Inf Retrieval (2008) 11:315–334
DOI 10.1007/s10791-008-9046-z

An outranking approach for information retrieval

Mohamed Farah · Daniel Vanderpooten

Received: 19 January 2008 / Accepted: 21 January 2008 / Published online: 16 February 2008

Abstract Over the last three decades, research in Information Retrieval (IR) shows performance improvement when many sources of evidence are combined to produce a ranking of documents. Most current approaches assess document relevance by computing a single score which aggregates values of some attributes or criteria. They use analytic aggregation operators which either lead to a loss of valuable information, e.g., the min or lexicographic operators, or allow very bad scores on some criteria to be compensated with good ones, e.g., the weighted sum operator. Moreover, all these approaches do not handle imprecision of criterion scores. In this paper, we propose a multiple criteria framework using a new aggregation mechanism based on decision rules identifying positive and negative reasons for judging whether a document should get a better ranking than another. The resulting procedure also handles imprecision in criteria design. Experimental results are reported showing that the suggested method performs better than standard aggregation operators.

Keywords Information retrieval · Relevance · Outranking approach · Multiple criteria · Aggregation

1 Introduction

Information Retrieval (IR) is concerned with situations where a user, having information needs, performs queries on a collection of documents to find a limited subset of the most relevant ones. Performances of IR systems is measured by its ability to search and retrieve *relevant* documents as efficiently and effectively as possible. In this paper, efficiency,

M. Farah (✉) · D. Vanderpooten
Lamsade, Université Paris-Dauphine, Paris, France
e-mails: farah@lamsade.dauphine.fr; Mohamed.Farah@riadi.rnu.tn

D. Vanderpooten
e-mail: vdp@lamsade.dauphine.fr

M. Farah
Riadi, Faculté des Sciences de Monastir, Monastir, Tunisia

which refers to the ability of a system to provide results within reasonable response times, is not our main concern. We primarily focus on retrieval effectiveness, which refers to the ability of a system to deliver the most relevant results first. Relevance is indeed the main challenge for most search engines as shown by several comparative studies reporting limitations in their performances (Hawking et al. 2001).

In the literature, a wide range of models have been proposed to rank documents according to their relevance to queries. They result in different rankings depending on the way they define relevance. In fact, relevance is reflected by the sources of evidence that are considered, as well as the way they are combined.

Most of the current approaches assess document relevance by computing a single score which aggregates values of elementary attributes related to the query terms, the document or the relationship between these two entities. For instance, in the Vector Space Model (Salton et al. 1975), the Okapi BM25 probabilistic model (Robertson et al. 1994) as well as language models (Cao et al. 2005), term frequency (*tf*), document frequency (*df*) and document length (*dl*) are the main attributes which come into play. These attributes are combined in the term weighting formulation which corresponds to a first aggregation phase. The resulting scores are in turn considered to compute document relevance status value (rsv) to queries, as a second aggregation phase.

With the advent of hypertext collections, such as the Web, attributes characterizing the hyperlink structure are considered and led to link-based measures such as Kleinberg's HITS scores (Kleinberg 1999), PageRank scores (Brin and Page 1998) and HostRank scores (Amento et al. 2000).

All these text- and link-based attributes can be combined to get better performance. A variety of aggregation operators have been used such as the min and max operators in (Fox and Shaw 1994) or the weighted linear operator in (Craswell et al. 2005). Other aggregation operators include similarity-based measures (Van Rijsbergen 1979; Salton and McGill 1983; Frakes and Baeza-Yates 1992), P-norms (Salton et al. 1983), or fuzzy-logic conjunctive and disjunctive operators (Dubois and Prade 1984).

In some cases, aggregation is performed in an ad-hoc manner. For instance, in (Kraaij et al. 2002) link-based attributes such as in-degree and URL, are used as priors in language models. Another way consists in aggregating evidence in two stages. In the first stage, text-based attributes are combined to get scores of documents. In the second stage, the resulting top ranked documents are re-ordered according to link information by using techniques such as spreading activation or probabilistic argumentation (Savoy and Rasolofo 2000). Thus, these approaches do not explicitly use link-attributes.

Each aggregation operator conveys a specific aggregation logic which reflects the degree of compensation we are ready to accept. In the IR literature, two main classes of operators are in use. The first class corresponds to a *totally compensatory logic*. It consists of building a single score using a more or less complex operator such as the weighted sum. For such operators, a very bad score on one criterion can be compensated by one or several good scores on other criteria. These operators often require inter-criteria information such as weights, which are sometimes difficult to define and interpret. Indeed, these weights aim at capturing at the same time the relative importance of criteria but also a normalization factor when criteria are expressed on different scales.

The second class corresponds to a *non-compensatory logic*. In this case, aggregation is mainly based on one criterion value such as the worst score or the score of the most important criterion. The remaining criteria are only used to discriminate documents with similar scores. This gives rise to min-based or lexicographic-based operators, variations of

which are the *discrimin* and *leximin* operators (Boughanem et al. 2005). A clear weakness of this class of operators is that a large part of the scores is ignored or plays a minor role.

Moreover, in both classes, we do not consider *imprecision* underlying criteria design resulting from the fact that there are many acceptable formulations of the same criterion: for instance, Anh and Moffat (2002) proposed four alternative formulations of the *tf* criterion. Therefore, it is important to give a limited interpretation to values, i.e., we should consider that slight differences in values are often not meaningful. This way, the resulting rankings are more *robust*.

In this paper, we propose a multiple criteria framework which combines any set of criteria while taking into consideration the imprecision underlying the criteria design process. We first put emphasis on the importance of the design of good criterion families capturing *complementary* aspects of relevance and give clues to the design of such families. Then, we describe ranking procedures based on natural decision rules.

Multiple criteria techniques were previously used in IR, especially in information filtering (Pasi et al. 2007) as well as in data fusion (Bordogna et al. 2003; Bordogna and Pasi 2004). Nevertheless, the proposed methods basically use fuzzy sets theory. In this paper, we use a different kind of aggregation mechanisms.

The paper is organized as follows. We first introduce the multiple criteria framework where we describe the overall approach and its component phases (Sect. 2). Then, we highlight some specificities of the IR problem which are addressed in the proposed approach (Sect. 3). Section 4 deals with the modeling phase which consists in designing a set of relevance criteria. We present in Sect. 5, a filtering procedure whose purpose is to obtain a reduced set of potentially relevant documents. Section 6 shows how to aggregate such criteria and build the final ranking. The complexity of the whole approach is investigated at the end of this section. We report experimental results in Sect. 7 and provide conclusions in a final section.

