SETIT 2005
3rd International Conference: Sciences of Electronic,
Technologies of Information and Telecommunications
March 27-31, 2005 – TUNISIA

A multicriteria paradigm of relevance for the Web Information Retrieval problem

FARAH Mohamed* and VANDERPOOTEN Daniel*

Lamsade, University of Paris IX, France

farah@lamsade.dauphine.fr
vdp@ lamsade.dauphine.fr

Abstract : We consider the problem of ranking Web documents within a multicriteria framework and propose a novel approach for the purpose. We focus on the design of a set of criteria aiming at capturing complementary aspects of relevance, while acknowledging that each criterion has a limited precision. Moreover, we provide algorithmic solutions to aggregate these criteria to get the ranking of relevant documents, while taking into account the specificities of the Web information retrieval problem. We report on results of preliminary experiments that give first justification of the pertinence of the proposed approach to improve retrieval effectiveness.

Keywords : Decision Criteria, Information Retrieval, Relevance.

1 Introduction

The aim of a Web information retrieval (WIR) system is to *effectively* search and retrieve *relevant* documents from the Web.

Relevance has become a major area of research [Van Rijsbergen, 1979] [Salton, 1989] , nevertheless, it still remains a not well understood concept [Froehlich, 1994] [Cosijn and Ingwersen, 2000], especially because there are many kinds of relevance depending on the information resources used (Documents, Surrogates, etc.), the user problem representation (Information Need, Request, Query, etc.), as well as the general context of the search (Topic, Task, Knowledge, etc.) [Mizzaro, 1997].

At the same time, while efficiency, which refers to the ability of a system to provide results within reasonable response times, can be considered as satisfactory in most current search engines [Gudivada et al., 1997], retrieval effectiveness, which refers to the ability of a system to deliver the most relevant results first, remains a challenging issue. In fact, several comparative studies report limitations in many search engines performances. For instance, [Gordon and Pathak, 1999] evaluated 8 search engines using 33 requests and found that half of them returned only one relevant item. Moreover, the probability of the first page to be relevant was around 15%. From [Hawking et al., 2001], where 20 search engines were evaluated using 54 requests, precision for the 20 first documents returned was around 0.5 for the best search engine.

It is worthy to note that effectiveness depends widely both on which kind of relevance is measured by the WIR system and which kind of relevance judgement is adopted, which is the assignment of values of relevance by one or more judges.

In the current approaches, document relevance is conveyed by a score that possibly aggregates several factors, where aggregation is either based on a weighted sum or a lexicographic order operator [Belkin et al., 1995], or a ranking fusion heuristic [Cook and Kress, 1985] [Dwork et al., 2001].

In this paper, we propose a general *multicriteria framework* for the definition of relevance that we model using a family of adequate criteria aiming at capturing complementary aspects of relevance, while acknowledging that each family leads to a definition of a specific relevance and vice versa. Moreover, we give insights on the way these criteria should be aggregated in order to get effective ranking of the relevant documents.

We illustrate our approach by considering collections of Web documents and give first justifications on the pertinence of the proposed approach.

This paper is organized as follows. We begin with a brief review of the past research concerning factors used to derive the ranking of documents in response to a query (section 2). We then announce the mutilcriteria hypothesis (section 3) according to which relevance is best captured within a multicriteria framework and retrieval effectiveness is enhanced when these factors are cleverly considered. This framework is intro-

duced in section 4 where we try to model relevance by a set of criteria, while acknowledging that many relevances can be considered. Section 5 shows how to aggregate such criteria while considering specificities of the problem and how to build a comprehensive ranking. Experimental results are presented in section 6. Conclusions are provided in a final section.

2 Relevance factors

Relevance is reflected by the factors that are considered, as well as the way they are used to derive the final ranking.

Web documents consist mainly of text, therefore factors that have solidly been established as important in the Text information retrieval literature [Harman, 1986] [Salton, 1989] [Salton and Buckley, 1988] [Kilgarriff, 1996], can be considered as the basis for criteria design. Moreover, with the advent of the Web, Web-specific factors have proven to improve effectiveness, especially in the last Text REtrieval Conference (TREC) series [Voorhees and Harman, 2000, 2002,2003]. .

These factors can be split into the following categories depending on the sources of evidence that are considered :

2.1 Text-based factors

These factors exploit the occurrence properties of the terms in the documents and the collection.

