

# Use reformulated profile in information filtering

**Ciro M. Santos**

Computer Science College of Caratinga  
168 Joo Pinheiro Av. - Caratinga - MG - Brazil

**Newton J. Vieira**

Federal University of Minas Gerais  
6627 Antonio Carlos Av. - Belo Horizonte - MG - Brazil

## Abstract

The aim of this article is to present a proposal to improve the quality of query results in information retrieval systems for the Web. The vector-space model is used for the information retrieval. Then, the retrieved information is filtered taken as base the user's profile. After obtaining classification through the filtering process, the profile is reformulated taking into account the user's captured information by using the method of word contribution (K. Hoashi & Hashimoto 1999)(K. Hoashi & Hashimoto 2000a)(K. Hoashi & Hashimoto 2000b). This process introduces a certain level of knowledge about the user into the system.

Through filtering, considering the user's initial profile as well as its reformulation, there is an improvement in the quality of the results returned by the search engines. Results from experiments show that, in place where the control and interaction with the user are possible, this method, results in an improvement in information retrieval system performance.

## Introduction

World Wide Web became, through the years, an important source of bibliographical resources. This happened, not just for the enormous amount of available information, but also for ease of recovery of this information. Such ease is supplied by search engines like Altavista<sup>1</sup>, Google<sup>2</sup>, Yahoo<sup>3</sup>, and many others, which work with problems of storing and organizing information so that its retrieval is fast, efficient and fulfill the researchers' needs.

The traditional information retrieval systems are composed of three basic processes: the collection, the indexing and the similarity calculation. The collector is responsible for collecting available information in the Web. After collect the information is processed, indexed and stored as an inverted file. The similarity calculation is made by the search engine starting from an analysis of the query content and the existing information in the inverted file index. This process returns a list of ranked documents, taking into account the occurrence of terms in the query and the documents. However, the context in which the query is inserted,

or even the user's profile which could be used to improve the ranking of retrieval documents, are not usually used by the search engines.

## Related works

The problem of filtering information using retrieval techniques has been researched thoroughly by the community of Computer Science. Next we present a brief revision of the main works in this direction, as well as other works that originated important techniques that were considered in this article.

In Salton (G. Salton & Yu 1982) (G. Salton & Yu 1975), a theory was presented to classify terms in agreement with discrimination capacity among documents that are part of a collection. That theory called term importance for discrimination value analysis, gives value to a term according to how well they are able to discriminate documents of a collection from each other; that is, the value of a term depends on how much the average separation between individual documents changes when the given term is assigned for content identification. The best terms are those that achieve the greatest separation. The system developed in this work also gives values to the terms of documents depending on the relevance of the terms for the documents.

The *word contribution* method (K. Hoashi & Hashimoto 2000a) (K. Hoashi & Hashimoto 2000b) calculates a measure to express the influence of the word to the similarity between query and document. This measure represents the contribution of the term for the relevance of the document. All terms that have positive weight above an established threshold are added to the user's profile. Words with great positive contribution are words that happen in the query and in the documents. That algorithm was used for reformulation of the user's profile.

Bruce Krulwich (Krulwich & Burkey 1997), points out that one of the most important aspects in the filtering task is to create the user's profile. This process is implemented through techniques of profile reformulation using information from relevant documents. However, which information from the text can be used for the reformulation? This question is crucial for this process. The use of the whole text allows that irrelevant terms are used to compose the new profile, harming the selection of relevant documents. Although, in our work the amount of terms inserted in the profile will

Copyright © 2004, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

<sup>1</sup>www.altavista.com

<sup>2</sup>www.google.com

<sup>3</sup>www.yahoo.com

depend on its importance, according to results obtained in the application of the *word contribution* method (K. Hoashi & Hashimoto 2000a) (K. Hoashi & Hashimoto 2000b).

