Extract-biased pseudo-relevance feedback

Wilson dos S. Batista Junior and Lucia Helena Machado Rino

Departamento de Ciência da Computação – UFSCar
Rodovia Washington Luiz, km 235
São Carlos – SP (Brazil)
{wilson_batista,lucia}@dc.ufscar.br

Abstract

Successfully retrieving a web document is a twofold problem: having an adequate query that can usefully and properly help filtering relevant documents from huge collections, and presenting the user those that may indeed fulfill his/her needs. In this paper, we focus on the first issue – the problem of having a misleading user query. The aim of the work is to refine a query by using extracts instead of full documents. Extracts, in our context, are actually summaries of documents of a hitlist produced by an extractive automatic summarizer. Automatic summarization of single and multi-documents is explored through GistSumm, our **Gist Summ**arizer, which is based on the gist of a document, hence its name. Results on pseudo-relevance feedback for the Portuguese CHAVE collection show that gist-based extracts may improve information retrieval.

Keywords: Automatic Summarization for Information Retrieval, Pseudo-Relevance Feedback, Blind Feedback.

1. Introduction

Information Retrieval (IR) aims at presenting the user a set of documents that may satisfy her/his information needs. Usually, IR systems do so through searching for the words presented by the user in the query in one (or several) document collection. However, it is quite common that linguistic choices of both, the authors of the documents and the users, in their queries, do not coincide, even if they convey the very same concepts or referring entities. This is quite troublesome when the system uses direct matching between the query and the document words to produce a list of query-related documents. Usually, relevant ones may be unrecognized, even when the user poses an adequate query. In this case, adequately indexing relevant documents may be problematic due to its length: few words may be not enough to help filtering out irrelevant documents; extending the query may be impractical for the user, mainly if s/he is not knowledgeable enough to improve its expressiveness or comprehensiveness. In sum, two bottlenecks are posed, for the system to identify relevant documents: the way queries are built and the way query and documents are matched.

The Relevance Feedback or RF [1] method has been proposed to tackle the above problems by adding information to the original queries and, thus, improve IR. In RF, information indicated by the user is added to the query in the following way: after the system presents her/him a hitlist, i.e., a list of ordered retrieved documents, s/he pinpoints those documents that are of interest. Their relevant content gives rise, thus, to new terms that are added to the original query, yielding a new, modified, query. Then, the IR process is repeated, to produce another, final hitlist. This process is quite effortful because it still depends on the user judgment. To overcome that, the Pseudo-Relevance Feedback (PRF) method is intended to give feedback to the system without the user intervention. By adopting

ISSN: 1137-3601.

an artificial, selective strategy, it pinpoints which documents should be relevant to improve the original query. Usually, the top retrieved ones are considered.

Following the PRF idea given above, we present in this paper an IR system that considers extracts of documents of a hitlist for PRF, instead of considering the hitlist documents themselves. In our context, extracts are summaries produced entirely on an extractive basis, i.e., on a *copy-and-paste* strategy of full text segments. These, in turn, are the ones pinpointed as the most significant text segments of a document, or the ones that present the highest similarity measures.

The extracts are automatically built by GistSumm [2], an automatic summarizer that uses the gist sentence of a document to produce its extract. GistSumm can be customized to generate either generic or query-based extracts, derived from mono or multi-documents. Generic extracts are those resulting from detecting and summarizing the main topics of a full document and usually mirror its author viewpoint. Query-based ones are those that convey only sentences related to the user query components, narrowing the choice of topics to the user preferences. Extracts of multi-documents mirror the main topics across them. Those three distinct types of extracts have been considered for PRF in our work. Results show that query-based extracts are more effective than generic ones and multi-document query-based ones still outperform single query-based ones.

In what follows, first we outline work on summary-biased indexing and PRF (Section 2), then we describe the proposed IR system architecture (Section 3). An assessment on IR based upon extracts-based PRF follows in Section 4, which considers the Portuguese CHAVE corpus used in CLEF2004 (http://www.clef-campaign.org/). Final remarks are presented in Section 5.