Documents are considered as bags of terms which have varying informative values depending mainly on the following factors : a) Term-Document frequency (tf) which is the number of term repetitions within a document [Luhn, 1958]. It is, in fact, widely accepted that frequent terms are more important in describing the semantic content of a document, b) Term-Collection frequency (cf) is the number of documents in the collection, in which the term occurs. It is assumed that terms with high cf do not help discriminating between relevant and non-relevant items. So infrequent terms are more valuable. In the literature, this factor is always measured by the transformation $\log(\frac{N}{cf})$, namely the inverse document frequency (idf), where N is the size of the collection [Sparck Jones, 1972], c) Term discrimination value is the difference of the average distance between documents with and without the term. A term has a good discrimination value if the documents to which the term belongs become better distinguished from the rest of the collection when the term is assigned to them. The average distance (AvD) before and after the assignment of terms can be computed in various ways such as

$$AvD = \frac{1}{N(N-1)} \sum_{\substack{i=1 \\ j=1 \\ i \neq j}}^{N} sim(d^i, d^j)$$

where $sim(d^i, d^j)$ represents the similarity coefficient between documents d^i and d^j according to one document representation. d) Signal-noise ratio (sn) is given by

$$sn = \log_2(TF) - \sum_{i=1}^{cf} (\frac{tf_i}{TF} \times \log_2(\frac{TF}{tf_i}))$$

where tf_i is the frequency of occurrence of the term in document d^i and $TF = \sum_{i=1}^{cf} tf_i$ is the total number of occurrences of the term in all the documents. It favors terms that are perfectly concentrated in particular documents as well as terms which occur unevenly in the documents of the collection. This measure is borrowed from the theory of information where the least predictable terms carry the greatest information value [Shannon, 1951].

Many other factors have been used attempting to improve search performances : a) Term locality (tl) tries to give more credits to terms present in specific structured portions of the documents such as the title, b) Term proximity (tp) captures the nearness of query terms within documents. It is, in fact, assumed that documents in which query terms are close together, should best meet the information need behind the query, c) Document length (dl) is the size of documents. We may need to retrieve short documents indeed, since they usually describe less topics and are therefore more specific to the query.

2.2 Probabilistic factors

These factors need user input in terms of relevance judgements with respect to query results. These judgements are used to estimate the probabilities that terms occur in relevant and non relevant documents :

Let :

- R : 'The document is relevant to the query',
- $\overline{\mathrm{R}}$: 'The document is **ir**relevant to the query',
- T : 'The query term t belongs to the document',
- r = # documents relevant to the query,
- $\mathrm{r_t}$ = # documents relevant to the query *and* contain term t, and

Then the previous probabilities are computed as follows :

$$P(T|R) = \frac{r_t}{r}$$
$$P(T|\overline{R}) = \frac{cf - r_t}{N - r}$$

and are used to build measures such as the term-relevance (tr) weight [Robertson and Spark Jones, 1976] :

$$tr = \frac{P(T|R) \times P(\overline{T}|\overline{R})}{P(T|\overline{R}) \times P(\overline{T}|R)} = \frac{\frac{r_t}{r-r_t}}{\frac{cf-r_t}{N-cf-(r-r_t)}}$$

2.3 Link-based factors

The hyperlink structure of the Web graph encodes a considerable amount of latent human judgments. In [Kleinberg, 1999], it is claimed that this type of judgments encompass the notion of authority which is very useful for document ranking. In fact, by including a link to document d', the author of document d is supposed to have conferred authority to d'. Roughly speaking, authority aims at capturing the notion of the quality of a Web document, where quality rather means popularity.

The analysis of similar link structures with the goal of understanding their social or informational organization has its roots in a number of areas such as in social networks (e.g. [Katz, 1953], [Hubbel, 1965]) or bibliometrics (e.g. [Garfield, 1978], [Pinski and Narin, 1976]).

In the context of the Web, Kleinberg's HITS algorithm [Kleinberg, 1999] as well as PageRank [Brin and Page, 1998] and Clever [Chakrabarti et al., 1998] are the basic link-based algorithms proposed in the literature. Factors pertaining to the popularity or visibility of documents can be induced by simple measures such as documents in-degrees or out-degrees as well as the scores computed by such algorithms.

For instance, the HITS algorithm assigns to each document two scores; a hub score (hub), which refers to its status as a pointer to other documents, and an authority score ($auth$), which refers to its status as a pointed document from outside. These scores are computed using an iterative algorithm with the following update functions at iteration k :

$$auth(d)^{<k>} = \sum_{d' \in D} b(d', d) \cdot hub(d')^{<k-1>}$$
$$hub(d)^{<k>} = \sum_{d' \in D} b(d, d') \cdot auth(d')^{<k>}$$

where $b(d, d')$ is 1 if d points to d' and 0 otherwise.