### Filtering of information

We can describe a filtering information system as being an automatic mechanism with the capacity of monitoring a continuous flow of documents and ability to select documents considering its relevance for a certain user or users' groups, according to its needs. These needs are represented through a profile of interests associated to the user or users' group. The ability to select relevant documents is associated with the mechanisms of retrieval information that calculate the value of similarity between documents of the collection and the profiles. Documents of great similarity with the profiles are considered important for the user or users' group. Nevertheless, due to personal or professional reasons, a user's interests may shift or change. These changes may happen in a relatively short duration of time or over a long period of time. The shifts can affect the user's interests partially or fully. To cope with this problem, it should be possible to do reformulation on the user's profile. This actualization is made through information sent to the system about the relevance of the received documents.

The filtering system proposed in this article presents a different aspect in relation to the generic filtering system, once it intends to recover relevant documents for a certain query. The terms of the query will give the scope of the information filtering. This filter intends to improve the ranking of documents previously retrieved, taking into account the user's profile and not to add a new document to fulfill its needs.

### The vector-space model

The vector-space model uses a vector of terms to represent the frequency in which each term appears in the document. That model has been widely tested and has gotten good results in the retrieval of documents, because it considers partial matching and the proximity of documents in relation to the terms of the query (Ribeiro-Neto & Baeza-Yates 1998) (I. Witten & Bell 1999).

In vector-space model, documents and querying are represented through vectors of terms in a vector-space so that documents with similar content have similar vectors, as shown in the Figure 1. Dimension of this space is given by the number of terms differently from the collection and the coordinates vector that are represented by the importance of each document from the collection for the query.

The document  $d_j$  is represented by a vector  $\vec{d}_j = (w_{j,1}, w_{j,2}, \dots, w_{j,n})$ , where  $n$  represents the number of different terms from the collection and  $w_{j,t}$  is the weight of the term  $t$  in the document  $d_j$ . The querying is represented by a querying vector  $\vec{q} = (w_{q,1}, w_{q,2}, \dots, w_{q,n})$ , where  $w_{q,t}$  is the weight of the term  $t$  in the query  $q$ . The profile is represented by a profile vector  $\vec{p} = (w_{p,1}, w_{p,2}, \dots, w_{p,n})$ , where  $w_{p,t}$  is the weight of the term  $t$  in the profile  $p$ .

Similarity among a document  $d_j$  and a query  $q$  or profile  $p$  is defined as a correlation among vectors  $\vec{d}_j$  e  $\vec{q}$  or  $\vec{d}_j$  and  $\vec{p}$ .

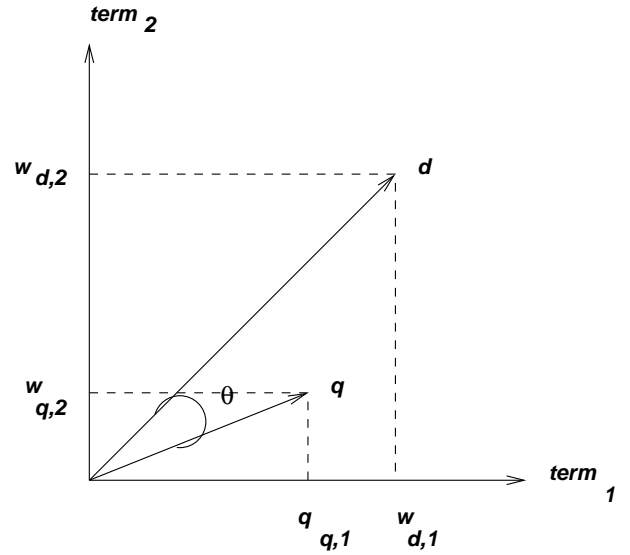


Figure 1: Representation of query and documents vectors in a space of 2 dimensions.