2. Indexing and PRF through Automatic Summarization

According to [3], in IR the most significant words of a document must be used as index terms. We generalize that approach by considering the most significant sentences of a document, in order to proceed to indexing, instead of considering only independent words. Our argument here is that, once AS (Automatic Summarization) aims at producing automatic summaries that convey the main information of the corresponding source, the sentences wording in a summary will also be significant for indexing. For the same reason they can also be used to tackle PRF into any query provided by the user.

Concerning sentence indexing, a parallel with full-document indexing can be made: although documents can yield more profitable hitlists, reducing the index through summaries can be almost as effective as those for IR, when they indeed mirror the main information of the full documents. Concerning the use of summaries in PRF, instead of single words, the resulting query may yield a better filtering of relevant documents than the original user query. The reason for this is that summaries usually convey the most important information of the documents signaled in the first hit. Moreover, they shall present interconnected information, which may help matching more significant portions of the intended retrieved documents.

Several researchers pinpoint the appropriateness of considering summaries as good means for indexing: at the same time that they aim at preserving fundamental information, they aim at discarding peripheral document segments. For example, [4] and [5] use generic summaries for document indexing. They adopt two strategies to extract sentences from the documents: considering only lead sentences or selecting those with high TF-IDF [6]. Additionally, [5] also combine both approaches, yielding a third strategy for summarization.

Brandow et al. [4] used extracts of fixed size (60, 150 and 250 words) for indexing. Their assessment showed that IR was improved in precision, but got worse in recall. Their recall measures also varied significantly according to the compression rate: the larger the number of words in an extract, the higher the recall. According to our view, the reasons for missing important information may be the following: (a) inadequate compression rates prevent extracts to convey important information; (b) extractive methods do not perform well in filtering significant terms for indexing; (c) information that may be relevant to a given query can be peripheral to the document and, thus, could be ignored when producing its summary.

Partly, the above evaluation for indexing was not worthy due to the use of the Boolean model for matching. To overcome that, Sakai and Sparck Jones [5] adopted the probabilistic model instead. They carried out several assessments with different compression rates and also used the author abstracts of the documents. As a result, they concluded that abstracts and extracts for indexing were as effective as full text indexing, in the search for highly relevant documents. However, for those documents that would be considered relevant in TREC assessments, their approach did not perform well. This was due to the fact that some query topics had marginal relevance in the actual relevant documents. Generic extracts, in this case, could not overcome that problem and, thus, they could not accomplish the indexing task. Using TREC-7 and TREC-8 data, Sormunen [7] also verified the same: ca. 33% of the relevant documents for 38 topics had marginal relevance. These findings make evident the inadequacy of using generic extracts for indexing because they decrease the IR performance (see reason (c) above).

So far, we made it clear that using summaries for PRF is in opposition to using them for indexing, in that they do not actually act on retrieving relevant documents. Instead, they only aim at providing more information to improve the query, which is then used by the traditional document indexing and retrieval processes. This approach was undertaken in both [8] and [5]. Lam-Adesina and Jones used both query-based and generic summaries. Sakai and Sparck Jones, as already referred to, used only generic ones for PRF, besides using them for indexing.

To generate their extracts of single documents with 85% compression rate, Lam-Adesina and Jones selected highly scored sentences according to the following metrics: word frequency, presence of title words in a sentence, sentence position in a document, and the presence of query words in a sentence. According to them, specifying the compression rate is rather delicate: there is a compromise between the length of an extract and the inclusion of terms that can be beneficial to IR. In turn, the number of non-relevant terms must be minimum, if they are not totally suppressed from the extract. So, determining the proper terms to expand the query is the main issue here.

