Highly Frequent Terms and Sentence Retrieval

David E. Losada and Ronald T. Fernández

Departamento de Electrónica y Computación,
Universidad de Santiago de Compostela, Spain
dlosada@dec.usc.es, ronald.teijeira@rai.usc.es

Abstract. In this paper we propose a novel sentence retrieval method based on extracting highly frequent terms from top retrieved documents. We compare it against state of the art sentence retrieval techniques, including those based on pseudo-relevant feedback, showing that the approach is robust and competitive. Our results reinforce the idea that top retrieved data is a valuable source to enhance retrieval systems. This is especially true for short queries because there are usually few query-sentence matching terms. Moreover, the approach is particularly promising for weak queries. We demonstrate that this novel method is able to improve significantly the precision at top ranks when handling poorly specified information needs.

Keywords: Information Retrieval, Sentence Retrieval, Term Frequency.

1 Introduction

Retrieval of sentences that are relevant to a given information need is an important problem for which an effective solution is still to be found. Many Information Retrieval (IR) tasks rely on some form of sentence retrieval to support their processes [13]. For example, question answering systems tend to apply sentence extraction methods to focus the search for an answer on a well-selected set of sentences or paragraphs [5]. In query-biased summarization, there is a large body of work dedicated to building summaries using sentences extracted from the documents [12]. Topic detection and tracking (TDT) is another task where the availability of effective sentence retrieval techniques is crucial in order to isolate relevant material from a dynamic stream of texts (e.g. news) [19]. In web IR, a good ranking of sentences, in decreasing order of estimated relevance to the query, can also act as a solid tool to improve web information access [25]. Sentence retrieval is therefore a core problem in IR research and advances in this area could potentially trigger significant benefits across the field.

We adopt the sentence retrieval problem as defined in the TREC novelty tracks [8,18,17]. Given a textual information need, an initial ranking of documents is produced using some effective retrieval method and, next, the systems should process the retrieved documents to locate the sentences that are relevant

N. Ziviani and R. Baeza-Yates (Eds.): SPIRE 2007, LNCS 4726, pp. 217–228, 2007.

to the information need[1]. This is a realistic task. For instance, typical web search engines have to address similar problems when building query-biased summaries.

The problem of sentence retrieval is far from easy. Sentences are very short pieces of texts and, therefore, the sentence-query matching process has difficulties. Many strategies, techniques and models have been proposed to address this problem. Despite the variety of the approaches investigated, very simple variations of tf/idf measures can be labeled as state of the art sentence retrieval methods [2,11].

We are not interested here in proposing a new model for sentence retrieval, but we will extend a competitive sentence retrieval method to incorporate the influence from terms which are highly frequent in the retrieved set of documents. This is inspired by well-known research in query-biased summarization [20] that estimated the set of significant words of a document using the number of term occurrences within each document. We adapt this intuition to our current retrieval scenario. Rather than focusing on a single document to estimate which words are significant, we will estimate the significant terms from the top retrieved documents. The set of highly ranked documents for a given query is a very valuable source of information and, ideally, it provides a vocabulary focused on the query topics. Hence, we compute term statistics globally in the retrieved set of documents and adjust the tf/idf scores to take into account the contribution from the most significant terms. In this way, sentences that do not match any query term can still be retrieved provided that they contain some significant terms.

In the experiments we show that the approach is simple but very effective. It outperforms significantly a competitive baseline under three different benchmarks. The method is able to improve both precision at 10 sentences and the F measure. We also compare the relative merits of this method against pseudo-relevance feedback, showing that our approach is much more robust, especially when queries are poor.

The rest of the paper is organized as follows. Section 2 reviews some papers related to our research. In section 3 we explain the foundations of the method proposed and the evaluation conducted is reported in section 4. The paper ends with some conclusions and future lines of work.