Factors presented in sections (2.1), (2.2) and (2.3) are used differently in a wide range of ranking algorithms leading to various definitions of relevance. For instance, in the vector space algorithm, relevance is measured by the cosine of the angle between vectors representing documents and queries [Salton, 1968] [Salton et al., 1975] :

$$\frac{\sum_{t \in q} w(d,t)\, w(q,t)}{\sqrt{\sum_{t \in d}(w(d,t))^2 \sum_{t \in q}(w(q,t))^2}}$$

where $w(d,t)$ ($w(q,t)$) are the weights assigned to a document (a query) due to query term t. These weights use mainly tf and idf factors, in the following way : $w(d,t) = tf \times idf$.

Many heuristics have been developed in order to consider a lot of these factors in the ranking. Some heuristics use more or less complex formulas to combine individual scores induced by various factors [Belkin et al., 1995]. Others use factors in an ad-hoc manner such as considering search results induced by some scoring functions and reranking them using techniques such as spreading activation or probabilistic argumentation as well as using some variations of link-based algorithms [Singhal and Kaszkiel, 2001] [Savoy and Rasolofo, 2001].

3 The multicriteria hypothesis

Different factors are available for capturing relevance of documents. It seems to be widely acknowledged that the way these factors are combined is either unknown, especially for most of the commercial search engines, or theoretically questionable.

Moreover, each factor family suffers from specific difficulties. Text-based factors have difficulties when dealing with the vocabulary problem, given the underlying ambiguity of natural languages [Furnas et al., 1988]. Maybe the most difficult vocabulary issues are synonymy and polysemy. Synonymy refers to situations where the same concept can be expressed using various terms. Polysemy refers to situations where the same term has different and even contradictory meanings and interpretations. These problems and many others render the process of retrieving documents, matching query terms, an uncertain process. In addition, document content do not encompass information that allows to properly assess its quality.

Link-based factors suffer from 'incomplete' link topology induced by 'emerging' communities, which penalizes recently introduced Web documents as well as high specific domain documents. Besides, these approaches are faced with noisy link topology since they fail in distinguishing various link types. Indeed, except for citation links, all navigational, commercial, and spam links are spurious informational links and have nothing to do with the conferral of authority.

In this work, we agree 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 deterio-

rate retrieval effectiveness when used by themselves. We therefore claim that being able to make effective use of all the possibly available and computable information is a desirable pattern, and can significantly improve retrieval effectiveness. This will be referred to as the 'multicriteria hypothesis'.

We hereafter present a formal framework to model relevance.

4 Relevance framework

According to the multicriteria hypothesis, relevance is multidimensional and should be modeled by a criteria family.

A criterion is the basis for *partial relevance assessment* as to whether a document is better or worse than some other document. It is restricted to a specific *point of view* where many factors can be considered and combined.

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 :

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

Hereafter, we will see that there is no one single definition of relevance and each criteria family models a specific kind of relevance.

Besides, even deemed acceptable to capture relevance, these criteria have a limited significance in the sense that a slight difference in the performance for two documents should not lead to discriminate them, especially when different formulations and implementations of the criterion are acceptable.

Also, we can distinguish criteria types that are specific for the WIR problem and need more thorough analysis.

4.1 Relevance dimensions

Each criteria family leads to a specific kind of relevance depending on the following three dimensions : 1) The information resources used in the relevance evaluation process. We have three entities indeed : a) Documents which are the physical entities that the user will see after querying the system., b) Surrogates which are the representations of documents (e.g. title, keywords, etc), and c) Information which is the difference in the user's knowledge after reading documents. 2) The representation of *the visible part of the user's problem*. Actually, the user is faced with a problematic situation that needs to be solved with information beyond his knowledge scope (Belkin speaks about the 'Anomalous states of knowledge' of the user [Belkin et al., 1982]). He perceives this *information need* and get a cognitive perception of the need. This is the *perceived need*. Then the user expresses this perceived need in a *request* using a natural language. Finally, he formalises this request in a *query* according to the system language, such as the boolean formalism. 3) The representation, if any, of *the **in**visible part of the user's problem* which corresponds to the general context of the search. It could be the activity the user will perform with the retrieved documents, the subject area which is the only interesting for the user, the documents already examined, etc.

Each 3-tuple of the above mentioned dimensions influence the different levels of the *relevance assessment process*.

In the first level, which is mainly a *filtering level*, *potentially relevant documents* can be derived using both of the parts of the user's problem representations. These will reduce the *noise* in the retrieved set of documents, where noise is the percentage of the non relevant items that are retrieved.

The second level consists of the identification of the various factors affecting relevance understanding giving the information sources that are considered. Thus, if only titles are considered for document representations, it is not possible to use factors relating to term occurrences or document characteristics. Besides, retrieving images or video sequences differs greatly from retrieving textual documents since each kind of information have specific features that others do not posses. For example, for images, factors such as the texture can be considered whereas texts do not have this feature.