In their experiments, the first five retrieved documents were used to produce the extracts and, then, determine those terms. This process was based on the rsv measure [9]. Scoring the extracts terms considers their distribution in both the pseudo-relevant documents and in the full collection. Those highly scored in the former set and non-significantly scored in the latter one get higher scores. The top twenty terms are thus used to expand the queries in the PRF process. The results showed that:

(i) query expansion using selected terms from query-based extracts improved retrieval effectiveness;
(ii) query-based extracts could also be efficient when the source documents also comprised irrelevant documents;
(iii) selecting terms for query expansion was more effective than using all the terms of a document, or terms that were conveyed only by generic extracts.

Differently from the above approach, Sakai and Sparck Jones used generic extracts and abstracts derived from five pseudo-relevant documents for query expansion. The pseudo-relevant documents were withdrawn based upon either lead sentences or those selected through their TF-IDF measures[1]. Generic extracts were generated under varied compression rates (95%, 90%, 70%, and 50%). However, document titles were also included, exceeding such rates. Both extracts and abstracts (e.g., author summaries, or summaries produced by an actual writing task, instead of an extractive method) were considered. Their results showed that generic extracts and abstracts improved IR effectiveness, mainly when they were used for indexing in the first run (i.e., to produce the hitlist that gives rise to AS) and, subsequently, a full document indexing took place for the final search.

3 IR Based on Pseudo-Relevance Feedback

Similarly to the above approaches, ours focuses upon using extracts produced by GistSumm to refine a user query for IR. A query refinement differs from expansion in that terms are not only included in the query, but may be also promoted or demoted. In this section first the architecture of the proposed IR system is presented, followed by the description of the AS *scenario* used for PRF.

[1] *They are so-called pseudo-relevant documents because it is not the user who pinpoints their relevance, but artificial methods based on decisions as the ones referred to here.*

3.1 The IR System Architecture

After pre-processing a collection of documents and the posed query (Q), our IR system content vector (Q') is matched against the index file to proceed either to PRF or to the intended IR task (Figure 1). Vectoring is accomplished by other usual pre-processing tasks (stopwords removal and stemming), yielding document indexing based on the stems of significant lexical items. A Portuguese-dependent stemmer [10] has been used at this point. Actually, the very same matching is carried out twice and independently: first to refine the query based upon the extracts produced by GistSumm (option yes in the figure), then to actually retrieve the final documents and conclude the search (option no).

Matching is based on the Dice coefficient. Firstly, a weight is calculated for each term of both, the document and the query, based upon the respective frequencies. The first matching retrieves five documents with the higher Dice similarity with the query, to proceed to PRF. This strategy mirrors the ones described in the previous section, i.e., [5] and [8]. Then, GistSumm generates extracts for those documents under a 90% compression rate. Following [8], we also used the rsv measure to select extract terms for PRF. Ten terms with more high rsv value were considered.

The rsv value for a given term i is calculated according to the following formula:

$$\mathrm{rsv(i)} = \mathrm{r(i)} * \mathrm{rw(i)} \quad (1)$$

for r(i) as the number of pseudo-relevant documents that convey term i and rw(i) as the relevance weight [11] of the same term, which is defined as

$$\mathrm{rw(i)} = \log\left(\frac{(\mathrm{r(i)}+0.5)*(\mathrm{N}-\mathrm{n(i)}-\mathrm{R}+\mathrm{r(i)}+0.5)}{(\mathrm{n(i)}-\mathrm{r(i)}+0.5)*(\mathrm{R}-\mathrm{r(i)}+0.5)}\right) \quad (2)$$

for n(i) being the total number of documents conveying term i; R, the total number of relevant documents for Q (five, in our case); and N, the total number of documents.