2 A multiple criteria framework for IR

Many studies argued that the reason why no consensus has been reached on the relevance concept is that there are many kinds of relevance, not just one, as stated by Borlund (2003). Moreover, different sources of evidence are contributing to capture the relevance concept. Therefore, being able to make effective use of these sources of evidence can significantly improve retrieval effectiveness.

We propose a formal approach for IR where relevance is explicitly defined as multidimensional (by a set of criteria) and ranking is derived from pairwise comparisons of document performance vectors (*document profiles*) using decision rules identifying positive and negative reasons for judging whether or not a document should get a better ranking than another. The overall approach can be split into four phases (see Fig. 1) which will be detailed in the following sections:

- The *modeling phase* consists in identifying various attributes affecting relevance. These attributes are used to develop a set of appropriate decision criteria which model different aspects of relevance. Each criterion will give rise to a *partial preference relation* (binary relation) modeling the way two documents are compared, according to that criterion.
- The *filtering phase* aims at identifying the set of *potentially relevant documents* with respect either to the query structure or to the criterion family. In the first case, a boolean

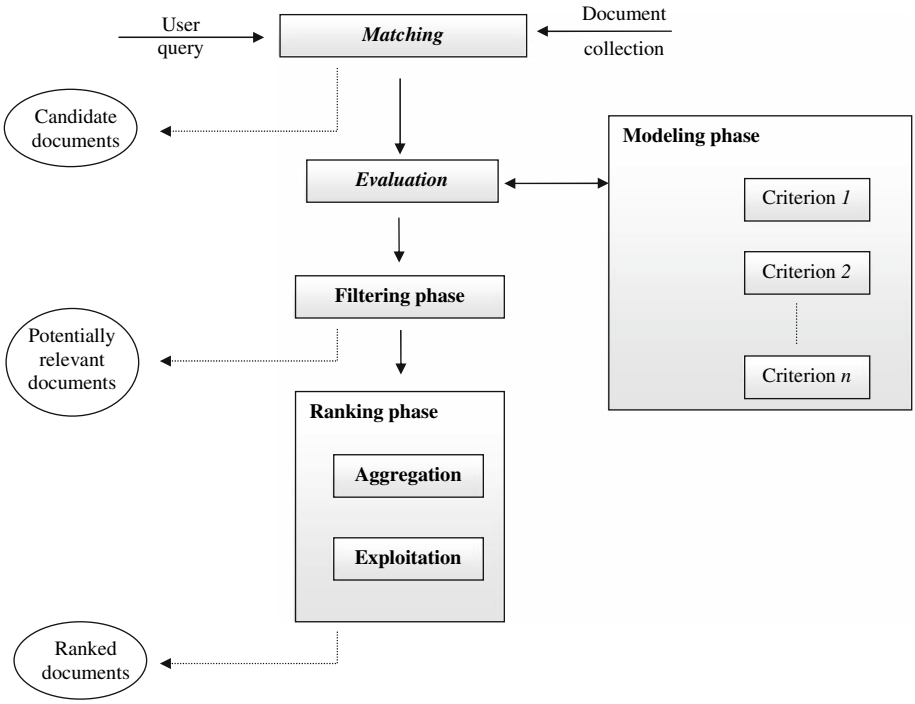

Fig. 1 Overall approach

filter selects documents that match query terms and query formula. In the second case, a profile-based filter selects documents that satisfy an *acceptance profile* defined by minimal required values on some or all criteria.

– The *aggregation phase* aggregates partial preference relations derived from pairwise comparisons of documents with respect to each criterion, into one or several *global preference relations*. A global preference relation indicates how two documents are compared with respect to all the considered criteria.
– The *exploitation phase* processes global preference relations resulting from the previous phase in order to derive the final ranking.

The last two phases correspond to the *ranking phase*.

It is worth noting that the proposed method is collection- and representation-independent to some extent. It can thus be used for any type of collection and combined with the best representation available. In fact, the context is mainly considered in the modeling phase in order to devise relevant criterion families.

3 Specificities of the IR problem

The IR problem can be considered as a multiple criteria decision problem when we explicitly consider the multidimensional nature of relevance. Nevertheless, it has some particularities that have an impact on the modeling phase as well as on the aggregation and exploitation phases.

3.1 Specificities for the modeling phase

Specificity 1: Two kinds of criteria need to be considered to assess documents relevance: query-dependent and query-independent criteria.

Query-dependent criteria measure semantic proximity between documents and queries and are derived from attributes about the form of occurrences of query terms in the document and the collection. Examples of such attributes are term frequency (*tf*) and document frequency (*df*).

The evaluation of query-dependent criteria depends on the structure of the query. In fact, we should distinguish *one-term queries* from *multi-terms queries*. Some criteria are only relevant in the second case. Moreover, for multi-terms queries, two evaluation levels are required: (i) evaluation for each term of the query, and (ii) aggregation of these evaluations. Therefore, the design of such criteria deserves thorough analysis. This is addressed in Sect. 4.1.

Query-independent criteria mainly refer to characteristics of the document and the collection. They can be evaluated independently of the query. Examples of such criteria are document length (*dl*) and PageRank. We need such criteria to better help discriminating between documents. In fact, the query frequently consists of two or three terms in average, and this cannot be sufficient to rank thousands or millions of documents.

Specificity 2: Criteria can play different roles depending on which phase they are used in. In the filtering phase, they are primarily used to build acceptance profiles which help separating potentially relevant documents. In the ranking phase, they are used for pairwise comparisons.

3.2 Specificities for the ranking phase

Specificity 3: Criteria to be used to establish relevance are not specified by the user. They are rather based on attributes evidenced to best capture relevance by the IR community. Consequently, it is difficult to get precise preference information regarding their relative importance. In this case, we assume that each criterion is neither prevailing nor negligible. Therefore, we should use appropriate ranking procedures.

Specificity 4: The query is too poor to justify a precise ranking of documents. One can expect that many of the 'most relevant' documents should be present in the head of the ranking, but their exact ranking is meaningless. This can also be justified in terms of users behavior when interacting with the results pages of search engines. In fact, research in *eye-tracking* analysis of users behavior has shown that once users have started scrolling, rank becomes less of an influence for attention (Granka et al. 2004). Therefore, even if a ranking is a handy way of presenting results, its significance should not be overemphasized.

4 Modeling phase

In our context, a criterion models relevance between documents, regarding a specific point of view. It is represented by a real-valued function g defined on the set of documents and aims at comparing any pair of documents d and d', on a specific point of view, as follows:

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

For instance, considering the term frequency criterion (*tf*), it is always common to consider that when one query term occurs more frequently in the body of document d than in document d', then d is judged more relevant than d', ceteris paribus: $tf(d) \geq tf(d') \Rightarrow d$ 'is at least as relevant as' d' according to criterion *tf*.