The third level, which is the ranking level, tries to aggregate pairwise comparisons of documents with respect to each criterion to get a *global relevance assessment*. Having information on the general context of the search would vafour some criteria upon others, for instance.

Although there are many definitions of criteria families, based on the same factor set, we should nevertheless try to fulfill the following desirable requirements :

- Each criterion should be intelligibly designed so that comparisons remain meaningful.
- All the factors deemed to be important in comparing documents should be captured by the criteria. In fact, we should avoid situations where two documents are considered equivalent with respect to the criteria family, whereas considering individual factors leads to discriminate them.
- We should avoid redundancy, i.e. we should not consider the same factors more than once and therefore, it is better to have independent criteria in order not to unduly favour factors upon others.
- Criteria should be low in number in order to have a synthetic vision of each document, namely its profile.

4.2 Modelling imprecision

It is often inadequate to consider that slight differences in evaluation should give rise to clear-cut distinctions.

In order to model imprecision underlying criteria design, we use different thresholds [Roy, 1989] :

- An indifference threshold allows for two close-valued documents to be judged as equivalent although they do not have exactly the same performances 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.

A criterion g_j, having an indifference threshold q_j and a preference threshold p_j, is called pseudo-criterion.

Comparing two documents d and d' according to a pseudo-criterion g_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 $dS_jd' \Leftrightarrow g_j(d)+q_j \geq g_j(d')$, whose semantics is 'is at least as relevant as'.

4.3 Specific WIR criteria

In the WIR context, we can distinguish two types of criteria :

- *Query-dependent criteria* whose evaluation depends on the specific query formulation, and
- *Query-independent criteria* which mainly refer to characteristics of the document such as its length.

It is obvious that the design of query-dependent criteria is more complex than the design of query-independent criteria and thus needs more thorough analysis, especially when the query language is the boolean formalism. In fact, the valuation of such criteria depends on the query type :

- One-term query : The query is a word or a phrase.
- Conjunctive query : The query is a conjunction of terms.
- Disjunctive query : The query is a disjunction of terms.
- Complex query : Any combination of the previous types.

We can distinguish 3 levels of computations to evaluate query-dependent criteria :

- Level 1 : For one-term queries, criteria building has no specific difficulties. For example, if g defines a term frequency criterion, computing the score of a document according to this criterion consists roughly in counting the number of term occurrences within it.
- Level 2 : For conjunctive or disjunctive queries, each literal of the query formula gives rise to a sub-criterion computation carried in the same way as in the first level. To get an overall evaluation, we should perform a sub-aggregation of these partial measures.
- Level 3 : For complex queries, we have more than one logical operator in the query formula. To evaluate a document accordingly, we proceed as follows :
  - We transform the query formula into its disjunctive or conjunctive normal form, and
  - We evaluate documents upon each component in the same way as in level 2, then sub-aggregate the resulting outcomes accordingly.

Carrying sub-aggregations could be monitored by setting properties that should be met by the aggregation operators or procedures to be used.

Some natural properties could be required such as continuity, monotonicity and idempotence. Some other natural properties could be questioned such as symmetry. This happens when we accept that the order of query terms has a specific significance.

Specific properties should be met depending on the query type. We can chose between three classes of aggregation operators, each possessing distinct behaviour : conjunctive operators, disjunctive operators and compensative operators [Dubois and Prade, 1984].

Moreover, depending on the type of the criterion scale, some aggregation operators cannot be used. In fact, aggregation on ordinal scales should be limited to operations involving comparisons only, such as medians, while linear combinations are allowed for interval scales.

For example, let consider a conjunctive query $q = t_i \wedge t_j$ and suppose we are computing the score of a document d according to the normalized version of the term frequency criterion (tf/tf_{max}) which is evaluated on a ratio scale between 0 and 1. If we decide to aggregate the partial measures $g(d,t_i) = x_i$ ($g(d,t_i)$

being the performance of d with respect to the query $q' = t_i$) and $g(d, t_j) = x_j$ using an aggregation operator h which respects the natural properties, including symmetry, as well as the following ones :

$$h(1,1) = 1$$
$$h(0,0) = 0$$
$$h(x_i, x_j) \leq min(x_i, x_j)$$

then operators like $min(x, y)$ and $x \times y$ are acceptable whereas others like $max(x, y)$ and $x + y$ are not.

At level 3, we have to transform the query into a normal form first such as the disjunctive normal form (DNF) :

$$q = \bigvee_{i=1}^{s} cc_i$$

where

$$cc_i = \bigwedge_{j=1}^{t} t_j$$

is a conjunctive component, i.e. a conjunction of terms.