That similarity is measured through calculation of the cosine between the two vectors. The equation below is used to calculate the similarity among a document  $d_j$  and query  $q$ , through the cosine of the angle among the vectors  $\vec{d}_j$  e  $\vec{q}$ :

$$\text{sim}(\vec{d}_j, \vec{q}) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|} = \frac{\sum_{t=1}^n w_{d_j,t} w_{q,t}}{\sqrt{\sum_{t=1}^n w_{d_j,t}^2} \sqrt{\sum_{t=1}^n w_{q,t}^2}}$$

### Creation and administration of the user's profile

The process to create a profile that expresses the user's interest is complex and it depends on characteristics that represent user, for example, personal preferences, knowledge degree, vocabulary, information on the content of the collection, knowledge on the type of used index, etc. Those characteristics give to the profile a subjective context that can be used by search engines.

Administration of the profile consists on the most sensitive part of the filtering system, because it is from the current state of the profile that the system determines which information is relevant for a certain user. The current content of the profile is used by the system in the calculation of similarity among user profile and rank documents returns by the system. The actualizations in the profile, with basis in the user's behavior and assimilate the user's typical interests. The continuous actualization provides a gradual and constant adaptation to the user's profile. In the next section the theoretical foundations will be presented on what was implemented by the module for profile administration.

**Method of word contribution** The method of word contribution (*WC*) is a measure, which expresses the influence of a word to similarity between the query and a document. In mathematical terms, the word contribution can be described through this Equation:

$$Cont(w, q, d_j) = Sim(q, d_j) - Sim(q'(w), d'_j(w))$$

Where  $Cont(w, q, d_j)$  is the contribution of the word  $w$  for the similarity between the query  $q$  and the document  $d_j$ ,  $Sim(q, d_j)$  is the similarity among the query  $q$  and the document  $d$ , and  $Sim(q'(w), d'_j(w))$  is the similarity among the query  $q$  and the document  $d$ , excluding the word  $w$  of the query  $q$  and the document  $d$ ,  $q'(w)$  is the query  $q$ , excluding the word  $w$ , and  $d'_j(w)$  is the document  $d$ , excluding the word  $w$  in the querying and in the document.

As mentioned by Hoashi (K. Hoashi & Hashimoto 2000a) (K. Hoashi & Hashimoto 2000b), most of the terms of a document has its contribution close to zero for the relevance of the document for the query. A small number of words have negative influence and another small amount of terms have positive contribution for the relevance of documents. The terms with positive contribution, or some of them, are used by the reformulation module for the profile actualization.

Initially the word contribution of all the terms belonging to the query and the document are calculated using the WC presented previously. The terms  $w$  with smaller  $Cont(w, q, d_j)$  values, will be used for the calculation of the *Score* through the Equation below:

$$Score(w) = wgt \times \sum_{d_j \in Drel(q)} Cont(w, q, d_j)$$

where  $wgt$  is a parameter with a negative value (since the contribution of the extracted word is also negative), and  $Cont(w, q, d_j)$  is the *WC* of word  $w$  to the similarity of profile  $p$  and document  $d$ . On this procedure, the calculated score is regarded as the TF (term frequency) element of the word. Finally, all extracted words and their weights are added to the profile. The value of  $wgt$  depends on the characteristics of each collection, and  $Drel(q)$  are the relevant documents sent by the user.

### Implementation of the filtering system

The filtering system implemented, which is schematized in Figure 2, consists mainly of four processing stages: registration, authentication, querying and profile reformulation. First it is made the user's registration through the registration module, in which the user supplies personal information that will be used in the authentication process and creation of the initial profile. After the registration, the user makes the authentication and then he is able to submit a query through the query module. That module receives and processes the requisition returning a ranking list with the relevant documents. With base in those documents the filtering module retrieves the terms of the user's profile and it makes the calculation of the similarity among the terms of the profile and the documents returned by the query. The result is a ranking list. When receiving documents from the filtering system the user evaluates the results and then sends to the reformulation module relevant documents. That module receives the documents transmitted by the user, recovers the original query

and applies the algorithm of word contribution between the documents, the query and the profile, storing in the user's profile the terms that more contributed to the relevance of the documents according to the user's classification.

