Novel approaches in Text Information Retrieval

Experiments in the Web Track of TREC 2004

Mohamed Farah and Daniel Vanderpooten
Lamsade, University of Paris Dauphine, France

{**farah,vdp**}**@lamsade.dauphine.fr**

Abstract : In this paper, we report our experiments in the mixed query task of the Web track for TREC 2004. We deal with the problem of ranking Web documents within a multicriteria framework and propose a novel approach for information retrieval. We focus on the design of a set of criteria aiming at capturing complementary aspects of relevance. Moreover, we provide aggregation procedures that are based on decision rules, to get the ranking of relevant documents.

1 Introduction

The TREC 2004 mixed query task of the Web track consists in searching a collection of about 1.25 million .Gov documents to find responses to 225 queries : 75 homepage finding (HP) queries, 75 named page (NP) finding queries and 75 topic distillation (TD) queries.

For the HP and NP finding sub-tasks, there is mainly one single document, either a homepage or not, that meets the information need behind the query. Therefore, precision-like measures such as rank reversal, are of the utmost interest. For the TD sub-task, many 'good entry points' to relevant sites can meet the information need. Consequently, a good precision-recall compromise is searched and measures such as average precision are relevant.

To deal with this task, two main approaches are conceivable. The first option is to devise a method that works whatever is the query. The second option is to conceive specific methods for each query type. This last option needs a filtering step to find to which type each unlabeled query belongs to.

Many factors or sources of evidence can be used to derive a ranking. Such factors are the document content and/or structure, the anchor text, the hyperlink structure of the collection, the URL features, etc. Existing approaches generate a ranking of documents based on a score that possibly combines many of these factors. This aggregation is done using either quite simple weighted sum or lexicographic order aggregation operator, or difficult to interpret measures.

In this first participation to the TREC conference series, we submitted one official run. We used a multicriteria approach where we focus on the design of a set of criteria aiming at capturing complementary aspects of relevance. We use 'natural procedures' to aggregate these criteria and rank documents accordingly. We conducted the experiments using a retrieval system designed and developed from scratch to support any search function that might be useful. In the experiments reported here, we process all the queries without a filtering process (option 2).

This paper is organized as follows. We first introduce the multicriteria framework where we describe the overall approach (section 2). In section 3, we describe our system specificities. Experimental results are presented in section 4. Conclusions are provided in the final section.

2 The multicriteria framework

In this work, we acknowledge that the Web search environment is rich with multiple sources of evidence, all of which have strengths that presumably complement one another and weaknesses that can significantly deteriorate retrieval effectiveness when used by themselves. We therefore claim that being able to make effective use of the available information can significantly improve retrieval effectiveness.

We hereafter propose a formal approach for Web information retrieval where relevance is explicitly defined as multidimensional. The overall approach could be split into three phases.

2.1 The modelling phase

It consists firstly of the identification of the factors affecting relevance.

In the text information retrieval context, some factors have solidly been established as important, through extensive testing in the information retrieval literature, and could therefore be considered as the basis for criteria design [4, 6, 10, 11]. Such factors are :

- Term frequency (tf) which is the number of occurrences of terms in documents. It is in fact widely accepted that the more often a term occurs within a document, the more likely it is important for that document.
- Term locality (tl) which captures the presence of terms in specific structured portions of the documents such as the title or the keywords.
- Document length (dl) which evaluates the information contained in documents.

– Term proximity (tp) which captures the nearness of query terms within documents since it is assumed that when query terms are close to each other, the document should best correspond to the information need behind the query.

Moreover, with the advent of the Web, factors pertaining to the popularity or visibility of documents have proven to improve performance. Such factors could be the scores induced by algorithms such as PageRank [1] or HITS [5] as well as simple measures such as documents in-degrees or out-degrees [2].

These factors are then used to develop a set of appropriate decision criteria. These criteria stand for the basis of pairwise comparisons among the *potentially relevant candidate documents*. Each criterion will give rise to a *partial preference relation* modelling the way two documents are compared.