2 Related Work

Sentence retrieval is an active research area where many researchers have proposed different alternatives to tackle the problem. Many studies applied query expansion either via pseudo-relevance feedback [6] or with the assistance of a terminological resource [27,9]. Nevertheless, the effect of pseudo-relevance feedback is very sensitive to the quality of the initial ranks and it is quite difficult to apply it effectively across different collections and types of queries [26]. More

[1] The TREC novelty tracks propose two different tasks: retrieval of relevant sentences and retrieval of relevant and novel sentences but we are only interested here in the first task.

evolved methods, such as selective feedback [1], are more stable but they usually require training data. On the other hand, to expand queries with synonyms or related terms from a lexical resource is problematic because noisy terms can be easily introduced into the new query [21]. Moreover, a large terminological resource, with good coverage, is not always available. As a matter of fact, lexical expansion was not significantly better than purely statistical expansion methods in the sentence retrieval tasks of the TREC novelty tracks [8,18,17]. Other expansion approaches based on co-occurrence data have been proposed. For instance, in [27] the authors expand the query with terms that co-occur highly with the query terms in the retrieved documents. Co-occurrence statistics from a background corpus have been applied in [15]. Nevertheless, there is not much evidence that these approaches can outperform the standard pseudo-feedback methods.

Rather than expanding queries with new terms, other studies have focused on improving the matching process by analyzing carefully the nature of the sentence components. In this line, in [11], patterns such as phrases, combinations of query terms and named entities were identified into sentences and the sentence retrieval process was driven by such artifacts. Although this technique was effective for detecting redundant sentences, it was not significantly better than a regular tf/idf baseline for finding relevant sentences. In [7], terms in sentences were categorized into four query-based categories, namely: highly relevant, scarcely relevant, non-relevant and highly non-relevant. Nevertheless, this classification was mainly guided by the topic subfield in which the term occurs (title, descriptive, narrative). Unlike this work, our approach uses the number of term occurrences in the retrieved set of documents as the main factor to estimate the significance of a term.

The work by Zhang and his colleagues [28] deserves special attention. They presented a high performing sentence retrieval system which combined query expansion and sentence expansion using Wordnet. Besides the linguistic-based expansion, they proposed two *impact factor* measures to estimate the significance of the query terms. The utility of these measures in removing some query terms (low impact factor words) was empirically evaluated but no benefits were observed. One of the impact factor measures was based on the frequency of the term in relevant documents. The intuition was that highly frequent terms are very significant and, therefore, they should not be removed from the query. We pursue a similar idea but with a different objective. Rather than polishing queries, our aim is to define a measure of how central a sentence is in the context of the retrieved documents. The highly frequent words in the retrieved set are the foundations of this measure.

In query-biased summarization, the notion of significant word (term with high frequency in the document) was applied for sentence scoring purposes [20]. This was combined with aspects such as term location (the presence of a term in the title of the document or in the leading paragraphs produced a positive weight) and some query-oriented weights to produce a final score for the sentences. Our method to adjust the basic tf/idf method is inspired by the notion

of significant word applied in [20] but, rather than estimating the significant terms for a single document, we are interested in computing the significant words in a set of documents. As a matter of fact, the set of highly ranked documents for a given query is a good container of topic-related terms, as demonstrated by the success of different expansion methods for document retrieval [23]. Some papers have applied the results of [20] to other environments. For instance, in [24], the effects of query-biased summaries in web searching were studied. The summaries are constructed from sentences scored using title, location, relation to the query and text formatting information. However, the notion of significant words, as proposed in [20], has not received major attention since then. We believe that it is interesting to study the effect of such notion for sentence retrieval.