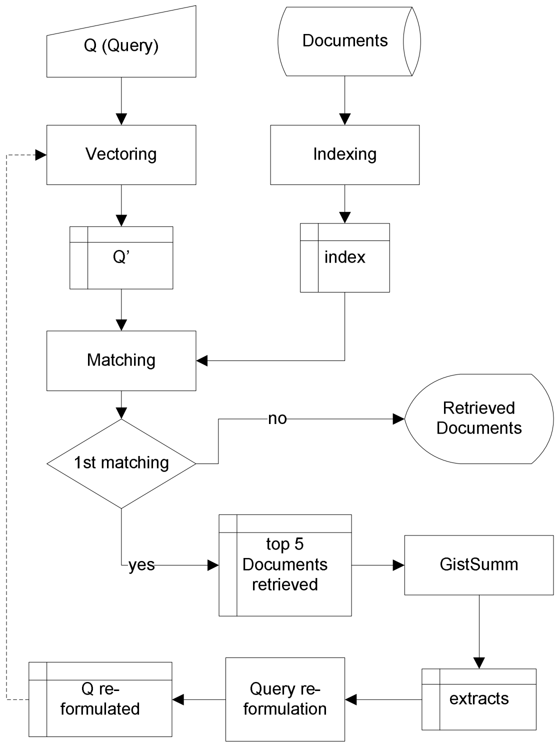

Figure 1. The IR System Architecture

Once selected the ten terms, the query is refined by reweighting its original terms. This is carried out according to the following formula [12], which shows that the new weight of term i, indicated by w(i)QRe, is based upon the weight of the non-refined query, i.e., w(i)QnonRe, and its rsv value, rsv(i):

$$\mathrm{w(i)QRe} = \alpha * \mathrm{w(i)QnonRe} + \beta * rsv(i) \quad (3)$$

The coefficients α and β are defined empirically and aim at stressing that query terms should be more important than extract terms. In our experiments, the values of α and β were 1 and 0.2, respectively. After refining the query, the second matching takes place to retrieve the intended documents.

3.2 The AS Scenario for IR

As already mentioned, GistSumm is an extractive automatic summarizer based on the gist of a source text. The gist is assumed to be conveyed by just one sentence of the document. To build an extract, GistSumm first determines the gist sentence, including it in the extract. Then it adds other sentences that better correlate to the gist one. As usual, the choice of extra sentences is constrained by the intended compression rate. GistSumm has been chosen to explore PRF in our research due to its usefulness measure at DUC'2003 (http://duc.nist.gov/): generic extracts of single documents with an average length of ca. 20 words yielded its 3.12 usefulness score in a 0-4 scale (0: summary of no use; 4: summary as good as the full text).

Our hypothesis in using GistSumm for PRF was that the terms of the extracts would be more expressive than single terms extracted from full documents, yielding a more effective retrieval. Clearly, such a refinement relies on GistSumm main assumption, i.e., that extracts convey the most important information related to the gist of their corresponding documents.

Currently, GistSumm also provides facilities to perform query-based single or multi-documents AS. For this reason, PRF is investigated under those three types of extracts, as already mentioned in the introduction. The construction of the corresponding extracts is further detailed below.

3.2.1 Generic Extracts of Single Documents

To generate generic extracts, GistSumm signals the gist sentence as the most frequent one, amongst all the sentences of the source text. Frequency calculation is based upon the words stems. Additional sentences to include in the extract are those that convey coinciding word stems to those of the gist sentence.

3.2.2 Query-Based Extracts of Single Documents

Differently from the former approach, to generate query-based extracts of single documents the gist sentence is determined through the cosine similarity measure [13]. It is the sentence of the source text that is most similar to the posed query, i.e., the one that achieves the biggest similarity score. This measure has nothing to do with the Dice coefficient used at the indexing phase and has been adopted just because it is the similarity measure embedded in GistSumm, which is used as a blackbox system, in what concerns the AS methodology. Additional sentences of a source document are chosen to include in the extracts as before.

3.2.3 Query-Based Extracts of Multi-Documents

In this case, GistSumm performs more simply than proper multi-documents AS (e.g., [14] or [15]), but following similar work (e.g., [16]). It considers the set of every document in the collection as if it were a unique source document. Then, it summarizes such a source in the same way as it does in the query-based single approach, i.e., building an extract of a "single document". As such, clearly GistSumm embeds a quite rudimental multi-document AS procedure. It differs, and ignores, most of the more sophisticated multi-document AS proposals, such as [17].