Then to compute the score of one document, we apply sub-aggregation of level 2 for each conjunctive component then for q.

5 Ranking procedure

In our context and in order to get a global relevance model on the set of documents, we have two main alternatives.

The first alternative consists in building a single synthesizing criterion

$$g(d, q) = h(g_1(d, q), g_2(d, q), \ldots)$$

where h is a more or less complex operator such as the weighted sum. In this case, we need precise inter-criterion information such as weights.

We discard this approach for the following reasons :

- Assigning weights to criteria is largely arbitrary. In fact, each criterion should model a different point of view. Therefore, criteria could be very heterogenous, and thus difficult to compare using such weights corresponding largely to substitution rates.
- When we use such approaches, we should accept the structural property stipulating that substitution rates are constant all over the scale. This implies that all the scales are ratio scales which is not always true.
- Imprecision underlying criteria measurements could not be considered. In fact, criteria design is an uncertain process and documents performances on each criterion are approximate values.

We pursue a different approach of ranking which consists of two phases.

5.1 Aggregation phase

This phase takes the partial preference assessments induced by the criteria family and aggregates them into one or more overall preference assessments using *preference relations* that define relevance. Such binary relations represent a global preference model allowing the comparison between each pair of documents.

Outranking approaches are relevant in this context since they permit considering imprecision in document evaluations. They are based on a partial compensatory logic using the notions of concordance and discordance [Roy and Bouyssou, 1993].

More precisely, we aim at constructing an overall outranking relation S, whose meaning is 'is at least as relevant as'. In order to accept the assertion dSd', the following conditions should be met :

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

In order to define such conditions, we usually need information on the relative importance of criteria. In the context of the Web, this information is usually difficult to grasp from the user and giving any values for such weights would be largely arbitrary (even assigning identical values). Instead we propose to build outranking relations under the assumption that each criterion of the family is neither prevailing nor negligible. Therefore, in order to accept the assertion dSd' we shall use simple conditions referring to the number of criteria *supporting* or *refuting* this assertion. Obviously, the conditions for defining this support may be more or less demanding, resulting in different relations. We hereafter present an approach which has its roots in [Roy and Hugonnard, 1982], where importance of criteria could not be assessed.

In order to define various outranking relations, let

- $\{g_1, \ldots, g_p\}$ be a family of p criteria,
- H be an overall preference relation (where H is P, Q or I),
- 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 report here 4 outranking relations based on decreasing restrictive conditions :

– S_1:

$$dS_1d' \Leftrightarrow card(C(dSd')) = p$$

In this case, all the criteria should be concordant with the assertion.

– S_2:

$$dS_2d' \Leftrightarrow card(C(dPd')) \geq card(C(d'Qd)) \quad and\ C(d'Pd) = \emptyset$$

To accept dS_2d', 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 $d'Pd$.

– S_3:

$$dS_3d' \Leftrightarrow card(C(dPd')) \geq card(C(d'P \cup Qd))$$

To accept dS_3d', there should be more criteria concordant with dPd' than criteria supporting a strict or weak preference in favor of d'.

– S_4:

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

To accept dS_4d', there should be more criteria concordant with dPd' than criteria concordant with $d'Pd$.

It is easy to establish that :

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

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

5.2 Exploitation phase

When we use a single synthesizing criterion, we get a complete preorder of the set of potentially relevant documents.

Outranking relations, in contrast, do not lend themselves to immediate exploitation to get the ranking as they are not necessarily transitive. Therefore, additional exploitation procedures should be devised to complement aggregating procedures. These procedures consist in elaborating a complete preorder Z based on the already built outranking relations. To do so, we partition the set of potential candidates D into r ranked classes where class C_h encloses documents that are considered ex aequo. C_1 stands for the best class.

Let

– $E \subseteq R$ be a subset of potential relevant documents for the query,
– $F_i(d, E) = card(\{d' \in E : dS_id'\})$ be the set of documents in E that could be considered as 'worse' than d according to S_i,
– $f_i(d, E) = card(\{d' \in E : d'S_id\})$ be the set of documents in E that could be considered as '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* on the base set $E_0 = R \setminus (C_1 \cup \ldots \cup C_{h-1})$. This process iterates over the outranking relations, starting from a well established relation, to get a reduced subset, called distillate, of documents with equivalent relevance.

At iteration i, we use the outranking relation S_i to get a reduced distillate E_i from E_{i-1}, using the following procedure :

1. Compute for each $d \in E_{i-1}$ its qualification 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 last distillate gives rise to a new class of ex-aequo that can consist of a single document. The distillation process stops immediately when $card(E_i) = 1$.