3 Highly Frequent Terms and Sentence Retrieval

The main component in our method consists of extracting a set of significant words from the set of documents retrieved for a given query. Ideally, these terms characterize well the query topics and, therefore, they can be used for improving sentence retrieval. Given a set of highly ranked documents, R_q, terms occurring more than a certain number of times in this set are collected into a *set of significant terms* (ST_q) as follows[2]:

$$ST_q = \{t \in V_q | tf_{t,R_q} > mno\} \tag{1}$$

where V_q is the set of unique terms appearing in R_q, tf_{t,R_q} is the term count of t in R_q ($tf_{t,R_q} = \sum_{D \in R_q} tf_{t,D}$) and mno is a parameter that determines the minimum number of occurrences required for a term to be considered as significant.

In order to select the sentences within the retrieved set of documents which are relevant to the query, a combination of a regular tf/idf score and a significant term score is applied. The tf/isf score (isf stands for inverse *sentence* frequency) is:

$$tf\_isf(s,q) = \sum_{t \in q} log(tf_{t,q} + 1) log(tf_{t,s} + 1) log(\frac{n+1}{0.5 + sf_t}) \tag{2}$$

where sf_t is the number of sentences in which t appears, n is the number of sentences in the collection and $tf_{t,q}$ ($tf_{t,s}$) is the number of occurrences of t in q (s). This formula was applied successfully in TREC novelty tracks for sentence retrieval purposes [2]. Along this paper this method will be referred to as TF/ISF.

Given a query q, the significant term score of a sentence s, $htf(s,q)$, is defined as:

$$htf(s,q) = \sqrt{|V_s \cap ST_q|} \tag{3}$$

[2] We include q as a subindex to stress the fact that this set is query-dependent. In our experiments, R_q was fixed to be the set of documents supplied by the track's organizers (set of highly ranked documents associated to each query).

where V_s is the set of unique terms appearing in s. The score depends on the number of highly frequent terms in the sentence and the square root was introduced to get the tf_isf and htf scores on a similar scale and, therefore, combine them adequately.

The sentence-query similarity is simply defined as the sum of the tf_isf and htf scores. Hereafter, this sentence retrieval method will be referred to as the HTF method. The combination method applied here is intentionally simplistic and we consider it as a first attempt to mix these factors. More evolved (and formal) methods will be studied in the near future.

Sentences with a poor overlap with the query can still be retrieved provided that they contain terms which are significant in the retrieved set. This method promotes the retrieval of sentences which are somehow *central* for the query topics. It is also a way to tackle the well-known vocabulary mismatching problem in retrieval engines. The notion of term significance is equated here to high term frequency in the retrieved set. Note that in our experiments stopwords were removed. This is important because, otherwise, the ST_q sets would be likely populated by many common words, which would introduce much noise. Of course, many other alternatives could have been proposed to estimate term significance (e.g. based on successful methods such as Local Context Analisis [26] or Divergence from Randomness [3]). Nevertheless, the present study is focused on a simple term frequency-based method which has proved successful in query-biased summarization [20]. A complete comparison of different alternatives to estimate term significance for sentence retrieval is out of the scope of this paper.

The parameter mno determines the size of the set ST_q and, consequently, influences directly the weight of the htf component in the similarity formula. Low mno values (e.g. $mno < 5$) do not make sense because the set ST_q would contain most of the terms in the retrieved set. This set would have many non-significant terms and the htf component in the formula above would be merely promoting long sentences. On the other hand, very high mno values are not advisable either because ST_q would be very small (empty in the extreme case) and, therefore, the method would be roughly equivalent to the basic TF/ISF technique. In this work, we test the effectiveness of our method with varying number of minimum occurrences and analyze the sensitiveness of the method with respect to the size of the set of significant terms. In the future, we will also study other alternatives, such as query-dependent mno values estimated from R_q.