The following abilities, suggested in [17], amongst others, deserve attention: (a) clustering, to group together both similar documents and passages that help finding relevant information; (b) targeting coverage adequacy, to deal with the main issues across documents; (c) minimizing redundancy, to recognize singularities across documents and convey only the most relevant passages in the summary; (d) identifying source inconsistencies (e.g., typos or incorrect information), to prevent their inclusion in a summary and, thus, avoid the decrease of the IR efficacy. From these, the only one that GistSumm aims to tackle is (b), in that it correlates information of all the documents with the gist sentence aiming at covering the main issues of the involved documents. However, since it does so considering that there is only one full document, coverage adequacy across documents cannot be actually guaranteed. Although documents are not clustered for AS, and, for this reason, ability (a) is not considered in GistSumm, our indexing approach undertakes it instead. Abilities (c) and (d) have been completely ignored in GistSumm.

4 Assessing the Extract-Biased PRF Approach

In assessing the extract-biased PRF approach, we used the CHAVE collection [18] to replicate the CLEF2004 monolingual Portuguese ad hoc track. The CHAVE corpus amounts to 55,070 articles of the Portuguese newspaper 'Público', comprising 50 different topics, with ca. 15 relevant documents per topic. However, for many topics there are at most three relevant documents, and for four topics there are no relevant documents at all.

The title and description fields of each topic were used to automatically build the queries. Five runs were used, hereafter named RDoc, RFGenS, RFQBS, RFQBM, RFFull-Doc[2]. Apart from RDoc, the other runs differ in the ways extracts are generated for PRF. Their description follows:

- RDoc is a baseline that does not perform PRF. So, it does not use extracts and relevant documents are retrieved and exhibited to the user in the first hit.
- RFGenS produces generic extracts of the top 5 documents for RF.
- RFQBS produces query-based extracts of single documents for RF.
- RFQBM produces only one query-based multi-document extract for RF.
- RFFullDoc, similarly to RDoc, differs from the others in that it does not use extracts, but full documents, to proceed to RF. In this case, the rsv measure is used to select the top 10 terms of the hitlist documents to refine the query[3].

When applicable, the extracts so defined are query-based because they are generated from the top 5 documents of the hitlist that is produced in the first matching through the user query (see Figure 1).

Table 1 shows retrieval precisions at 5 and 10 document cutoffs (P5 and P10) and the Mean Average Precision (MAP) for all the systems. Bracketed percentages signal variations between the baseline and the related run.

When considering the precision at 5 cutoff, the performance of the baseline is very good, although very few relevant documents are used. However, using query-based extracts outperforms the baseline, being PRF based on the multi-documents extract still better than that on single-documents extracts. Noticeably, PRF through generic extracts (RFGenS) yields the worst result. So, improving the query through query-based AS indeed improves IR of documents in the CHAVE collection.

The precision at 5 for the RFFullDoc system shows that using extracts is expressively better than using only the rsv measure for PRF. Both P10 and MAP values for all the assessed runs but RFQBS make evident that query-based extracts for PRF improve IR effectiveness. However, this was statistically supported only for the comparison between RFQBM and RDoc, whose differing numbers rendered statistically significant (sign test and t-test were used).

Although the precision values for RFQBM and RFQBS differed, it is noticeable that the terms selected for PRF by both presented little variation. This may be explained by the fact that the gist sentence of the multi-document extract coincided with the gist sentence of one out of the five extracts produced by the RFQBS system. Extracts of single documents may also have coinciding sentences with those that complement the gist sentence in the multi-document extract.

Comparing our results for P5 with the corresponding official runs in CLEF2004 (reminding that we mirrored the very same track), RFQBM occupied the 12th position, amongst 23 runs, some of them also considering PRF. In other words, at the same time that RFQBM performance signals a significant room for improvement, it occupies an expressive position, when compared to the performance of the other CLEF systems. Other runs on CLEF2004 were also carried out but were less significant than the ones reported here.