Formally, a criterion is a real-valued function g defined on the set of candidate documents which aims at comparing any pair of documents d and d', on a specific point of view, as follows :

$$g(d) \geq g(d') \Rightarrow d \text{ 'is at least as relevant as' } d'$$

Evaluating document according to each criterion can be affected by imprecision, uncertainty, or inaccurate determination. Moreover, There is no one best way to build criteria and different formulations can be acceptable. Therefore, a slight difference in the scores of two documents, with respect to one criterion, should not lead to discriminate them. In order to model imprecision underlying criteria design, we set the following discrimination thresholds [7] :

– An indifference threshold allows for two close-valued documents to be judged as equivalent although they do not have exactly the same performance on the criterion. The indifference threshold basically draws the boundaries between an indifference and a preference situations.
– A preference threshold is introduced when we want or need to be more precise when describing a preference situation. Therefore, it establishes the boundaries between a situation of a strict preference and an hesitation between an indifference and a preference situations, namely a weak preference.

Thresholds can either be set as fixed or variable along the criterion scale.

Comparing two documents d and d' according to a criterion g_j with an indifference threshold q_j and a preference threshold p_j, leads to the following preferential situations :

– $dI_jd' \Leftrightarrow -q_j \leq g_j(d) - g_j(d') \leq q_j$
– $dQ_jd' \Leftrightarrow -q_j < g_j(d) - g_j(d') \leq p_j$
– $dP_jd' \Leftrightarrow g_j(d) - g_j(d') > p_j$

where I_j, Q_j and P_j represent respectively *indifference*, *weak preference* and *strict preference* relations restricted to criterion g_j. These 3 relations could be grouped into an outranking relation S_j such that $dS_jd' \Leftrightarrow g_j(d) + q_j \geq g_j(d')$.

Each criterion can be endowed with specific indicators of importance, namely its *relative importance* which gives an idea on what criteria are more important than others.

2.2 The aggregation phase

This phase takes the partial preference structures induced by the criteria family and aggregates them into one or more *overall preference relation*(s) that model relevance. To do so, we use partial compensatory aggregation mechanisms that are based on the following basic principles to compare two documents d and d' [8] :

– A *concordance* principle : d 'is as least as relevant as' d' if the majority of criteria agree with this assertion, and
– A *discordance* principle : d 'is as least as relevant as' d' if none of the discordant criteria strongly refutes this assertion.

Formally, we build an overall outranking relation S, whose meaning is 'is at least as relevant as'. Therefore, dSd' if and only if the two above-mentioned conditions hold. These conditions may be more or less demanding, resulting in different outranking relations.

Let

– $F = \{g_1, \ldots, g_p\}$ be a family of p criteria,
– H be an overall preference relation, where H is P, Q, I or S,
– H_j be a partial preference relation, i.e. restricted to criterion g_j,
– $C(dHd') = \{j \in F : dH_jd'\}$ be the concordance coalition of criteria in favor of establishing dHd'.

We distinct two basic families of outranking relations.

If we do not have information on the relative importance of the criteria, we use conditions referring to the number of criteria *supporting* or *refuting* the outranking such as the following outranking relations :

$$dS_1d' \Leftrightarrow C(dSd') = F \tag{1}$$

which is a well established relations.

$$\begin{aligned} dS_2d' \;\Leftrightarrow\; & card(C(dPd')) \geq card(C(d'Qd)) \\ & and\ C(d'Pd) = \emptyset \end{aligned} \tag{2}$$

where there should be more criteria concordant with dPd' than criteria concordant with $d'Qd$. At the same time, there should be no criterion concordant with $d'Pd$.

$$dS_3d' \Leftrightarrow card(C(dPd')) \geq card(C(d'P \cup Q\, d)) \quad (3)$$

where there should be more criteria concordant with dPd' than criteria supporting a strict or weak preference in favor of d'.

$$dS_4d' \Leftrightarrow card(C(dPd')) \geq card(C(d'Pd)) \quad (4)$$

where there should be more criteria concordant with dPd' than criteria concordant with $d'Pd$.

where

$$S_1 \subset S_2 \subset S_3 \subset S_4$$

since when we move from S_j to S_{j+1}, the credibility of the comparisons gets weaker.

If the relative importance of criteria can be assessed, we can use more elaborated outranking relations that allow compensation in some directions but not others. More precisely, we can accept that a 'relatively' bad performance on one criterion g_j can be compensated by a good performance on a more important criterion g_i.

2.3 The exploitation phase

In order to derive the final ranking, we need exploitation procedures that enrich the outranking relations elaborated in the aggregation phase. In fact, these relations are not necessarily transitive. To do so, we use information on how each document is compared to the other documents according to each outranking relation.