The HTF method can be actually regarded as a form of pseudo-relevance feedback (at the document level) aimed at estimating the importance of the sentences within a retrieved set (with no query expansion). Term selection is done before sentence retrieval and the selected terms are not used for expansion (the terms are used for estimating the significance of a sentence). On the other hand, the standard query expansion via pseudo-relevance feedback or local co-occurrence is strongly sensitive to the quality of the original query [26]. If we use a few top ranked sentences to expand the query then it is likely that we end up introducing some noise in the new query. We expect that the highly

frequent terms approach is more robust because, rather than doing first a sentence retrieval process to get some new terms, it analyzes globally the set of retrieved documents to locate important sentences. It seems reasonable to think that the behaviour of this method is more stable. One can rightly argue that an original poor query will also be harmful in our approach because the quality of the document ranking will be low. Nevertheless, once we have the rank of documents, the adjustment that we propose (htf score) does not involve the query text. On the contrary, pseudo-relevance feedback methods apply subsequently a sentence retrieval process from the query terms and, hence, they are much more dependent on the quality of the queries (the query text is used twice). Other evolved expansion methods, such as Local Context Analysis [26], which has proved to be more effective and robust than pseudo-relevance feedback for document retrieval, require also a second usage of the original query text for retrieving passages and, next, phrases are selected from the top passages to expand the query. The process is thus rather complicated whereas the method applied here is much simpler.

4 Experiments

The performance of the HTF method has been tested using three different collections of data. These datasets were provided in the context of the TREC-2002, TREC-2003 and TREC-2004 novelty tracks [8,18,17]. There are no newer TREC collections suitable for our experiments because we need relevance judgments at the sentence level. This sort of judgments is only available in the novelty track, whose last edition took place in 2004. The novelty track data was constructed as follows. Every year there were 50 topics available. In TREC-2002, the topics were taken from TRECs 6, 7 and 8 (the complete list of topics chosen for the novelty track can be found in [8]). In 2003 and 2004, the topics were created by assessors designated specifically for the task [18,17] (topics N1-N50 and N51-N100). For each topic, a rank of documents was obtained by NIST using an effective retrieval engine. In 2002 and 2003 the task aimed at finding relevant sentences in relevant documents and, therefore, the ranks included only relevant documents (i.e. given a topic the set of relevant documents to the topic were collected and ranked using a document retrieval engine). On the contrary, the TREC-2004 ranks contained also irrelevant documents (i.e. the initial search for documents was done against a regular document base, with relevant and irrelevant documents). Note that this means that the irrelevant documents are close matches to the relevant documents, and not random irrelevant documents [17]. In any case, the ranks of documents contained at most 25 relevant documents for each query.

The documents were segmented into sentences, the participants were given these ranks of sentence-tagged documents and they were asked to locate the relevant sentences. The relevance judgments in this task are complete because the assessors reviewed carefully the ranked documents and marked every sentence as relevant or non-relevant to the topic. In TREC-2002, very few sentences were

judged as relevant (approximately 2% of the sentences in the documents). In TREC-2003 and TREC-2004 the documents were taken from the AQUAINT collection and the average percentage of relevant sentences was much higher than in 2002 (approximately 40% in 2003 and 20% in 2004). This therefore shapes an assorted evaluation design in which we can test the HTF method under different scenarios and conditions. We focus our interest on short queries, which are by far the most utilized ones, especially in environments such as the web [16]. In our experiments, short queries were constructed from the title tags of the TREC topics.

We used two different evaluation measures: precision at 10 sentences retrieved and the F-measure, which was the official measure in the TREC novelty experiments[3]. Regarding statistical significance tests, we applied two different tests, the t-test and the Wilcoxon test, and we only concluded that a given difference between two runs was significant when both tests agree (with a 95% confidence level).

