: bringing order to the Web, In technical report available at http://www-db.stanford.edu/~backrub/pageranksub.ps.
7. Boughanem, M., Chrisment, C. and Soulé-Dupuy, C. (1999). Query modification based on relevance back-propagation in ad hoc environment, Information Processing & Management 35, 121-139.
8. Buckley, C., Salton, G. and Allan, J. (1994). The effect of adding information in a relevance feedback environment, Conference on Research and development in Information Retrieval (SIGIR).
9. Buckley, C., Mitra, S. and Singhal, A. (1998). Improving automatic query expansion, Proceedings of the 21st ACM SIGIR Conference on Research and development in Information Retrieval.
10. Chakarabarti, S., Dom, B., Gibson, D. and Kleinberg J. M. (1998). Automatic resource compilation by analyzing hyperlink structure and associated text, Proceedings of the 7th World Wide Web Conference.
11. Croft, W.B. and Xu, J. (1996). Query expansion using local and global document analysis, Proceeding of the 19th Annual International ACM SIGIR Conference on research and development in Information retrieval SIGIR 96.
12. Croft, W.B. and Xu, J.(2000). Improving effectiveness of information retrieval with local context analysis, ACM Transaction on Information systems 18(1), 79-112.
13. Gery, M. (1999). Recherche de zones de pertinence sur le World Wide Web, INFORSID.
14. Kleinberg, J. M. (1998). Authoriative sources in a hyperlinked environments, Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms.
15. Eguchi, K. (2000). Incremental query expansion using local information of clusters, Proceedings of the 4th World Multiconference on systemics, Cybernetics and informatics (SCI 2000), 2, 310-316.
16. Mothe, J. (2000). Correspondance analysis method applied to document re-ranking, Internal report IRIT/00-22 R.
17. Rocchio, J. J. (1971). Relevance feedback in information retrieval, In G. Salton, editor, The Smart retrieval System, Experiments in Automatic Document processing.

18. Savoy, J., Picard, J. (2000). Retrieval effectiveness on the Web. Information, Processing & Management 37, 543-569.
19. Sparck, J. (1971). Automatic keywords classification for information retrieval, Buterworths, London.
20. Sun, J. (2000). World Wide Web information retrieval using Web connectivity information, http://citeseer.nj.nec.com/448145.html.
21. Yonggang, Q. and Frei, H. F. (1993). Concept based query expansion, Proceedings of the 16th ACM SIGIR Conference on Research and development in information retrieval, 160-169.