Choosing the right criterion family depends on the task at hand as well as the type of information that documents encompass. In fact, retrieving images or video sequences differs greatly from retrieving textual documents since each kind of information encompass specific features. This choice should be undertaken with great care since it has an important impact on the final ranking.

Although many candidate criterion families could be derived from the same considered relevant attributes, we should nevertheless try to fulfill the following desirable requirements:

- each criterion should be concerned with a specific point of view,
- all attributes deemed to be important in comparing two documents should be captured by the set of criteria,
- we should avoid redundancy, i.e., we should not consider the same attribute more than once and therefore, it is better to have independent criteria in order not to favor attributes upon others, and
- while building the criterion family, we should have in mind the way it will be used in the ranking process.

It is worth noting that many formulations of the same criterion are possible. Therefore, we should not overemphasize the criterion scores of documents. We briefly discuss two important issues of the modeling phase.

4.1 Evaluation of query-dependent criteria

To build some query-dependent criteria, such as the *tf*-like criterion, we need to make a clear distinction between one-term and multi-terms queries. For one-term queries, building criteria has no specific difficulties, but to deal with multi-terms queries, i.e., conjunctive and/or disjunctive queries, we can proceed in two steps:

- build a *sub-criterion* corresponding to each term of the query. Each literal of the query formula can therefore be evaluated accordingly,
- select an aggregation operator corresponding to each query-type (conjunctive query, disjunctive query or a combination of both). This *sub-aggregation* step aggregates *homogeneous* partial measures derived from the previous step.

Since elements being aggregated in the sub-aggregation step are homogeneous, we can use analytic aggregation operators like conjunctive, disjunctive or compensatory operators (Dubois and Prade 1984), depending on the aggregation logic we wish to use and on the interpretation given to the juxtaposition of terms.

For instance, let us suppose that we want to assess the relevance of documents to some query $q = t_1 t_2 \ldots t_{n_q}$ according to the *tf* criterion, where t_k is a query term. In the first step, we compute the score of each document d for each query term t_k, i.e., $tf(d, t_k)$. In the second step, we combine these different scores into one single score using some aggregation operator such as the average operator, i.e., $tf(d) = \frac{tf(d,t_k)}{n_q}$.

4.2 Modeling imprecision

It is often inadequate to consider that slight differences in evaluation should give rise to clear-cut distinctions. This is particularly true when different formulations of criteria are acceptable. Imprecision underlying criteria design can be modeled using the following discrimination thresholds (Roy 1989):

- An *indifference threshold* allows for two documents with close criterion values to be judged as equivalent. The indifference threshold basically draws the boundary between an indifference and a preference situation.
- A *preference threshold* is introduced when we want or need to be more precise when describing a preference situation. Therefore, it establishes the boundary between a situation of a strict preference and an hesitation between an indifference and a preference situations, namely a weak preference.

A criterion g_j, having indifference and preference thresholds, q_j and p_j, respectively ($p_j \geq q_j \geq 0$), is called a *pseudo-criterion*. Comparing two documents d and d' according to a pseudo-criterion g_j leads to the following partial preference relations:

$$\begin{cases} dI_j d' \Leftrightarrow |g_j(d) - g_j(d')| \leq q_j \\ dQ_j d' \Leftrightarrow q_j < g_j(d) - g_j(d') \leq p_j \\ dP_j d' \Leftrightarrow g_j(d) - g_j(d') > p_j \end{cases}$$

where I_j, Q_j and P_j represent respectively *indifference, weak preference* and *strict preference relations* restricted to criterion g_j. These three relations could be grouped into an *outranking relation* $S_j = (I_j \cup Q_j \cup P_j)$ such that $dS_j d' \Leftrightarrow g_j(d) - g_j(d') \geq -q_j$ which corresponds to the assertion *d'is as least as relevant as'd'* with respect to the aspects covered by criterion g_j.

To model situations where a very low score of a document d' with respect to d, according to some criterion g_j, cannot be compensated by a good score on one or several other criteria, we use a *veto threshold*v_j ($v_j \geq p_j$) and define the following *veto relation* $V_j : dV_j d' \Leftrightarrow g_j(d) - g_j(d') > v_j$. In this case, d' cannot be considered as *'at least as relevant as'd*, whatever the scores on other criteria.

Figure 2 summarizes the different preference situations that can be derived from the comparison of two documents d and d'.

We illustrate these different preference relations using the following example. Let us consider Table 1 which gives the scores of five documents evaluated according to a pseudo-criterion g. Table 2 gives the different thresholds of this criterion. In this illustration, we denote $g_{ij} = g(d_i) - g(d_j)$ which corresponds to the difference of the scores of documents d_i and d_j according to criterion g. Table 3 reports the differences of document scores and Table 4 gives the relational interpretation of such differences. For instance,

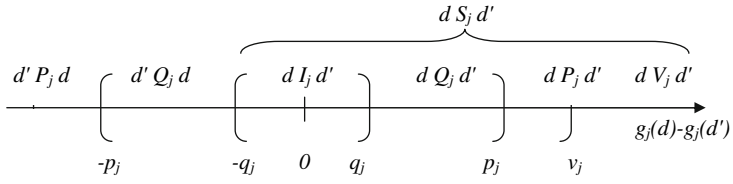

Fig. 2 Preference relations

Table 1 Documents scores according to g

	g
d_1	0.8
d_2	0.7
d_3	0.5
d_4	0.4
d_5	0.1

Table 2 Threshold values of g

	q	p	v
g	0.2	0.4	0.6

Table 3 Difference of document scores w.r.t. g

g_{ij}	d_1	d_2	d_3	d_4	d_5
d_1	–	0.1	0.3	0.4	0.7
d_2	–	–	0.2	0.3	0.6
d_3	–	–	–	0.1	0.4
d_4	–	–	–	–	0.3
d_5	–	–	–	–	–

Table 4 Partial preference relations between documents w.r.t. g

	d_1	d_2	d_3	d_4	d_5
d_1	–	I	Q	Q	P,V
d_2	–	–	I	Q	P
d_3	–	–	–	I	Q
d_4	–	–	–	–	Q
d_5	–	–	–	–	–

since $q \leq g_{13} = 0.3 \leq p$ the weak preference relation holds between d_1 and d_3. Moreover, since $g_{15} > v > p$ both the strict preference relation as well as the veto relation hold between d_1 and d_5. This involves, in particular, that criterion g imposes its veto to the assertion 'd_5 is at least as good as d_1', whatever the scores on other criteria.