To ensure that the baseline was competitive we ran some initial experiments with other popular retrieval methods. We experimented with Okapi BM25 [14] and a Language Modeling approach based on Kullback-Leibler Divergence (KLD) as described in [10] (with Dirichlet smoothing). The performance of BM25 is influenced by some parameters: $k1$ controls the term frequency effect, b controls a length-based correction and $k3$ is related to query term frequency. We tested exhaustively different parameter configurations ($k1$ between 0 and 2 in steps of 0.2, b between 0 and 1 in steps of 0.1 and different values of $k3$ between 1 and 1000). Similarly, we experimented with the KLD model for different values of the μ constant, which determines the amount of smoothing applied (μ = 10, 100, 500, $1k$, $3k$, $5k$). Results are reported in Table 1. A run marked with an asterisk means that the difference in performance between the run and TF/ISF is statistically significant. In all collections, there was not statistically significant difference between the TF/ISF run and the best BM25 run. We also observed that BM25 was very sensitive to the parameter setting (many BM25 runs performed significantly worse than TF/ISF). On the other hand, KLD was inferior to both TF/ISF and BM25. These results reinforced previous findings about the robustness of the TF/ISF method [2,11] and demonstrated that this method is a very solid baseline.

We also tried out different combinations of the standard preprocessing strategies (stopwords vs no stopwords, stemming vs no stemming). Although there was no much overall difference, the runs with stopword processing and no stemming were slightly more consistent. Since the TF/ISF method takes the idf statistics from the sentences in the documents available for the task (which is a small set of sentences), we were wondering whether better performance may be obtained using idf data from a larger collection. To check this, we indexed a large collection

[3] The F-measure is the harmonic mean (evenly weighted) of sentence set recall and precision. In the TREC-2002 experiments, we computed the F-measure using the top 5% of the retrieved sentences and in the other collections we used the top 50% of the retrieved sentences. Similar thresholds were taken in TREC experiments [2].

Table 1. Comparing different baselines

	TF-ISF	BM25	KLD
	TREC-2002		
best run		$k1 = .4, b = 0, k3 = 1$	$\mu = 3000$
P@10	.19	.19	.16
F	.188	.190	.172*
	TREC-2003		
best run		$k1 = .6, b = 0, k3 = 1$	$\mu = 1000$
P@10	.74	.76	.73
F	.512	.512	.510*
	TREC-2004		
best run		$k1 = .2, b = 0, k3 = 1$	$\mu = 500$
P@10	.43	.44	.41
F	.370	.371	.369

Table 2. Evaluation results

		HTF, mno=				PRF, # exp terms=			
	TF/ISF	7	10	15	20	5	10	20	50
		TREC-2002							
P@10	.19	.23*	.22*	.22	.22	.19	.20	.17	.20
F	.188	.197	.197	.197	.192	.190	.181	.181	.178
		TREC-2003							
P@10	.74	.78*	.78	.79*	.79*	.78*	.79*	.78	.77
F	.512	.560*	.559*	.555*	.554*	.535*	.550*	.558*	.560*
		TREC-2004							
P@10	.43	.51*	.50*	.50*	.50*	.48*	.48*	.49*	.46
F	.370	.391*	.389*	.387*	.388*	.376	.382	.386*	.392*

of documents (the collection used in the TREC-8 adhoc experiments [22]) and ran some experiments where the idf statistics were taken from this index. The original TF/ISF method computed at the sentence level over the small document base was superior. It appears that the small index of sentences is good enough for sentence retrieval (at least for these short queries). We therefore set the baseline to be the original TF/ISF approach with stopword and no stemming[4].

In the HTF experiments we took the set of retrieved documents for each query and computed the set of highly frequent terms (ST_q). Several experiments were executed with varying minimum number of occurrences (*mno*) namely, 7, 10, 15, 20[5]. Results are shown in Table 2 (columns 3-6). The results are very encouraging. All the HTF runs produced better performance than the baseline's performance. The improvements are small in TREC-2003. This is not surprising because there is a high population of relevant sentences in this collection (note that the baseline performance is very high, e.g. P@10=74%) and, therefore, it is difficult to get further benefits. There are so many relevant sentences that there is

[4] We use short queries, while the groups participating in the TREC novelty tracks were allowed to use the whole topic. This means that the results presented here are not comparable to any of the results reported in the novelty tracks.