6 Experiments and results

6.1 System description

To facilitate empirical investigation of the proposed methodology, we developed a prototype search engine implementing a basic multicriteria approach, named Wires. The software is entirely implemented in java SDK1.4.1 using JBuilder 6 Enterprise Trial IDE. It has mainly three agents : a filter, a spider and a ranking agent. The filter determines the 'base set' of documents to be ranked by Wires for each query. In fact, in the experiments, we selected a widespread commercial search engine to retrieve the potentially relevant documents, and retain a reduced subset of the 300 best ranked HTML documents. This is the base set. The spider retrieves these documents directly from the Web and the ranking agent operationalized the muticriteria algorithm.

In order to design the criteria family, 1) we considered many document surrogates : a) the URL, b) the title, c) the keywords tag, d) the description tag, e) the body terms, f) the term positions within the body text, g) the anchor terms used in the hyperlink when linking to the current document, and h) the document indegree; 2) Queries are considered as boolean formulas where conjunctions are retained since it is the default operator of the considered search engine; and 3) We supposed that no information on the context of the search is available, especially to derive the relative importance of criteria or orient research in a specific direction. Therefore we considered the following criteria family :

– 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 conjunctive queries, we retain the $\min$ operator, whereas for disjunctive queries, we use the $\max$ 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 one-term queries, we first compute a partial performance with respect to each location L_i :

$g_2(d, t, L_i) =$ *the 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 :

$$g_2(d, q) = \frac{\sum_{i=1}^{5} k_i g_2(d, q, L_i)}{\sum_{i=1}^{5} k_i}$$

where $k_1 = k_2 = k_3 = 2$ and $k_4 = k_5 = 1$. This is because the keywords and description tags are not always available for documents.
For conjunctive, disjunctive and complex queries, the overall score is computed by

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

where n_q in the number of query terms.
We do not make distinctions between conjunctions and disjunctions because we judge we should not penalize documents where query terms are missing in some specific locations.

– g_3 is a popularity measure. Currently, only in-degrees are used :

$$g_3(d, q) = \frac{\log(indeg(d) + 1)}{\log(indeg_{\max} + 1)}$$

Another alternative would be to compute authority scores using the HITS algorithm.

– g_4 uses the dl factor. We consider, in fact, that relevant documents should be neither too short, in which case, there is nothing to learn therein, nor too long, in which case, the document is too general and contains much irrelevant information. g_4 is given by :

$$g_4(d, q) = \frac{1}{\log(|dl - \hat{d}| + 1)}$$

where $\hat{d}(= 1500)$ is the best length for a relevant document.

– g_5 is a proximity criterion. It is inversely proportional to the minimal span of query terms for a document, i.e. the smallest text excerpt from the document that contains all the query terms. It is not applicable for one-term queries.

g_1 and g_2 are query-dependent criteria and are set so that they respect some desired properties that we judge relevant.

Each criterion is supposed to be a pseudo-criterion where indifference and preference thresholds are set respectively to 20% and 40%. Setting such values is partly arbitrary but reflects our willing to limit the discrimination power of criteria. Ignoring such thresholds would be much more arbitrary.

The four outranking relations S_1, S_2, S_3 and S_4 are considered for the aggregation. To get the ranking, we retain the exploitation procedure presented in section 5.2.

6.2 Results

In the experiments, we consider the 300 first documents retrieved by the Reference Search Engine (RSE) with respect to some example queries. In this paper, we mainly comment the following queries : 'Web information retrieval' (Q1), 'multiple criteria methods' (Q2) and 'java tutorial' (Q3). Each time, we translate queries into conjunctive formulas since the RSE uses AND semantics.

After running each query, we analysed document rankings given by the two algorithms and try to detect and explain similarities and noticeable rank reversals.

First of all, we observed that both rankings were significantly different across all the queries (20 queries are used). Spearman's rank correlation coefficient was 25% in average. Moreover, the average intersection in the top 100 was about 45% whereas the average intersection in the top 20 was about 20%.

We report in table 2 statistics from the comparisons of both rankings with respect to queries Q1, Q2 and Q3.

	Q1	Q2	Q3
r_s	26,56%	29,91%	9,43%
Wires^{20} | RSE^{20}	5	5	2
RSE^{20} | Wires^{20}	4	5	5
Wires^{20} | RSE$_{20}$	0	0	2
RSE^{20} | Wires$_{20}$	1	0	0
rank$\searrow$	(16,300)	(11,211)	(12,253)
rank$\nearrow$	(191,14)	(289,33)	(283,2)

Table 2: Comparison of the rankings

where r_s is the Spearman's rank correlation.