5 Filtering procedure

In this section, we show how it is possible to get the top k best relevant documents using acceptance profiles. In fact, acceptance profiles allows us to discriminate documents that can be considered better than the acceptance profile. Different procedures can be used to obtain the top k best relevant documents. We give one such procedure.

Suppose that we have a set D of n documents, possibly resulting from the application of a first boolean filter, and we need to retain only the top k best documents, using an acceptance profile, i.e., acceptance thresholds a_j on each criterion g_j. The problem is to define these values $a_j\,(j = 1, \ldots, p)$ such that the set of acceptable documents $A = \{d \in D | g_j(d) \geq a_j\ (j = 1, \ldots, p)\}$ has an approximate cardinality of k. A simple way of

setting and adjusting values a_j $(j = 1,\ldots,p)$ is to use a single parameter α corresponding to a percentile used on all criteria scales. Considering that we want to retain, for each criterion, a proportion α of the documents so as to retain globally a proportion $\frac{k}{n}$ of documents from D, α can be set to an initial value of $\sqrt[p]{\frac{k}{n}}$. Using a dichotomic procedure, α can be adjusted so as to obtain the required size for the filtered set A.

6 Ranking procedure

In order to get a global relevance model on the set of documents, we use *outranking approaches* (Roy 1991), which are quite appropriate regarding the specificities of Sect. 3.2 and are based on a *partial compensatory logic*. They consist of two phases: an aggregation phase and an exploitation phase. We hereafter give details about both phases. In Sect. 6.3, we illustrate how they work precisely using a simple example.

6.1 Aggregation phase

Outranking approaches take as input the partial preference relations induced by the criterion family and aggregate them into one or more global preference relation(s) S. They are particularly relevant in our context since they (i) permit considering imprecision in document evaluations, (ii) can handle criteria expressed on heterogeneous scales, (iii) use all the available information on document performances, and (iv) do not necessarily require inter-criteria information, such as weights.

In order to accept the assertion dSd', stating that 'document d is at least as relevant as document d'', the following conditions should be met:

- a *concordance* condition which ensures that a majority of criteria are concordant with dSd' (*majority principle*).
- a *discordance* condition which ensures that none of the discordant criteria strongly refutes dSd' (*respect of minorities principle*).

In this paper, we suppose that there is no available information on the *relative importance of criteria*. In this case, to accept the assertion dSd', we use decision rules based on the criteria *supporting* (positive reasons) or *refuting* (negative reasons) this assertion. Obviously, the rules for defining this support may be more or less demanding, resulting in different outranking relations. For example, let

- $F = \{g_1,\ldots,g_p\}$ be a family of p criteria,
- H be a global preference relation, where H is P, Q, I, V or S,
- H^- be a relation such that $dH^-d' \iff d'Hd$,
- 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', and
- $c(dHd')$ the number of items in $C(dHd')$

A candidate outranking relation is:

$$dS^1d' \Leftrightarrow C(dSd') = F \tag{1}$$

which is a well established, but usually poor, relation since it only holds if all the criteria are concordant with dSd'.

We can also use less demanding outranking relations such as:

$$dS^2d' \Leftrightarrow c(dPd') \geq c(dP^- \cup Q^-d') \quad \text{and} \quad C(dV^-d') = \emptyset \tag{2}$$

To accept dS^2d', there should be more criteria concordant with dPd' than criteria supporting a strict or weak preference in favor of d'. This corresponds to the concordance condition. At the same time, no discordant criterion should strongly disagree with this assertion. This corresponds to the discordance condition.

$$\begin{aligned} dS^3d' \Leftrightarrow\ & c(dPd') \geq c(dP^-d') \\ & \text{and } c(dP \cup Qd') \geq c(dP^- \cup Q^-d') \\ & \text{and } C(dV^-d') = \emptyset \end{aligned} \tag{3}$$

To accept dS^3d', only criteria that are concordant with dPd' can conceal criteria supporting a strict preference in favor of d' but criteria supporting a weak preference in favor of d' can be concealed by criteria concordant with either a strict or a weak preference in favor of d. At the same time, no discordant criterion should strongly disagree with this assertion.

Observe that these three relations get richer and richer, i.e., we have $S^1 \subseteq S^2 \subseteq S^3$, but less and less well-established.

It is worth noting that the proposed aggregation mechanism, which compares the size of various coalitions of criteria, does not require that criteria are defined on a common scale. Therefore, normalizing criterion scales, which is always somewhat arbitrary, is unnecessary in our approach.

6.2 Exploitation phase

Outranking relations are not necessarily transitive and do not lend themselves to immediate exploitation to get the final ranking. Therefore, we need exploitation procedures in order to derive the final document ranking. We propose the following procedure which finds its roots in (Roy and Hugonnard 1982). It consists in partitioning the set of documents into r ranked *classes* where each class C_h contains documents with the same score. This is coherent with specificity 4 of Sect. 3.2. Considering that s outranking relations $S^1 \subseteq \cdots \subseteq S^s$ have been defined, let:

- R be the set of potential relevant documents for a query,
- $F_i(d,E) = card(\{d' \in E : dS^id'\})$ be the number of documents in $E (E \subseteq R)$ that could be considered 'worse' than d according to the global relation S^i,
- $f_i(d,E) = card(\{d' \in E : d'S^id\})$ be the number of documents in E that could be considered 'better' than d according to S^i,
- $s_i(d,E) = F_i(d,E) - f_i(d,E)$ be the *qualification* of d in E according to S^i.

Each class C_h results from a *distillation process*. It corresponds to the last distillate of a series of sets $E_0 \supseteq E_1 \supseteq \cdots \supseteq E_r$ $(r \geq 1)$, where $E_0 = R \setminus (C_1 \cup \cdots \cup C_{h-1})$ and E_i is a reduced subset of E_{i-1} resulting from the application of the following procedure:

1. compute for each $d \in E_{i-1}$ its qualification according to S^i, i.e., $s_i(d, E_{i-1})$,
2. choose $s_{\max} = \max_{d \in E_{i-1}}\{s_i(d, E_{i-1})\}$, then
3. $E_i = \{d \in E_{i-1} : s_i(d, E_{i-1}) = s_{\max}\}$

The distillation stops either when $card(E_r) = 1$ or when $r = s$.

6.3 Illustrative example

This section tries to illustrate the concepts and procedures introduced previously. Let us consider a set of candidate documents $R = \{d_1, d_2, d_3, d_4, d_5\}$. Table 5 gives the performance vectors of the documents of R w.r.t. a family of four pseudo-criteria $F = \{g_1, g_2, g_3, g_4\}$. Indifference, preference and veto thresholds of these criteria are summarized in Table 6.