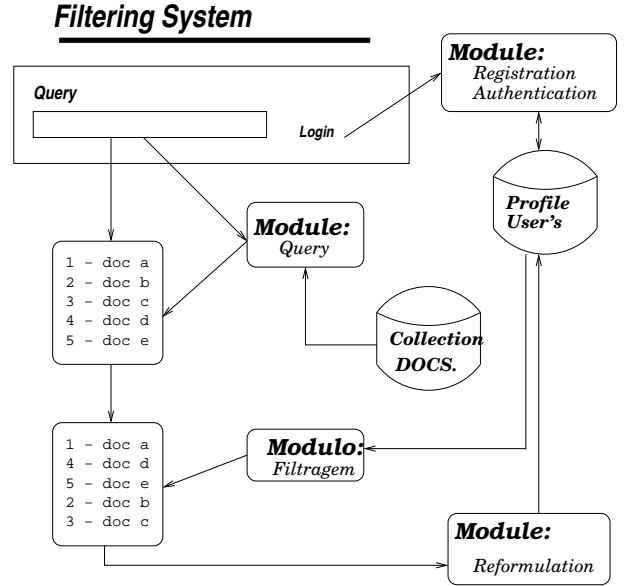


Figure 2: Architecture for the filtering system.

### Querying Process

The querying process in the filtering system is executed in two stages, described next. The first stage consists of the application of the vector-space model to calculate the similarity of documents from the collection with the terms of the query. The obtained result is a list of rank documents in a decreasing order according to the value of the similarity. The cutoff point  $k$  is the fraction of the top  $k$  ranked documents that are relevant to the query forming a subcollection of documents. The value of  $k$  depends on the characteristics of the collection. This subcollection is kept to be accumulated with the results of a second stage in the filtering process.

In the second stage the vector-space model is used for the calculation of the similarity between the terms of the user's profile and the subcollection created on the first stage of the filtering system. The rank of the documents is made through the sum of the similarities obtained in the two stages of the filtering process.

The final result returned by the filtering system is a list of documents orderly decreasing according to the values of similarities between the terms of the query, the profile and the documents of the collection.

When analyzing the documents returned by the system, the user sends a feedback that will be used for reformulation of the profile, applied the algorithm of word contribution, in accordance with (K. Hoashi & Hashimoto 1999) (K. Hoashi & Hashimoto 2000a) (K. Hoashi & Hashimoto 2000b).

## System Evaluation

In order to test our filtering system implemented in this work, we use two collections one called *FCCCMG* and the other *CACM* which are show on Table 1

Collections	FCCCMG	CACM
Documents	1.750	3.204
Querying	197	12
Terms	8729	7121
Average docs size	48,45	24,26
Profile	35	12
Size of the Collection	1,740 Mbytes	1,8 Mbytes

Table 1: Collection used for the experiments.

### Collection FCCCMG

The *FCCCMG* collection, shows on Table 1, was created with the participation of 35 students from Computer Science College of Caratinga. It was composed by research document on search engines called Todobr<sup>4</sup>, Google<sup>5</sup>, and Cad<sup>6</sup>, on the subjects movies, cookery, education, sport, computer science, politics, health, safety and tourism, for 60 days. The search engines and topics choice were done by the students considering their ability and knowledge. The rules to create the collection were:

- each student had a subject for research;
- the querying could only happen on the search engines above;
- each student retrieval 50 documents on the average;
- each student classified the 10 most important documents according to his/her preference.

### Collection CACM

The second database we have used in our experiments is a collecton of *CACM*, extracted from the TREC project (D. A. Hull & Schutze 1996) (Hull 1998) (Hull 1999). This collection are available in the *LATIN - Laboratory for Information Treatment of DCC/UFMG*, and are discriminated according to Table 1.