[5] In [20], terms occurring seven or more times in a document were regarded as significant for building a query-biased summary.

no need to apply an evolved method to get a reasonably good top 10. On the other hand, when the retrieved documents have less relevant sentences (2002 and 2004) the HTF retrieval approach produces very significant improvements. It should be noted that there was not a single HTF run yielding worse performance than the baseline. Actually, the method is quite insensitive to the minimum number of occurrences for a term to be considered highly frequent. The four values tested yielded very similar performance. This is a nice feature of the approach as it does not seem problematic to set a good value. Furthermore, the improvements are apparent in both performance ratios. This means that the HTF technique works effectively not only for supplying ten good sentences but also consistently provides relevant sentences across the whole rank.

We also wanted to study the relative merits of HTF against expansion via pseudo-relevance feedback (hereafter, PRF). Although HTF and PRF are intrinsically distinct and both alternatives could actually complement each other, it is still interesting to analyze how robust the HTF method is in comparison with PRF. We experimented with query expansion using an standard PRF technique [4] which consists of adding the most frequent terms from the top ranked documents. This technique, adapted to sentence retrieval (i.e. selecting expansion terms from the top ranked sentences), has proved to be successful to improve the performance of sentence retrieval [10]. There is no empirical evidence that any other advanced query expansion method (e.g. Local Context Analysis [26]) works better than PRF in sentence retrieval.

Since the characteristics of the collections are very different to one another, the comparison between HTF and PRF is general enough, with varying conditions of the type of data and the amount of relevant material. For instance, we expected modest improvements from PRF in 2002 data due to the scarcity of relevant material, and much better PRF results in 2003 and 2004 collections because there are more relevant sentences. Another interesting point to study is the effect of the number of expansion terms. This number affects performance in a critical way (many terms imply usually too much noise in the new query). Therefore we want to check this effect in sentence retrieval and compare it against the behaviour of the HTF approach.

A pool of pseudo-relevance feedback experiments was designed as follows. Given the TF/ISF rank, queries were expanded with the 5, 10, 20 or 50 highest frequent terms in the top 10 sentences. Next, the set of sentences was re-ranked using the new queries. Results are reported in Table 2 (columns 7-10). PRF is much less consistent than HTF. The average performance is clearly lower than the HTF's average performance. The number of PRF runs whose performance is significantly better than the baseline's performance is smaller than the corresponding number of HTF runs. Furthermore, some PRF runs performed worse than the baseline. Summing up, the HTF method works at least as well as PRF techniques and it is less sensitive to variations in its parameter. It is important to note that the HTF method is also convenient for efficiency reasons. It does not require an initial sentence retrieval process and the ST_q sets can be easily

Table 3. Poorest queries

	TF/ISF	HTF	PRF
TREC-2002			
P@10	.03	.09	.03
		7(+) 0(-) 8(=)	1(+) 2(-) 12(=)
TREC-2003			
P@10	.39	.52	.50
		9(+) 3(-) 3(=)	10(+) 2(-) 3(=)
TREC-2004			
P@10	.13	.21	.15
		8(+) 2(-) 5(=)	4(+) 4(-) 7(=)

computed from a regular inverted file (summing the term counts in the retrieved documents).

Let us now pay attention to the behaviour of these sentence retrieval methods when handling poor queries. Such queries are very problematic for retrieval systems and it is important to analyze them in depth. For each collection, we analyzed the 15 queries that had yielded the lowest P@10 figures with the baseline method[6]. In Table 3 we report the P@10 values obtained with the TF/ISF run and the best HTF and PRF runs. For the HTF and PRF runs, we also show the number of queries whose P@10 is better than (+), worse than (-) or equal to (=) the P@10 obtained with the baseline run. The results are very conclusive. In TREC-2003 data, where the 15 worst topics retrieve a reasonably good number of relevant sentences in the top 10 (39% on average), both methods perform roughly the same. On the contrary, in TREC-2002 and TREC-2004, where the initial ranks are quite poor (3% and 13% of relevant sentences on average, respectively), the HTF method is much more robust and performs significantly better than PRF. The HTF method is consistent even in the presence of weak queries. With such queries, pseudo-relevance feedback suffers from poor performance (in some cases, it does not outperform the baseline), whilst the HTF runs still provide solid improvements. This is very valuable because it is usually easy to improve performance when the initial rank has a good population of relevant material but, on the other hand, it is quite hard to improve performance when the initial rank is not good enough. The ability of the HTF method to improve significantly the precision at top ranks when handling poor queries is an important property of this sentence retrieval approach. This suggests that this method is especially suitable for real applications with poorly specified information needs where users only want to go through a small number of sentences.