We retain the outranking relations of Eqs. 1 and 2 to carry pairwise contests of the aggregation phase. We hereafter give details about the computation of the outranking relations S^1 and S^2: we thus give in Tables 7–10 all matrices corresponding to the outranking (S_j), weak preference (Q_j) strict preference (P_j) and veto (V_j) relations for each considered criterion g_j. All these matrices derive directly from Tables 5 and 6. For instance, since $|g_1(d_2) - g_1(d_1)| = |0.7-0.8| = 0.1 \leq q_1$, then $d_2 S_1 d_1$ holds.

For the aggregation phase, let us consider the outcome of pairwise comparison of document d_2 against d_5: on the one hand, we have $C(d_2 S d_5) = \{g_1, g_2, g_4\}$ since we have $d_2 S_1 d_5$, $d_2 S_2 d_5$ and $d_2 S_4 d_5$, i.e., criteria g_1, g_2 and g_4 are concordant with $d_2 S d_5$. On the other hand, criterion g_3 does not support this assertion as shown in the matrix corresponding to relation S_3. Therefore, $d_2 S^1 d_5$ does not hold according to Eq. 1.

Criteria g_1 and g_2 are concordant with the assertion $d_2 P d_5$, therefore $c(d_2 P d_5) = 2$. At the same time, according to criterion g_3, d_5 is strictly preferred to d_2, therefore $c(d_2 P^- \cup Q^- d_5) = 1$. Moreover, performance difference w.r.t. the same criterion g_3 is larger enough to let veto occurs: $c(d_2 V^- d_5) = 1$. Finally, according to the definition of outranking relation S^2 of Eq. 2, although the concordance condition is met $(c(d_2 P d_5) \geq c(d_2 P^- \cup Q^- d_5))$ and is for establishing $d_2 S^2 d_5$, the discordance condition refutes this assertion $(c(d_2 V^- d_5) \neq 0)$, therefore, $d_2 S^2 d_5$ does not hold.

The computation of the global outranking relations S^1 and S^2 is reported in Table 11.

We now move to the illustration of the exploitation phase which is responsible for building the final ranking of the documents.

Observing that $F_k(d_i, R)$ (resp. $f_k(d_i, R)$) is given by summing the values of the ith row (resp. column) of the matrix of outranking relation S^k, the consensus ranking is obtained as follows:

Table 5 Documents profiles

	g_1	g_2	g_3	g_4
d_1	0.8	0.6	1	0.1
d_2	0.7	0.9	0.1	0.6
d_3	0.5	0.6	0.6	0.5
d_4	0.4	0.3	0.3	0.6
d_5	0.1	0.2	0.9	0.3

Table 6 Threshold values

	q_j	p_j	v_j
g_1	0.2	0.4	0.6
g_2	0.2	0.5	0.7
g_3	0.3	0.3	0.5
g_4	0.1	0.3	0.6

Table 7 Partial outranking relations

S_1	d_1	d_2	d_3	d_4	d_5
d_1	1	1	1	1	1
d_2	1	1	1	1	1
d_3		1	1	1	1
d_4			1	1	1
d_5					1

S_2	d_1	d_2	d_3	d_4	d_5
d_1	1		1	1	1
d_2	1	1	1	1	1
d_3	1		1	1	1
d_4				1	1
d_5				1	1

S_3	d_1	d_2	d_3	d_4	d_5
d_1	1	1	1	1	1
d_2		1		1	
d_3		1	1	1	1
d_4		1	1	1	
d_5	1	1	1	1	1

S_4	d_1	d_2	d_3	d_4	d_5
d_1	1				
d_2	1	1	1	1	1
d_3	1	1	1	1	1
d_4	1	1	1	1	1
d_5	1				1

Table 8 Partial weak preference relations

Q_1	d_1	d_2	d_3	d_4	d_5
d_1			1	1	
d_2				1	
d_3					1
d_4					1
d_5					

Q_2	d_1	d_2	d_3	d_4	d_5
d_1				1	1
d_2	1		1		
d_3				1	1
d_4					
d_5					

Q_3	d_1	d_2	d_3	d_4	d_5
d_1					
d_2					
d_3					
d_4					
d_5					

Q_4	d_1	d_2	d_3	d_4	d_5
d_1					
d_2					1
d_3					1
d_4					1
d_5	1				

Iteration 1: To get the first class C_1, we compute the qualifications of all the documents of $E_0 = R$ with respect to S^1. They are respectively 0, 1, 2, -2 and -1. For instance, since $F_1(d_3, R) = 3$ and $f_1(d_3, R) = 1$, we have $s_1(d_3, R) = 3 - 1 = 2$. Therefore $s_{\max}$ equals 2 and $C_1 = E_1 = \{d_3\}$ since d_3 is the only document of the first distillate.

Iteration 2: To run a new iteration and compute the next class C_2, we first remove document d_3 from the outranking matrices by removing its corresponding rows and columns in both matrices. We compute the new qualifications of the documents of the new starting set $E_0 = R \setminus C_1 = \{d_1, d_2, d_4, d_5\}$. They are respectively 0, 1, -1 and 0. Therefore, document d_2 having the maximum qualification 1 constitutes the only document of the second class C_2.

Iteration 3: To get the third class, we remove d_2 from both matrices of S^1 and S^2. The remaining documents d_1, d_4 and d_5 have the same qualification value 0. Thus, the first

Table 9 Partial strict preference relations

P_1	d_1	d_2	d_3	d_4	d_5
d_1					1
d_2					1
d_3					
d_4					
d_5					

P_2	d_1	d_2	d_3	d_4	d_5
d_1					
d_2				1	1
d_3					
d_4					
d_5					

P_3	d_1	d_2	d_3	d_4	d_5
d_1		1	1	1	
d_2					
d_3		1			
d_4					
d_5		1		1	

P_4	d_1	d_2	d_3	d_4	d_5
d_1					
d_2	1				
d_3	1				
d_4	1				
d_5					

Table 10 Partial veto relations

V_1	d_1	d_2	d_3	d_4	d_5
d_1					1
d_2					
d_3					
d_4					
d_5					

V_2	d_1	d_2	d_3	d_4	d_5
d_1					
d_2					
d_3					
d_4					
d_5					

V_3	d_1	d_2	d_3	d_4	d_5
d_1		1		1	
d_2					
d_3					
d_4					
d_5		1		1	

V_4	d_1	d_2	d_3	d_4	d_5
d_1					
d_2					
d_3					
d_4					
d_5					

distillate of this class is $E_1 = \{d_1, d_4, d_5\}$. We use the second outranking relation S^2 to reduce this set. The qualifications of the documents of E_1 are respectively 2, -1 and -1. Therefore, the second distillate is $E_2 = \{d_1\}$ and corresponds to the third class C_3.