### Utility

We used the scaled *Utility*, shown in the Equation 1, to evaluate the results of the query using the collection *FCCCMG*. That measure supplies a value associated to four important categories. However, we only have knowledge of two categories:

- the group of recovered relevant documents;
- the group of recovered non-relevant documents.

<sup>4</sup>www.todobr.com.br

<sup>5</sup>www.google.com

<sup>6</sup>www.cade.com.br

The collections TREC-7 and TREC-8 (D. A. Hull & Schutze 1996) (Hull 1998) (Hull 1999) present results of the measure of utility, using several values as parameters of utility, among them, 3 and 2 presented satisfactory results; where  $A = 3$  are used for the group of recovered relevant documents and  $B = 2$  is for the group of relevant documents not recovered, according to Equation 1.

$$Utilidade = (A) \times R_+ - (B) \times N_+ \quad (1)$$

### Recall-precision

We used the *Recall-precision* to test the accuracy of the filtering system when using the *CACM* collection.

Typically, the precision decreases as recall increases and vice-versa. A precision-recall graph is usually generated to graphically represent the relationship between the two measures.

Those measured, in general, Are used in systems that present its answers as an orderly group of documents. The evaluation is usually done by observed the evolution of the precision in function of the recall, represented through a recall-precision curve.

The expressions to calculate the recall-precision is:

$$Precision = \frac{\text{Number of relevant docs retrieval}}{\text{total of retrieval docs}}, \text{ and}$$

$$Recall = \frac{\text{Number of relevant docs retrieval}}{\text{total of relevant docs}}.$$

### Results obtained with the collection FCCCMG

Table 2 shows the utility results for 10 documents returned for each query evaluate by the user through the Equation 1. Where  $DRN1 = 0.66$ , indicate that, to 10% of all 198 query analyzed by the user, the relationship between relevant and not relevant documents without profile reformulation were 0.66 and  $DRN1 = 1.77$ , indicate that, to 20% of all 198 query analyzed by the user, the relationship between relevant and not relevant documents without profile reformulation were 1.77.

The obtained results are shown in the Table 2, according to the labels below:

- $DRN1$  : shows the Utility between relevant and not relevant documents, of the 10 evaluation documents by the user, for percentages described in the Table 2 with empty profiles.
- $DRN2$  : shows the Utility between relevant and not relevant documents, of the 10 evaluation documents by the user, for percentages described in the Table 2 after the first reformulation of profiles, repeating the same query.
- $DRN3$  : shows the Utility between relevant and not relevant documents, of the 10 evaluation documents by the user, for percentages described in the Table 2 after several reformulation of profiles, repeating the same query.
- $DRN$  : shows the Utility between relevant and not relevant documents, of the 10 evaluation documents by the user, for percentages described on Table 2 after several reformulation of profiles, using a new query.

Percent (%)	DRN1	DRN2	DRN3	DRN
10	0,66	1,77	2,89	1,55
20	1,77	3,73	5,51	3,28
30	4,71	6,58	7,68	5,79
40	6,65	8,71	9,47	7,66
50	6,14	8,02	8,82	7,04
60	6,34	8,33	8,85	7,34
70	6,04	7,46	8,36	6,55
80	6,39	8,00	8,87	7,03
90	6,95	8,71	9,53	7,65
100	6,51	8,14	9,07	7,16

Table 2: Utility results.

Next graph shows the results obtained in the execution of the query supplied by the user, taking in consideration the reformulation or not of the user's profile.

In Figure 3, we have the results of the query, through the information submitted by the students involved in the evaluation of the filtering system, according to data presented in Table 2.