In table 2, 'Wires^{20} | RSE^{20}' gives the number of documents retrieved in the 20 first items in Wires

among the 20 first positions in the RSE. In the same way, 'Wires^{20} | RSE$_{20}$' gives the number of documents retrieved in the 20 first items in Wires among those ranked in the 20 last positions in the RSE. The 'rank$\searrow$' ('rank$\nearrow$') gives the document associated with the highest rank fall (raise) where the first element is its ranking in RSE whereas the second gives its ranking in Wires.

For the query 'Web information retrieval', both systems return the same first document 'A survey on Web information technology - Huang (ResearchIndex)' which is clearly very relevant to the query as it gives a description of the paper of Lan Huang as well as some pointers to related topics. This document has the profile $(0.67,\ 0.2,\ 0.71,\ 0.33,\ 1)$ which corresponds to its performance on criteria $g_1, \ldots, g_5$ respectively and which is a good one indeed, *in comparison to the other profiles*.

The highest rank reversal occurred for document 16, that is ranked 16th in RSE. It is, in fact, ranked last in Wires. This document untitled 'WAIS access through the Web' is a short decription of WAIS. Its rank was lowered because of its low performance on criteria g1 (=0) and g4 (=0) since the query term 'Web' doesn't occur within the body text. At the same time, we observed rank reversal in the opposite direction since document 191 'Machine Learning Applied to Information Retrieval' is ranked 14th is Wires. It is very relevant and contains rich bibliographic references from Rick Belew, Peter Turney and Scott Weiss on information retrieval research. It has the following profile $(0.18,\ 0.08,\ 0.39,\ 0.69,\ 0.25)$ which is a good profile.

For the query 'multiple criteria methods', document 11 'Interactive multiple criteria decision methods : an investigation and an approach' is ranked 211 in Wires. In fact, it is a very short document (g4=0.3 since dl=71) without any valuable information.

For the query 'java tutorial', the first document returned by Wires is document 119 'Tutorials java help' which has profile $(0.4,\ 0.4,\ 0.5,\ 0.35,\ 1)$. It is a portal for tens of valuable tutorials on java.

Document 12 which is a tutorial about 'Vector Cross Product', is ranked 253 in Wires due its low performance on g1 (=0.08) and g4 (=0.3). Besides, document 283 is ranked 2nd in Wires. It gives 1,280 valuable links to online Java books, training slides, and tutorials. It has the profile $(0.34,\ 0.4,\ 0.2,\ 0.48,\ 1)$ which is a good one.

Our findings constitute a first justification of the value of the proposed approach. Particularly, we show that bad performances on one criterion could not be compensated by very good scores on the remaining criteria. In fact, each criterion is supposed to capture one dimension of relevance and thus each document should have enough good performances on each criterion to be in the top retrieved documents.

7 Conclusions

In this paper, a new multicriteria framework of relevance is presented where relevance is modeled by a family of criteria that differs depending on the information resources used, the user problem representation as well as the general context of the search. We illustrated our approach by considering collections of Web documents, but different document types can be considered, and gave first justifications on the pertinence of the proposed approach. We gave insights on the way these criteria should be aggregated in order to get comprehensive and effective ranking of the relevant documents.

We are currently conducting experiments with respect to more classic test collections such as the Text REtrieval Conference .GOV collection where sets of queries with their corresponding relevant documents are available for testing.

An interactive version of this approach is also under study in order to consider factors pertaining to the general context of the user, especially its cognitive status, such as his knowledge level and believes, as well as the situational or contextual environment such as the document 'style' to be retrieved (bibliography papers, overviews, etc.).

References

[Belkin et al., 1995] Belkin, N. J., Kantor, P., Fox, E. A., and Shaw, J. A. (1995). Combining evidence of multiple query representations for information retrieval. *Information Processing and Management*, 31(3):431–448.

[Belkin et al., 1982] Belkin, N. J., Oddy, R. N., and Brooks, H. M. (1982). Ask for information retrieval :part ii. results of a design study. *Journal of Documentation*, 38(3):145–164.

[Brin and Page, 1998] Brin, S. and Page, L. (1998). Anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International World Wide Web Conference*, pages 14–18, Brisbane, Australia.

[Chakrabarti et al., 1998] Chakrabarti, S., Dom, B., Gibson, D., Kumar, S., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1998). Experiments in topic distillation. In *ACM SIGIR Workshop on Hypertext Information Retrieval on the Web*, Melbourne, Australia.

[Cook and Kress, 1985] Cook, W. D. and Kress, M. (1985). Ordinal ranking with intensity of preference. *Management science*, 31(1):26–32.

[Cosijn and Ingwersen, 2000] Cosijn, E. and Ingwersen, P. (2000). Dimensions of relevance. 36(4):533–550.