Iteration 4: Computing the qualifications of the remaining documents d_4 and d_5 give the same value 0 in both S^1 and S^2. Therefore the last class C_4 corresponds to these documents: both relations do not permit building more refined ranked classes.

The consensus ranking is finally $\{d_3\} \rightarrow \{d_2\} \rightarrow \{d_1\} \rightarrow \{d_4, d_5\}$.

It is worth noting that using more standard aggregation operators lead to different rankings. For instance, supposing that document performances are normalized and that all criteria have similar weights, ranking documents according to the sum aggregation operator leads to the following ranking: $\{d_1\} \rightarrow \{d_2\} \rightarrow \{d_3\} \rightarrow \{d_4\} \rightarrow \{d_5\}$. More

Table 11 Global outranking relations

S^1	d_1	d_2	d_3	d_4	d_5
d_1	1				
d_2		1		1	
d_3			1	1	1
d_4				1	
d_5					1

S^2	d_1	d_2	d_3	d_4	d_5
d_1	1		1	1	1
d_2		1		1	
d_3		1	1	1	1
d_4				1	
d_5					1

particularly, document d_3 which is initially ranked first using the outranking approach is now ranked third according to the sum operator. This shows an important feature of outranking approaches: documents with acceptable and more balanced profiles (e.g., d_3) are preferred to documents with rather more contrasted profiles (e.g., d_1 and d_2) as shown in Fig. 3.

6.4 Complexity of our approach

Before presenting experimental results, we briefly investigate and comment the complexity of our approach.

Considering n documents to be processed, given a family of p fixed criteria, the computation of their scores on the criteria can be performed in linear time $O(n)$. The filtering phase also requires $O(n)$ time, whereas the ranking phase requires $O(n^2)$ time due to the computation of the various preference matrices during the aggregation phase. Remark, however, that the filtering phase drastically reduces the number of documents to be processed by the ranking phase, which makes the whole approach quite efficient.

7 Experiments and results

7.1 Test setting

To facilitate empirical investigation of the proposed methodology, we developed a prototype search engine, named WIRES, that implements a preliminary version of our

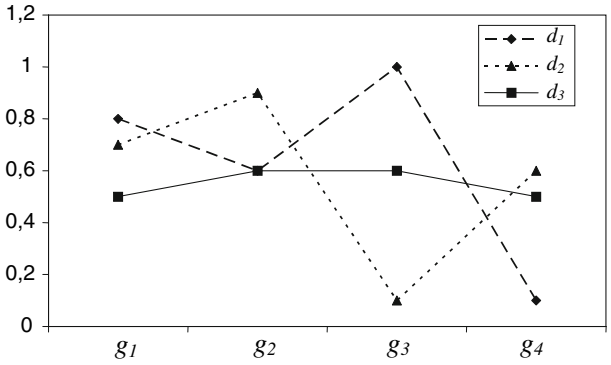

Fig. 3 Document profiles

multiple criteria approach. In this paper, we apply our approach to the Topic Distillation (TD) task of TREC-13 Web track (Craswell and Hawking 2004). In this task, there are 75 topics where only a short description of each is given. For the experiments, we translated each topic to a conjunctive query following most search engine strategies. We have built an inverted index of the '.GOV' TREC test collection where we consider word stems as index terms using the Porter stemming algorithm and discard common English stopwords. We also used the hyperlink structure of this collection to build link-based criteria.

At a first level, we had to define the set F of criteria, for which we used the following elementary features which are the main attributes used in the literature:

- tf_k: frequency of term t_k in document d,
- df_k: number of documents the term t_k occurs in,
- max tf: maximum frequency tf_k of all terms $t_k \in d$,
- $l_{k,a}$: a binary value which equals 1 if term t_k occurs in location L_a and 0 otherwise. The considered locations are the URL (L_1), the title (L_2), the keywords tag (L_3) and the description tag (L_4),
- $\Gamma^-(d)$: set of incoming hyperlinks to d,
- $Child(d)$: set of children documents of d. Document d' is in $Child(d)$ if it appears in a lower hierarchical level than d according to their site map,
- $prox$: proximity of query terms in document d. It corresponds to the size (number of terms) of the smallest text excerpt from the document that contains all the query terms. It equals 0 if not all the query terms are in d,
- ql: query length, i.e., the number of terms of the query,
- dl: document length, and
- $depth(d)$: depth of the URL of d, which is the number of intermediary sub-directories between document d and the root of its corresponding site map.

Based on these features, we defined the following candidate criteria:

- Frequency: For one-term queries (i.e., $q = t_k$), $g_1(d, t_k) = \frac{tf_k}{\max tf}$
- Position: For one-term queries, $g_2(d, t_k) = \sum_{a=1}^{4} l_{k,a}$
- Authority: $g_3(d) = card(\Gamma^-(d))$
- Prominence: $g_4(d) = card(Child(d))$
- Proximity: $g_5(d) = \frac{ql}{prox}$ if $prox \neq 0$, *and* 0 otherwise
- Document length: $g_6(d) = dl(d))$
- Rareness: For one-term queries, $g_7(d, t_k) = df_k$

For multi-terms queries, we used the average operator.

It is worth noting that the concrete choice of the features as well as the criterion family should be monitored to best correspond to the specific application context. For our experiments, we focused on features which capture the major well-known IR evidences. The exact definition of each criterion tries to capture intuitive preferences but remains somewhat arbitrary, thus we considered simple formulations, but more refined formulations can be used too.

In the TD task, a successful relevance ranking should favour 'good entry points' although they could contain little detailed information. This is captured by the prominence criterion (g_4).

For evaluation, we used the 'trec_eval' standard tool which is used by the TREC community to calculate the standard measures of system effectiveness which are Average precision (`AvP`),`R-precision` (`R-p`), Reciprocal rank (`r-r`) and Success@n (`S@n`) (see, e.g., Craswell and Hawking (2004)).

Our approach effectiveness is compared against some high performing official results from TREC-13 using the *paired t-test* which is shown to be highly reliable (more than the *sign* or *Wilcoxon* tests) according to Sanderson and Zobel (2005). In the experiments, significance testing is mainly based on the `t-student` statistic which is computed on the basis of the `AvP` values of the compared runs. In the tables of the following section, we have marked with an asterisk statistically significant differences.