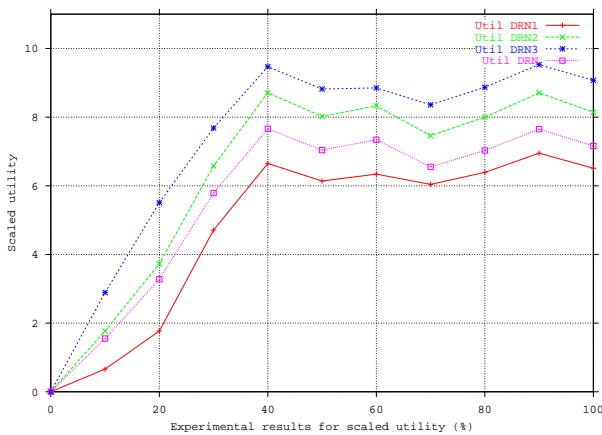


Figure 3: Representation of the measure of utility.

It was verified that a query with reformulated profile obtained a performance 25,3% greater if compared with the same query without reformulated profiles according to Table 2 columns *DRN1* and *DRN2*. However, after another reformulation using the same query the performance was 11% greater if compared with the same query after the first reformulation, according to Table 2 columns *DRN2* and *DRN3*.

The performance of a query with new terms over a reformulated profile is 9% greater to the obtained through the execution of the same query, according to Table 2 columns *DRN1* and *DRN*.

The Figure 4 shows the average performance of the relevant and not relevant documents, retrieves after successive submissions of querying and reformulated of profiles. With this graphic, we can make conclusion that, as the experiments are executing, consistent improvement in scaled utility for the amount of documents classified as relevant can be observed until it stabilized. However, we having a change in

the amount of non-relevant documents, which begins with a high rate and decreases while the execution of the experiments are happening.

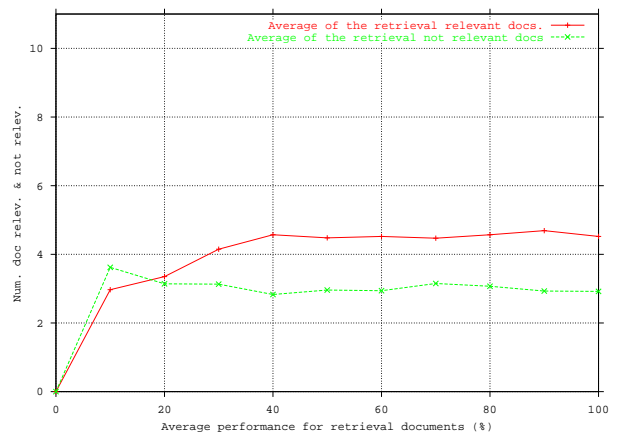


Figure 4: Average of recovered relevant and not relevant documents.

## Results obtained with the *CACM* collection

The experiments with the *CACM* collection were addressed in the following way: the terms of the query were submitted to the system. The system return the relevants documents to be analyzed and evaluated. First document relevant retrieval would be use by the system for the reformulation. That interaction happened for several times, increasing to each repetition the number of relevants documents used for the reformulation. For this collection was submitted 12query and 36 profiles reformulation.

The Tabela 3 shows results of experiments that we did with the collection *CACM*, described accordingly labels below:

- Case 1: medium precision for 11 points of recall 0,288, after the execution of 12 query.
- Case 2: medium precision for 11 points of recall 0,293, after the first profile reformulation.
- Case 3: medium precision for 11 points of recall 0,284, after several profile reformulation, submitting a new query.
- Case 4: medium precision for 11 points of recall 0,311, after several profile reformulation and aleatory reuse of a query previously submitted.

The Figure 5 shows the medium precision, considering the result obtained by the vector space model, and the submission of a new query after the reformulation of the profile, or the aleatory reuse of a query previously accomplished. As it can be observed, the performace of a query with new terms aguest a reformulated profile doesn't shows significant improvement, taking to believe that, as the relevants documents are judged without partiality, the application of the reformulation by itself do not improves the performace of the filtering system, once that the results obtained by the

Recall (%)	Case 1	Case 2	Case 3	Case 4
0	0,495	0,510	0,493	0,560
10	0,437	0,443	0,439	0,477
20	0,391	0,391	0,397	0,389
30	0,339	0,339	0,340	0,341
40	0,279	0,299	0,260	0,285
50	0,219	0,219	0,218	0,223
60	0,175	0,175	0,162	0,174
70	0,142	0,142	0,134	0,148
80	0,065	0,065	0,068	0,095
90	0,049	0,049	0,045	0,071
100	0,005	0,005	0,000	0,006

Table 3: Medium Precision for several recall levels.

vector-model in that type of situation are very good already. However, when compared with query executed previously, the improvement in the precision is significant for the base levels of recall.