Having demonstrated that the HTF technique is competitive and outperforms solid sentence retrieval methods when initial ranks are poor, it is important to mention that it is actually compatible with pseudo-relevance feedback. Our method is capable of estimating the centrality of sentences within a retrieved set

[6] For reproductability purposes, these queries were: **T2002**: 305,312,314,315,330, 377,381,406,411,420,432,323,325,326,339. **T2003**: 48,12,14,25,19,20,1,45,24,28,29,30, 5,22,36. **T2004:** 57,77,61,71,86,93,94,97,56,62,65,70,78,80,84.

of documents. This effect could be combined with an expansion approach from a few top ranked sentences and, therefore, retrieval engines can incorporate both techniques. With weak queries the application of pseudo-relevance feedback is not advisable but stronger queries could attain benefits from the combined use of centrality and expansion.

5 Conclusions and Future Work

We have proposed a novel sentence retrieval mechanism based on extracting highly frequent terms from a retrieved set of documents. The experiments reported in this paper demonstrate that this approach outperforms clearly state-of-the-art sentence retrieval methods in the context of a query retrieval use case. We also showed that our method is more robust than standard pseudo-relevance feedback methods and it is simpler because it does not require an initial sentence retrieval process. The method proposed here is especially valuable in terms of the precision at top ranks when queries are weak. In the future we will explore more formal methods to combine the centrality and retrieval scores. We will also study alternative ways to estimate term significance.

Acknowledgements

We thank Tassos Tombros, David Elsweiler and Alvaro Barreiro for their very helpful discussions and recommendations. This work was partially supported by projects PGIDIT06PXIC206023PN and TIN2005-08521-C02-01. David E. Losada belongs to the "Ramón y Cajal" program, whose funds come from "Ministerio de Educación y Ciencia" and the FEDER program.

References

1. Abdul-Jaleel, N., Allan, J., Croft, B., Diaz, F., Larkey, L., Li, X., Smucker, M., Wade, C.: UMass at TREC 2004: Novelty and Hard. In: Proc. TREC-2004, the 13th Text Retrieval Conference (2004)
2. Allan, J., Wade, C., Bolivar, A.: Retrieval and novelty detection at the sentence level. In: Proc. SIGIR-2003, the 26th ACM Conference on Research and Development in Information Retrieval, Toronto, Canada, pp. 314–321. ACM Press, New York (2003)
3. Amati, G., van Rijsbergen, C.J.: Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS) 20(4), 357–389 (2002)
4. Buckley, C., Singhal, A., Mitra, M., Salton, G.: New retrieval approaches using SMART: TREC 4. In: Harman, D. (ed.) Proc. TREC-4, pp. 25–48 (1996)
5. Cardie, C., Ng, V., Pierce, D., Buckley, C.: Examining the role of statistical and linguistic knowledge sources in a general-knowledge question-answering system. In: Proc. ANLP-2000, the 6th Applied Natural Language Processing Conference, Seattle, Washington, pp. 180–187 (2000)
6. Collins-Thompson, K., Ogilvie, P., Zhang, Y., Callan, J.: Information filtering, novelty detection, and named-page finding. In: Proc. TREC-2002 (2002)