7.2 Results

With the criteria described before, we performed several retrieval runs. In the first set of runs, we rank documents according to each criterion and report performances in Table 12. We aim at showing which criteria are really relevant for the TD task.

Table 12 shows that the run with the prominence criterion (g_4) performs significantly better than the others. Runs carried out with the first 4 criteria perform significantly better than runs carried out with the last 3. Moreover, the random run `random` performs better than the same 3. Therefore document length (g_5), proximity (g_6), and rareness (g_7) do not play an important role for the TD task.

In the second set of runs, we only considered the best four criteria, i.e., criteria g_1–g_4. In our baseline run (`mcm`), the set R of potentially relevant documents is obtained in two stages: we first use the boolean filter to identify a first set A which is then extended to a set A^+ that includes each document pointing to at least two documents in A. Many of the added documents should, in fact, correspond to good entry points to relevant sites. In the aggregating procedure of Sect. 6.1, each criterion is supposed to be a pseudo-criterion where indifference, preference and veto thresholds are set to 20%, 60% and 90%, respectively. These thresholds are set after some tunings carried with respect to TREC-12 Web track TD topics. We suppose that there is no information on the relative importance of criteria and use the outranking relation S^2 defined by (2). We implement the exploitation procedure of Sect. 6.2.

We now try to catch the impact of profile filtering on performance using the procedure presented in Sect. 5 which allows us to get a reasonably small set R of documents. We carried out some runs where we tried to get different numbers of filtered documents: for each run `mcm-filter-x`, x corresponds to the number of the filtered documents.

Table 12 Performances of single criterion runs

Run Id	AvP	R-p	r-r	S@1	S@5	S@10	Δ-AvP
prominence	12.37%	16.15%	46.42%	29.33%	74.67%	85.33%	–
authority	9.27%	10.41%	31.68%	18.67%	44.00%	57.33%	−25.03%*
position	7.30%	6.66%	21.62%	12.00%	29.33%	42.67%	−40.99%*
frequency	7.01%	6.49%	16.44%	6.67%	24.00%	37.33%	−43.36%*
random	3.17%	2.42%	9.90%	4.00%	10.67%	22.67%	−74.40%*
proximity	2.78%	2.14%	4.73%	0.00%	6.67%	9.33%	−77.56%*
rareness	2.27%	1.00%	4.24%	1.33%	2.67%	9.33%	−81.65%*
length	1.76%	0.22%	2.19%	0.00%	2.67%	2.67%	−85.74%*

* denotes statistically significant differences

Table 13 Impact of filtering procedure

Run Id	AvP	R-p	r-r	S@1	S@5	S@10	Δ-AvP
mcm	17.08%	18.37%	58.04%	45.33%	74.67%	81.33%	–
mcm-filter-1000	17.00%	18.37%	58.04%	45.33%	74.67%	81.33%	−0.46%
mcm-filter-800	16.83%	18.37%	58.04%	45.33%	74.67%	81.33%	−1.45%
mcm-filter-500	16.52%	18.34%	58.04%	45.33%	74.67%	81.33%	−3.26%
mcm-filter-50	15.65%	18.40%	58.04%	45.33%	74.67%	81.33%	−8.35%*

* denotes statistically significant differences

Table 14 Different aggregation strategies

Run Id	AvP	R-p	r-r	S@1	S@5	S@10	Δ-AvP
mcm	17.08%	18.37%	58.04%	45.33%	74.67%	81.33%	–
max	8.02%	7.70%	21.40%	8.00%	33.33%	50.67%	−53.02%*
min	10.74%	12.91%	47.20%	32.00%	70.67%	77.33%	−37.13%*
prod	12.06%	14.02%	53.66%	37.33%	74.67%	80.00%	−29.41%*
sum	13.45%	14.37%	51.78%	36.00%	66.67%	82.67%	−20.73%*

* denotes statistically significant differences

Table 13 shows that mcm-filter-x runs differs only with respect to AvP and R-p. All the other measures remain the same. This is because all these runs have the same ranking at the top. When we filter 50 documents, performance decreases rather significantly, whereas considering the R-p. measures, performance slightly increases. Performances do not significantly decrease with respect to those of mcm when we filter 1,000, 800 or 500 documents. We can conclude that filtering is beneficial for IR since it considerably reduces the size of the set of documents to be compared in the ranking procedure, and at the same time, it does not lead to significant performance drop.

We compare now our basic run mcm with other aggregation strategies.

In Table 14, we report performances of four aggregation operators which are max, min, sum and product operators. For these runs, documents performances are normalized so that

Table 15 Performance comparison with official runs

Run Id	AvP	R-p	r-r	S@1	S@5	S@10	Δ-AvP
mcm	17.08%	18.37%	58.04%	45.33%	74.67%	81.33%	–
uogWebCAU150	17.91%	20.30%	62.57%	50.67%	77.33%	89.33%	+4.84%
MSRAmixed1	17.80%	20.45%	52.79%	38.67%	72.00%	88.00%	+4.18%
MSRC04C12	16.45%	19.07%	53.39%	38.67%	74.67%	80.00%	−3.68%
humW04rdpl	16.28%	19.72%	55.31%	37.33%	78.67%	90.67%	−4.68%
THUIRmix042	14.66%	16.65%	39.54%	21.33%	58.67%	74.67%	−14.17%*
average	10.53%	12.84%	36.58%	23.87%	51.50%	61.82%	−38.37%*
median	11.52%	14.64%	39.99%	25.33%	58.00%	69.33%	−40.86%*

* denotes statistically significant differences

they range in the [0,1] interval. The best performing run is the `sum` run, but its performances are significantly worse than those of `mcm`. This shows that a total compensatory logic (e.g., `sum` and `prod` runs) as well as a non-compensatory logic (e.g., `max` and `min` runs) perform worse than a partial compensatory logic (e.g., `mcm` run) using outranking approaches for example.

We end this section by reporting performances of the official runs from TREC-13 (Craswell and Hawking 2004) and compare our approach accordingly.

In Table 15, we first report performances of the best runs of the first five teams which participated to the track. Then, we computed average and median performances of all the submitted runs. From this table, we can see that `mcm` has similar performances compared to those of the best ones. Moreover, `mcm` performs significantly better than the `average` or the `median` runs.

8 Conclusions

In this paper, we propose a multiple criteria framework for evidence combination where a set of candidate relevance criteria are proposed and used to determine how documents should be ranked using a set of decision rules.