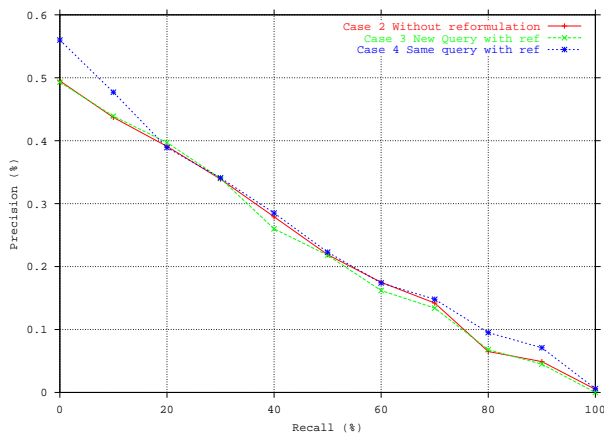


Figure 5: Curves of precision-recall.

## Conclusion

In this work a filtering system was implemented based on the vector-space model with profile reformulation, applying the algorithm of word contribution (K. Hoashi & Hashimoto 2000a) (K. Hoashi & Hashimoto 2000b). The use of the vector-space model for the query processing and the filtering of the result of querying considering the user's profile is justified because the model is efficient and of simple implementation. The results presented in this article demonstrate that, in place where possible the control and the interaction with the user, the use of profiles turns the results more efficient for search engines, increasing the degree of the users' satisfaction, the proposed methods are a viable alternative for implement information retrieval systems.

The experimental results obtained with the simulation of the filtering system show that the accomplishment of reformulation of the profile improves 25,3% on average the quality of the retrieval document when compared with the same

querying without profile reformulation, and in 9% compared with the submission of a new query.

## References

- D. A. Hull, J. O. P., and Schutze, H. 1996. *Method combination for document filtering*. Proceedings of the 19th Annual International ACM SIGIR CRDIR.
- G. Salton, C. Y., and Yu, C. 1975. *A theory of term importance in automatic text analysis*, volume 1. Journal of the ASIS.
- G. Salton, C. Y., and Yu, C. 1982. *Term weighting in information retrieval using the term precision model*, volume 24. Journal of the ACM.
- Hull, D. A. 1998. *TREC-7 Filtering track: description and analysis*. Proceedings of the Seventh Text Retrieval Conference (TREC-7).
- Hull, D. A. 1999. *TREC-8 Filtering track: description and analysis*. Proceedings of the Seventh Text Retrieval Conference (TREC-8).
- I. Witten, A. M., and Bell, T. 1999. *Managing gigabytes - compressing and indexing documents and images*. Morgan Kaufmann.
- K. Hoashi, K. M. N. I., and Hashimoto, K. 1999. *Experiments on the TREC-8 filtering track*. TRACK.
- K. Hoashi, K. M. N. I., and Hashimoto, K. 2000a. *Document filtering method using non-relevant information profile*. Athens, Greece: ACM SIGIR.
- K. Hoashi, K. M. N. I., and Hashimoto, K. 2000b. *Query expansion method on word contribution*. Berkeley CA-USA: ACM SIGIR.
- Krulwich, B., and Burkey, C. 1997. *The infofinder agent: learning user interests through heuristic phrase extraction*. Number 22-27. IEEE Expert.
- Ribeiro-Neto, B., and Baeza-Yates, R. 1998. *Modern information retrieval*. Addison Wesley.