The exploitation procedure we use has its roots in the [8, 9]. It consists in partitioning the set of potential candidates D into r ranked classes leading to the definition of a complete preorder. Each class C_h encloses documents that are considered as ex aequo where C_1 is the best class.

Each class C_h results from a *distillation process* which is an iterative process starting from the base set $E_0 = D \setminus \{C_1 \cup \ldots \cup C_{h-1}\}$. This process iterates over the outranking relations defined in the aggregation phase, starting from the well established relation till the less demanding one. Each iteration i tries to reduce the set of ex-aequo documents, starting from the distillate E_{i-1}, to get the distillate E_i, as follows :

1. Compute for each $d \in E_{i-1}$ its qualification $s_i(d, E_{i-1})$, which is the difference of the number of documents that could be considered as 'better' in the one side, and 'worse' in the other side, than d according to S_i,
2. $s_{\max} = \max_{d \in E_{i-1}}\{s_i(d, E_{i-1})\}$
3. $E_i = \{d \in E_{i-1} : s_i(d, E_{i-1}) = s_{\max}\}$

The distillation process stops when $card(E_i) = 1$, otherwise, the last class will consist of the last distillate produced by the use of the less demanding outranking relation.

3 System description

To facilitate empirical investigation of the proposed methodology, we developed a prototype search engine implementing a basic multicriteria approach. The software is entirely implemented in java SDK1.4.1. It mainly consists of five agents (see figure 1) :

- A text parser which processes the TREC collection to get the document surrogates. In fact, document representation is based on its internal content delimited by the body tags, as well as texts in the URL, the title, keywords and description tags. Moreover, we use link anchor texts present in Web documents pointing to the current document since it is shown in [3] that such texts often provide accurate descriptions of pointed documents content. We consider only stems as index terms using the Porter algorithm to reduce the size of the inverted index. We also discard cuurent english stopwords,
- A graph parser which builds the TREC graph where nodes are documents and edges are hyperlinks between documents,
- An indexer which produces an inverted index of the whole collection,
- A filtering agent which matches query terms with TREC documents to get the unordered base set of potentially relevant items, and
- A ranking agent which implements the muticriteria algorithm.

Five criteria are considered in this study. We distinguish one-term queries having single terms from complex queries.

- g_1 captures frequencies of query terms. For one-term queries, we use
$$\frac{tf}{tf_{max}}$$
where tf is the term frequency and tf_{max} corresponds to the most frequent term in the document. For complex queries, we retain either the product or the sum aggregation operator.
- g_2 captures the occurrences of query terms within the anchor text (location L_1), the URL ($L2$), the title (L_3), the keywords tag (L_4) and the description tag (L_5). For the one-term queries, we first compute a partial performance with respect to each location L_i :
$$g_2(d, t, L_i) = \textit{number of occurrences of t in } L_i$$
The overall score of each document with respect to this criterion is computed using the weighted sum operator as follows :
$$g_2(d, t) = \frac{\sum_{i=1}^{5} k_i g_2(d, t, L_i)}{\sum_{i=1}^{5} k_i}$$

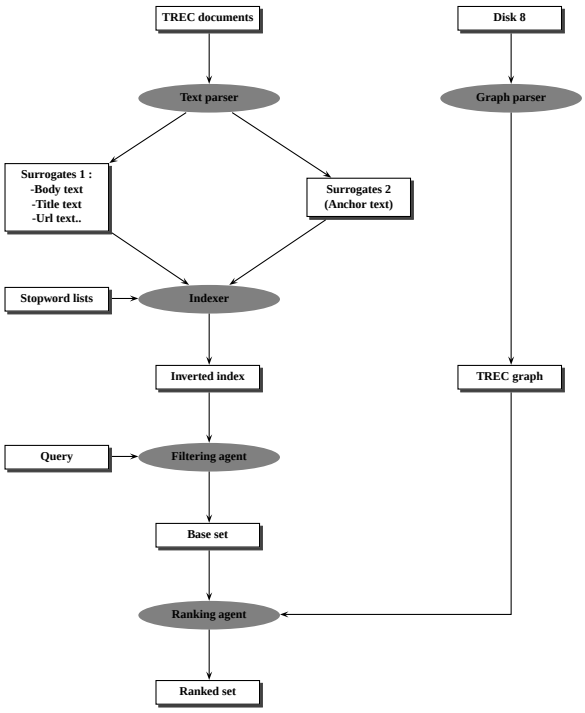

Fig. 1. System description

where k_i is an importance score for location L_i. In fact, we suppose that some locations are more important that others.