The proposed approach overcomes limits of classical analytical retrieval formulas which do not allow considering complex logics when aggregating various criteria. It is also straightforward to show that the proposed approach, based on decision rules, fulfill intuitive and desirable formal requirements that any reasonable retrieval method should satisfy according to Fang et al. (2004) work. Interestingly, these authors show that none of the formulas used in the vector space model, the probabilistic model, or the language model, satisfies these requirements unconditionally.

From the first TREC experiments, this work seems to have the potential for high impact in the field of IR, given the possible application of evidence combination. It presents the advantage that it is applicable whatever is the collection under consideration provided that a pertinent criterion family is used. It also overcomes criteria heterogeneity problems by using a set of decision rules which are easy to grasp. Moreover, the proposed approach easily helps considering domain and context specific criteria in a natural way, rather than using complex formula which are difficult to interpret.

Approaches from multiple criteria decision theory, and especially outranking approaches, are generally used as an aid for decision makers. In the TREC context, there are various assessors judging documents with different and even conflicting preferences. This is the main explanation why it seems to be difficult to have significantly better performances. At the same time, we can outline an advantage of the proposed approach since we can easily carry out the study from the user perspective by setting a criterion family according to his/her preferences, giving rise to a personalized and valuable aid.

Future work will consist of additional experiments to strengthen the results. More specifically, applying our method in a human centered context would be an interesting extension of our work. Also, in this paper we considered that each criterion is not prevailing not negligible. But when there is some evidence that some criteria are more important than others without being able to assign precise values, specific outranking approaches such as Melchior (Leclercq 1984) are more appropriate.

References

Amento, B., Terveen, L. G., & Hill, W. C. (2000). Does "authority" mean quality? predicting expert quality ratings of web documents. In N. J. Belkin, P. Ingwersen, & M. -K. Leong, (Eds.), *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 296–303). Athens, Greece, July 2000. ACM Press.

Anh, V. N., & Moffat, A. (2002). Vector space ranking: Can we keep it simple? In *ACDS 2002: Proceedings of the Seventh Australasian Document Computing Symposium*, December 2002.

Bordogna, G., & Pasi, G. (2004). A model for a soft fusion of information accesses on the web. *Fuzzy Sets and Systems, 148*, 105–118.

Bordogna, G., Pasi, G., & Yager, R. (2003). Soft approaches to distributed information retrieval. *International Journal of Approximate Reasoning, 34*, 105–120.

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology, 54*(10), 913–925.

Boughanem, M., Loiseau, Y., & Prade, H. (2005). Rank-ordering documents according to their relevance in information retrieval using refinements of ordered-weighted aggregations. In M. Detyniecki, J. M. Jose, A. Nürnberger, & C. J. van Rijsbergen, (Eds.), *Adaptive Multimedia Retrieval* (Vol. 3877, pp. 44–54). *Lecture Notes in Computer Science*, Springer-Verlag.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks, 30*(1–7), 107–117.

Cao, G., Nie, J. -Y., & Bai, J. (2005). Integrating word relationships into language models. In R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat & J. Tait, (Eds.), *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 298–305). Salvador, Brazil, August 2005. ACM Press.

Craswell, N., & Hawking, D. (2004). Overview of the TREC-2004 Web Track. In *Proceedings of TREC'2004*.

Craswell, N., Robertson, S. E., Zaragoza, H., & Taylor, M. J. (2005). Relevance weighting for query independent evidence. In R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, & J. Tait, (Eds.), *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 416–423). Salvador, Brazil, August 2005. ACM Press.

Dubois, D., & Prade, H. (1984). Criteria aggregation and ranking of alternatives in the framework of fuzzy set theory. In H. Zimmermann, L. Zadeh, & B. Gaines, (Eds.), *Fuzzy sets and decision analysis* (Vol. 20, pp. 209–240). TIMS Studies in the Management Sciences.

Fang, H., Tao, T., & Zhai, C. (2004). A formal study of information retrieval heuristics. In M. Sanderson, K. Järvelin, J. Allan, & P. Bruza, (Eds.), *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 49–56). Sheffield, UK, July 2004. ACM Press.

Fox, E. A., & Shaw, J. A. (1994). Combination of multiple searches. In *Proceedings of TREC'3*.

Frakes, W., & Baeza-Yates, R. (1992). *Information retrieval: Data structures and algorithms*. Prentice Hall.

Granka, L. A., Joachims, T., & Gay, G. (2004). Eye-tracking analysis of user behavior in WWW search. In M. Sanderson, K. Järvelin, J. Allan, & P. Bruza, (Eds.), *SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 478–479). Sheffield, UK, July 2004. ACM Press.

Hawking, D., Craswell, N., Bailey, P., & Griffihs, K. (2001). Measuring search engine quality. *Information Retrieval, 4*(1), 33–59.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of ACM, 46*(5), 604–632.

Kraaij, W., Westerveld, T., & Hiemstra, D. (2002). The importance of prior probabilities for entry page search. In *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 27–34). Tampere, Finland, August 2002. ACM Press.

Leclercq, J. P. (1984). Propositions d'extensions de la notion de dominance en présence de relation d'ordre sur les pseudo-critères. *Revue Belge de Recherche Opérationnelle, de Statistiques et d'informatique, 24*, 32–46.

Pasi, G., Bordogna, G., & Villa, R. (2007). A multi-criteria content-based filtering system. In W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, & N. Kando, (Eds.), *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 775–776). Amsterdam, The Netherlands, July 2007.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., & Gatford, M. (1994). Okapi at TREC-3. In *Proceedings of TREC'3*.

Roy, B. (1989). Main sources of inaccurate determination, uncertainty and imprecision. *Mathematical and Computer Modelling, 12*(10–11), 1245–1254.

Roy, B. (1991). The outranking approach and the foundations of ELECTRE methods. *Theory and Decision, 31*, 49–73.

Roy, B., & Hugonnard, J. (1982). Ranking of suburban line extension projects on the Paris metro system by a multicriteria method. *Transportation Research, 16A*(4), 301–312.

Salton, G., Fox, E. A., & Wu, H. (1983). Extended boolean information retrieval. *Communications of the ACM, 26*(11), 1022–1036.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. McGraw-Hill.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*(11), 613–620.

Sanderson, M., & Zobel, J. (2005). Information retrieval system evaluation: Effort, sensitivity, and reliability. In R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, & J. Tait, (Eds.), *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 162–169). Salvador, Brazil, August 2005. ACM Press.

Savoy, J., & Rasolofo, Y. (2000). Report on the TREC-9 experiment: Link-based retrieval and distributed collections. In *Proceedings of TREC'9*.

Van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). Dept. of Computer Science, University of Glasgow, London: Butterworths.