[Dubois and Prade, 1984] Dubois, D. and Prade, H. (1984). Criteria aggregation and ranking of alternatives in the framework of fuzzy set theory. In Zimmermann, H., Zadeh, L., and Gaines, B., editors, *Fuzzy Sets and Decision Analysis*, pages 209–240. Studies in the Management Sciences, vol. 20, North-Holland, Amsterdam.

[Dwork et al., 2001] Dwork, C., Kumar, S. R., Naor, M., and Sivakumar, D. (2001). Rank aggregation methods for the web. In *World Wide Web*, pages 613–622.

[Froehlich, 1994] Froehlich, T. J. (1994). Relevance reconsidered-towards an agenda for the 21st century: introduction to special topic issue on relevance research. *Journal of the American Society for Information Science*, 45(3):124–134.

[Furnas et al., 1988] Furnas, G. W., Deerwester, S., Dumais, S. T., Landauer, T. K., Harshman, R. A., Streeter, L. A., and Lochbaum, K. E. (1988). Information retrieval using singular value decomposition model of latent semantic structure. pages 465–480. ACM Press.

[Garfield, 1978] Garfield, E. (1978). Citation analysis as a tool in journal evaluation. *Science*, 178:471–479.

[Gordon and Pathak, 1999] Gordon, M. and Pathak, P. (1999). Finding information on the world wide web: The retrieval effectiveness of search engines. *Information Processing and Management*, 35(2):141–180.

[Gudivada et al., 1997] Gudivada, Venkat, N., Ragghavan, Vijay, V., Grosky, William, I., and Kasanagottu, R. (1997). Information retrieval on the world wide web. *IEEE Internet Computing*, 1(5):58–68.

[Harman, 1986] Harman, D. (1986). An experimental study of factors important in document ranking. ACM Press.

[Hawking et al., 2001] Hawking, D., Crasswell, N., Bailay, P., and Griffiths, K. (2001). Measuring search engine quality. *Information Retrieval*, 4(1):33–59.

[Hubbel, 1965] Hubbel, C. H. (1965). An input-output approach to clique identification. *Sociometry*, 28:377–399.

[Katz, 1953] Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18:39–43.

[Kilgarriff, 1996] Kilgarriff, A. (1996). 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.

[Kleinberg, 1999] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environnement. *Journal of the ACM*, 46(5):604–632.

[Luhn, 1958] Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2:159–165.

[Mizzaro, 1997] Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society of Information Science*, 48(9):810–832.

[Pinski and Narin, 1976] Pinski, G. and Narin, F. (1976). Citation influence for journal aggregates of scientific publications : Theory with application to the literature of physics. *Information Processing and Management*, 12:297–312.

[Robertson and Spark Jones, 1976] Robertson, S. E. and Spark Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146.

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

[Roy and Bouyssou, 1993] Roy, B. and Bouyssou, D. (1993). *Aide multicritère à la décision : Méthodes et cas*. Economica.

[Roy and Hugonnard, 1982] Roy, B. and 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, 1968] Salton, G. (1968). *Automatic Information organization and retrieval*. McGraw-Hill.

[Salton, 1989] Salton, G. (1989). *Automatic Text Processing*. Addison-Wesley.

[Salton and Buckley, 1988] Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523.

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

[Savoy and Rasolofo, 2001] Savoy, J. and Rasolofo, Y. (2001). Report on the trec-9 experiment: Link-based retrieval and distributed collections. In *Text REtrieval Conference*, pages 579–588.

[Shannon, 1951] Shannon, C. (1951). Prediction and entropy of printed english. *Bell Systems Technical Journal*, 30:50.

[Singhal and Kaszkiel, 2001] Singhal, A. and Kaszkiel, M. (2001). At&t at trec-9. In *Text REtrieval Conference*, pages 103–105.

[Sparck Jones, 1972] Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

[Van Rijsbergen, 1979] Van Rijsbergen, C. J. (1979). *Information Retrieval*. Dept. of Computer Science, University of Glasgow, London : Butterworths, second edition.

[Voorhees and Harman, 2000] Voorhees, E. and Harman, D., editors (2000). *Proceedings of the 9th Text REtrieval Conference (9)*. National Institute of Standards and Technology. Special Publication 500–249.

[Voorhees and Harman, 2001] Voorhees, E. and Harman, D., editors (2001). *Proceedings of the 10th Text REtrieval Conference (10)*. National Institute of Standards and Technology. Special Publication 500–249.

[Voorhees and Harman, 2002] Voorhees, E. and Harman, D., editors (2002). *Proceedings of the 11th Text REtrieval Conference (11)*. National Institute of Standards and Technology. Special Publication 500–249.