For complex queries, the overall score is computed as

$$g_2(d,q) = \sum_{t \in q} g_2(d,t)$$

- g_3 is a popularity measure. It is either the number of document in-links, or the authority score computed using the HITS algorithm.
- g_4 aims at favoring entry points to relevant sites by determining for each document the number of its relevant children.
- g_5 is a proximity criterion that is only considered for complex queries. It is inversely proportional to the minimal matching span of query terms in the document, i.e. the smallest text excerpt from the document that contains all the query terms.

It is obvious that these criteria play different roles for each specific task. Therefore, knowing to which type each topic belongs can lead to a definition of specific procedures for each query type. Since this is not the case, we consider that each criterion is neither predominant nor negligible and develop an aggregation procedure that is based on the outranking relations of equations (1)-(4).

4 Results

We report here performances of the official run we submitted to the mixed query task of the Web track, as well as other runs we carried out after an expansion of our system so that it can process queries with high number of potentially relevant documents.

We begin with a detailed performance analysis on a per query type basis, then discuss overall performance of each run.

Beforehand, we begin with a brief description of the runs.

4.1 The Runs

The various runs we discuss in this paper are the followings :

- LamMcm1 : This is the official run we submitted to TREC 2004. It implements the aggregating and exploitation procedures of sections (2.2) and (2.3). Criteria g_1 to g_4 are considered where the product aggregation operator is used for the criterion g_1. We suppose that no information on the relative importance of criteria is available. We therefore use the outranking relations of equations (1)- (4). The indifference and preference thresholds are considered as variable. They are set to 20% and 50%, respectively. For each query, we only considered the first 500 processed documents from those that match query terms. This is basically due to the limitations of the prototype version of the system.
- mcm1 : This is the same as LamMcm1 but uses another filtering mechanism which is based on criterion g_2. In fact, we considered the first 1000 documents according to his criterion. This is the basic run for the followings.
- mcm2 : Same as mcm1 but uses a different formulation of criterion g_1. We used the sum operator to aggregate frequencies of the terms of the query.
- mcm3 : Same as mcm1 but uses fixed thresholds. We used the following values : $q = (0.05; 0; 5; 3)$ and $p = (0.1; 0.25; 10; 6)$.
- mcm4 : Same as mcm1 but uses criterion g_5.

4.2 The HP/NP finding sub-tasks

HP/NP queries correspond to situations where users do not need large lists of relevant documents to search in but would rather prefer to get the URL of a specific home page (HP queries) or a non-home page document (NP queries).

Tables 1 and 2 depict various statistics based on relevance assessments of the HP and NP query tasks, respectively. They clearly show that there is one correct answer per HP/NP topic.

Tables 3 and 4 provide a summary description of the runs with respect to the HP/NP queries.

For the HP queries, using fix thresholds leads to the best run (mcm3) with MRR=56.77%, whereas using the sum aggregation operator for criterion g_1 retrieves more relevant

num_queries	75
num_rel_docs	83
Mean_rel_docs / query	1.106

Table 1. Relevance judgment statistics for HP queries

num_queries	75
num_rel_docs	78
Mean_rel_docs / query	1.040

Table 2. Relevance judgment statistics for NP queries

Run	MRR	S@1	S@5	S@10
mcm1	46.06	36.00	57.33	68.00
mcm2	55.67	**46.67**	65.33	74.67
mcm3	**56.77**	41.33	**74.67**	**80.00**
mcm4	35.33	29.33	41.33	49.33
LamMcm1	32.61	26.67	41.33	45.33

Table 3. Summary description of the Runs for the HP query finding task

Run	MRR	S@1	S@5	S@10
mcm1	**57.34**	**45.20**	73.97	75.34
mcm2	56.99	41.10	**79.45**	**82.19**
mcm3	52.74	41.09	67.12	75.34
mcm4	50.79	38.35	65.75	69.86
LamMcm1	32.26	21.33	44.00	54.67

Table 4. Summary description of the Runs for the NP query finding task

items in the first position (S@1 = 46.67% for mcm2). Run mcm3 is also the best run to retrieve rapidly most of the correct answers (S@5=74.67% and S@10=80.00%).