Run	**Precisions**		
	P5	**P10**	**MAP**
RFQBM	0.3720 (+4.5%)	0.2940 (+10.5%)	0.4371 (+4.0%)
RFQBS	0.3600 (+1.1%)	0.2840 (+6.8%)	0.4029 (-4.1%)
RDoc	0.3560	0.2660	0.4203
RFFullDoc	0.3551 (-0.2%)	0.2755 (+3.6%)	0.4273 (+1.7%)
RFGenS	0.3480 (-2.2%)	0.2880 (+8.3%)	0.4119 (-2.0%)

Table 1. Precision of 5 distinct runs

RFGenS results seem reasonable, if we consider that even for indexing, generic extracts did not yield significant improvements earlier (e.g., [4,5]) because very often documents considered relevant conveyed peripheral query topics. Certainly the use of generic extracts is recommended when the pseudo-relevant documents are highly relevant, as

---

[2] *Acronyms signal the main features of the systems, embedded in their letters, namely: Relevance Feedback (RF), generic (Gen) or query-based (QB), single (S) or multi-(M) documents extraction, a mere retrieval of full documents (RDoc) or RF through full documents (RFFullDoc).*

[3] *It should be observed that Figure 1 does not embed this alternative process.*

confirmed in [5]: generic extracts would be more expressive and, thus, PRF would improve IR. Using generic extracts at the indexing phase would also be more efficient, since the index file would be reduced, when compared with the full document index file.

In contrast with RFGenS and RFFullDoc, the use of GistSumm makes evident that query-based extracts improve IR, for the CHAVE collection. More importantly, using query-based extracts does not depend on the position of the query topics in the accessed documents. That is, even being peripheral, the topic conveyed by the query will be included in the extracts and so will it be in the refined query.

The difference between the RFQBM and RFQBS runs is explained by the fact that, to generate a multi-document extract, GistSumm calculates the gist sentence of the top 5 documents as a whole, whilst that sentence is most generally embedded in just one of them. For this reason, if there were an irrelevant document to the gist, its content would not be considered for PRF. This could be confirmed in our assessment: ca. 1.8 documents, out of the five pseudo-relevant ones, were indeed relevant. In this case, our approach has benefited from the choice of the CHAVE collection, because it has few relevant documents, as shown in CLEF2004.

When compared to using pseudo-relevant documents for PRF, using query-based extracts is still better because many of those documents may have few significant terms to refine the query. At the extreme case, none of them may actually be relevant. This could be confirmed for CHAVE topic 215, when our first matching took place. After refining the query, however, only one document relevant to that topic was retrieved, which was the one positioned third in the hitlist.

5 Final Remarks

We presented in this paper a pseudo-relevance feedback method for IR that relies on extracts produced by GistSumm, a system built at NILC[4] that is based on the gist of a document. The main idea behind the suggested approach is twofold: (a) keywords included in an index file are not enough to provide the means to retrieve relevant documents for the user, and (b) the user query may be significantly sparse or non-informative for identifying pertinent documents. Both features influence the effectiveness of an IR system. Using keywords, for example, may yield very large and non-discriminatory hitlists, besides reasoning on them being a very difficult problem [19].

In adopting an AS approach to PRF, we tried to overcome the above in that extracts already convey interdependent and significant terms of the corresponding documents. Actually, GistSumm has as its main premise the preservation of the main idea in the extracts, whilst excluding non-relevant information. So does our PRF approach.

GistSumm usefulness measure in DUC2003 has also been reassured in our assessment: the results suggest that extracts feature better the user’s interests than the original query. This is quite related to the positive cluster centroid of a document ([19], or [16]), if we consider that a gist sentence conveys the main idea of the document. As a consequence, gist-based extracts for PRF render a reformulated query with more satisfactory proximity with the main topics of the relevant documents. Oppositely, producing extracts still relies on blind hitlists, posing additional problems to the whole process. Despite this, our extract-biased PRF outperforms the ones that are not based on extracts, in spite of its simplicity. This may be due to limiting the index vector space to information that is closer to the document gist. However, our final results still suggest that the use of Portuguese data provides, after all, only new baselines for work in progress.