7. Dkaki, T., Mothe, J.: TREC novelty track at IRIT-SIG. In: Proc. TREC-2004, the 13th text retrieval conference (2004)
8. Harman, D.: Overview of the TREC 2002 novelty track. In: Proc. TREC-2002 (2002)
9. Kwok, K.L., Deng, P., Dinstl, N., Chan, M.: TREC 2002 web, novelty and filtering track experiments using PIRCS. In: Proc. TREC-2002, the 11th text retrieval conference (2002)
10. Larkey, L., Allan, J., Connell, M., Bolivar, A., Wade, C.: UMass at TREC 2002:cross language and novelty tracks. In: Proc. TREC-2002 (2002)
11. Li, X., Croft, B.: Novelty detection based on sentence level patterns. In: Proc. CIKM-2005. ACM Conf. on Information and Knowledge Management, ACM Press, New York (2005)
12. Mani, I., Maybury, M.T.: Advances in Automatic Text Summarization. MIT Press, Cambridge (1999)
13. Murdock, V.: Aspects of sentence retrieval. PhD thesis, Univ. Massachussetts (2006)
14. Robertson, S.E., Walker, S., Jones, S., HancockBeaulieu, M.M., Gatford, M.: Okapi at TREC-3. In: Harman, D. (ed.) Proc. TREC-3, the 3rd Text Retrieval Conference, pp. 109–127. NIST (1995)
15. Schiffman, B.: Experiments in novelty detection at columbia university. In: Proc. TREC-2002, the 11th text retrieval conference (2002)
16. Silverstein, C., Henzinger, M., Marais, H., Moricz, M.: Analysis of a very large web search engine query log. ACM SIGIR Forum 33(1), 6–12 (1999)
17. Soboroff, I.: Overview of the TREC 2004 novelty track. In: Proc. TREC-2004, the 13th text retrieval conference (2004)
18. Soboroff, I., Harman, D.: Overview of the TREC 2003 novelty track. In: Proc. TREC-2003, the 12th text retrieval conference (2003)
19. Stokes, N., Carthy, J.: First story detection using a composite document representation. In: Proc. of HTL-01, the Human Language Technology Conference, San Diego, USA (March 2001)
20. Tombros, A., Sanderson, M.: Advantages of query biased summaries in information retrieval. In: Proc. SIGIR-1998, the 21st ACM Int. Conf. on Research and Development in Information Retrieval, August 1998, pp. 2–10. ACM Press, New York (1998)
21. Voorhees, E.: Using Wordnet to disambiguate word senses for text retrieval. In: Proc. SIGIR-1993, Pittsburgh, PA, pp. 171–180 (1993)
22. Voorhees, E., Harman, D.: Overview of the eight text retrieval conference. In: Proc. TREC-8, the 8th text retrieval conference (1999)
23. Voorhees, E., Harman, D. (eds.): The TREC AdHoc Experiments, chapter The TREC AdHoc Experiments, pp. 79–97. MIT Press, Cambridge (2005)
24. White, R., Jose, J., Ruthven, I.: A task-oriented study on the influencing effects of query-biased summarisation in web searching. Information Processing and Management 39, 707–733 (2003)
25. White, R., Jose, J., Ruthven, I.: Using top-ranking sentences to facilitate effective information access. Journal of the American Society for Information Science and Technology (JASIST) 56(10), 1113–1125 (2005)
26. Xu, J., Croft, B.: Query expansion using local and global document analysis. In: Proc. SIGIR-1996, Zurich, Switzerland, July 1996, pp. 4–11 (1996)
27. Zhang, H.P., Xu, H.B., Bai, S., Wang, B., Cheng, X.Q.: Experiments in TREC 2004 novelty track at CAS-ICT. In: Proc. TREC-2004 (2004)
28. Zhang, M., Song, R., Lin, C., Ma, S., Jiang, Z., Jin, Y., Liu, Y., Zhao, L.: THU TREC 2002: Novelty track experiments. In: Proc. TREC-2002, the 11th text retrieval conference (2002)