For the NP queries, run mcm2 retrieves more relevant items in the top 5 and 10 retrieved list whereas run mcm1 is more efficient to get the relevant item in the first position (MRR=57.34% and S@1 = 45.20%).

Surprisingly, using criterion g_5 deteriorates performance for both the HP and TD queries.

4.3 The Topic Distillation sub-task

This task aims at retrieving 'key resources' on a given topic. There is no explicit definition of what constitutes a key resource. Nevertheless, it seems to be appropriate to return at first, father pages which have many relevant children. This is achieved by the criterion g_4.

Table 5 provides a summary description of the runs with respect to the TD queries.

Firstly, this table shows low performance of the official run. This is principally due to the following factors:

- The run was not considered for judgment in the pooling procedure, which means that the best ranked documents of the run may not have been assessed for relevance judgements,

Run	AvP	S@1	S@5	S@10
mcm1	**13.25**	36.00	69.33	80.00
mcm2	13.14	**40.00**	**74.67**	**81.33**
mcm3	11.53	22.67	65.33	74.67
mcm4	9.68	33.63	53.36	61.43
LamMcm1	4.87	17.33	40.00	46.67

Table 5. Summary description of the Runs for the TD task

- In the prototype version of our system, among all the documents matching the query (the base set), only a subset of at most 500 documents could be processed. Therefore a wide number of relevant items are skipped for evaluation. In HP/NP tasks, this limitation does not have significant impact on the system performance, but in the TD task, there is an important impact on performance,
- We did not use any training queries from last years to tune some parameters and especially thresholds.

Runs mcm1 and mcm2 have the best performances whereas performance falls down when the proximity criterion is used.

4.4 The mixed-query task

Table 6 shows that :

- Run mcm2 is significantly better than the other runs,
- Considering the proximity criterion deteriorates performance, and
- The use of variant thresholds performs better.

Run	AvP	S@1	S@5	S@10
mcm1	38.35	39.01	66.81	74.43
mcm2	**41.05**	**42.60**	**73.09**	**79.37**
mcm3	39.68	34.98	69.05	76.68
mcm4	31.19	33.63	53.36	61.43
LamMcm1	22.68	21.78	41.78	48.89

Table 6. Summary description of the Runs for the mixed-query task

5 Conclusions

In this paper, we proposed a novel vision of the information retrieval problem and described a system that implements a multicriteria methodology. The ideas developed for the TREC collection of textual documents can be adapted to any family of criteria deemed to be relevant and respecting some properties. It also provides a general framework within which different document types could be searched and retrieved.

Further experiments are currently undertaken to help judge which criteria are more important than others or which ones are more task specific.

References

1. S. Brin and L. Page. Anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International World Wide Web Conference*, pages 14–18, Brisbane, Australia, Apr. 1998.
2. J. Carriere and R. Kazman. Webquery: Searching and visualizing the web through connectivity. In *Proceedings of the 6th International World Wide Web Conference*, pages 7–11, Santa Clara, California, Apr. 1997.
3. N. Crasswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In *Proceedings ACM SIGIR 2001*, pages 250–257, 2001.
4. A. Kilgarriff. Which words are particularly characteristic of a text ? a survey of statistical approaches. In *Proceedings of the AISB Workshop on Language Engineering for Document Analysis and Recognition*, pages 33–40, Brighton, UK, Apr. 1996.
5. J. M. Kleinberg. Authoritative sources in a hyperlinked environnement. *Journal of the ACM*, 46(5):604–632, Sept. 1999.
6. S. E. Robertson and K. Spark Jones. Simple proven approaches to text retrieval. Technical report 356, Cambridge University Computer Laboratory, May 1997.
7. B. Roy. Main sources of inaccurate determination, uncertainty and imprecision. *Mathematical and Computer Modelling*, 12(10/11):1245–1254, 1989.
8. B. Roy and D. Bouyssou. *Aide multicritère à la décision : Méthodes et cas*. Economica, 1993.
9. B. Roy and J. Hugonnard. Ranking of suburban line extension projects on the Paris metro system by a multicriteria method. *Transportation Research*, 16A(4):301–312, 1982.
10. G. Salton. *Automatic Text Processing*. Addison-Wesley, 1989.
11. G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.