Our approach could be even improved if document titles were considered, which is not applicable to using the CHAVE collection. Generally, terms of title are related to the main topic of a document. Even the choice for this collection could be questioned, because, for many of its topics, there are at most three relevant documents; for four topics there aren’t relevant documents at all. So, other, more expressive, collections should be considered. In spite of this, extract-biased PRF through GistSumm still brought improvements to IR.

Improving our approach should also be explored by investigating the impact of the amount of non-relevant documents in the hitlist, which would demand scalable assessment procedures in the long run. More robust summarizers than GistSumm should also be considered.

[4] *Núcleo Interinstitucional de Lingüística Computacional (http://www.nilc.icmc.usp.br/nilc/index.html)*

Acknowledgments

The reported work has been funded by CAPES and CNPq.

References

[1] J.J. Rocchio Jr. Relevance feedback in information retrieval. In G. Salton (ed.): *The SMART Retrieval System* – Experiments in automatic document processing, Prentice Hall Inc, pages 324-336, 1971.

[2] T.A.S. Pardo, L.H.M Rino and M.G.V. Nunes. GistSumm: A summarization tool based on a new extractive method. *Lecture Notes in Artificial Intelligence 2721*, pages 210–218. Springer–Verlag, 2003.

[3] G. Salton. Associative document retrieval techniques using bibliographic information. *Journal of the ACM*, 10(4):440–457, 1963.

[4] R. Brandow, K. Mitze and L.F. Rau. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685, 1995.

[5] T. Sakai and K. Sparck-Jones. Generic summaries for indexing in information retrieval. *In the Proc. of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 01),* pages 190–198, 2001.

[6] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[7] E. Sormunen. Liberal relevance criteria of TREC: counting on negligible documents?. *In the Proc. of the 25th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 02),* pages 324–330, 2002.

[8] A.M. Lam-Adesina and G.J.F. Jones. Applying summarization techniques for term selection in relevance feedback. *In the Proc. of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 01),* pages 1–9, 2001.

[9] S.E. Robertson. On term selection for query expansion. *Journal of Documentation*, 46(4): 359–364, 1990.

[10] V.M. Orengo and C. Huyck. A Stemming Algorithm for Portuguese Language. *In the Proc. of the 8th Symposium on String Processing and Information Retrieval (SPIRE 2001),* pages 183–193, 2001.

[11] S.E. Robertson and K. Sparck-Jones. Relevance Weighting of Search Terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[12] C. Carpineto and G. Romano. Towards more effective techniques for automatic query expansion. *In the Proc. of the 3rd European Conference on Research and Advanced Technology for Digital Libraries*, pages 126–141, 1999.

[13] G. Salton. *The SMART Retrieval System* – Experiments in automatic document processing, Prentice Hall Inc, 1971.

[14] C. Lin and E. Hovy. Automated Multi-document Summarization in NeATS. In *Proceedings of the DARPA* (Human Language Technology Conference HLT), pages 50- 53, 2002.

[15] H. Daum III and D. Marcu. Bayesian Multi-document Summarization at MSE. In *Proceedings of the Workshop on Multilingual Summarization Evaluation at ACL 2005*. http://www.cs.utah.edu/~hal/docs/daume05mse.pdf, 2005.

[16] D.R. Radev, H. Jing and M. Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. *In the Proc. of the ANLP/NAACL Workshop on Summarization*, pages 21-30, 2000.

[17] J. Goldstein, V. Mittal, J. Carbonell and J. Callan. Creating and evaluating multi-document sentence extract summaries. *In the Proc. of the 9th international conference on Information and knowledge management (CIKM 00),* pages 165–172, 2000.

[18] D. Santos and P. Rocha. CHAVE: topics and question on the Portuguese participation in CLEF. In C. Peters and F. Borri (eds.): *Working Notes for CLEF 2004 Workshop*, Cross Language Evaluation Forum, pages 639–648, 2004.

[19] C. J. Van Rijsbergen. *Information Retrieval*. 2nd edn. Butterworths, London , 1999.