Multidimensional Relevance: Prioritized Aggregation in a Personalized Information Retrieval Setting

Célia da Costa Pereira, Mauro Dragoni, Gabriella Pasi

Multidimensional Relevance: Prioritized Aggregation in a Personalized Information Retrieval Setting

Célia da Costa Pereira[1], Mauro Dragoni[2], and Gabriella Pasi[3]

[1]) *Université de Nice Sophia-Antipolis/CNRS*
UMR-6070, Laboratoire I3S, 06903, Sophia Antipolis, France
celia.pereira@unice.fr
[2]) *Fondazione Bruno Kessler, FBK-irst*
Via Sommarive 18, Povo, I-38123 Trento, Italy
dragoni@fbk.eu
[3]) *Università degli Studi di Milano Bicocca, DISCO*
Viale Sarca, 336, I-20126 Milano (MI), Italy
pasi@disco.unimib.it

Abstract

A new model for aggregating multiple criteria evaluations for relevance assessment is proposed. An Information Retrieval context is considered, where relevance is modeled as a multidimensional property of documents. The usefulness and effectiveness of such a model are demonstrated by means of a case study on personalized Information Retrieval with multi-criteria relevance. The following criteria are considered to estimate document relevance: *aboutness*, *coverage*, *appropriateness*, and *reliability*.

The originality of this approach lies in the aggregation of the considered criteria in a prioritized way, by considering the existence of a prioritization relationship over the criteria. Such a prioritization is modeled by making the weights associated to a criterion dependent upon the satisfaction of the higher-priority criteria. This way, it is possible to take into account the fact that the weight of a less important criterion should be proportional to the satisfaction degree of the more important criterion.

Experimental evaluations are also reported.

Key words: Information Retrieval, Multidimensional Assessment, Relevance Ranking

1. Introduction

The problem of information overload on the Web leads to a demand for effective systems able to locate and retrieve information relevant to user's interests. Systems for content-based access to huge information repositories usually produce a ranked list of documents in response to a user's query; the ranking is based on the assessment of documents' relevance (or probability of relevance) to the user's interests expressed by the query.

Relevance assessment is usually based on the evaluation of multiple criteria (that in this paper we also call relevance dimensions), aimed to capture different aspects or properties of the considered document or document context. All the considered dimensions concur to estimate the utility of the document with respect to the considered user's query. The concept of *page popularity* in search engines is an example of a criterion that is usefully exploited in the process of documents' relevance estimate. As also witnessed by the recent literature, we may assert that relevance can be modeled as a *multidimensional* property of documents [35]. Following this interpretation, the computation of an overall relevance score to be associated with each retrieved document is based on the computation of several scores representing the satisfaction of the considered criteria. As a direct consequence, there is the need to aggregate single values (scores related to single criteria evaluations) into an overall score representing the overall relevance estimate (the so called Retrieval Status Value).

Despite of the fact that this aggregation step plays an important role on the document's final ranking, in the literature this problem has not raised the attention it deserves. In fact, several aggregation schemes may be adopted, giving a distinct role to the single relevance assessments and to their interplay in the definition of the overall relevance assessment.

Among the aggregation operators, traditional non-compensatory operators, such as the min operator and the max operator, used for example in [31], allow to set up a pessimistic (min) or optimistic (max) aggregation scheme. In fact, the min and max operators are essentially dominated by just one criterion value, the worst score and the best score respectively; the remaining criteria are only used to discriminate documents with similar scores. The main inconvenience of such operators is that a large part of scores is ignored or plays a minor role in the aggregation process.

On the contrary, traditional averaging aggregation operators are totally compensatory, i.e., a lack in the satisfaction of a criterion can be compensated by the surplus satisfaction of another one. This property is not very realistic in many real applications in general, and in particular in the Information Retrieval (IR) field. Suppose that a user is looking for articles about Information Retrieval (criterion C_1) which have been possibly written after 1958 (criterion C_2). By using a compensatory aggregation operator, the fact that an article has been written after 1958 can compensate the fact that it is not about IR. This is not realistic because the criterion of being an article on Information Retrieval prevails on the criterion of being an article written after 1958; in other words, the user is interested in an article which is, first of all, about Information Retrieval and, besides, is written after 1958. This example illustrates that there is a need to represent a priority order on the user interests.

The main aim of this paper is to propose and evaluate the role of prioritized aggregation schemes for multidimensional relevance assessment. In particular, two prioritized aggregation operators, originally proposed in [10, 11], are described as well as their formal properties. Moreover, their impact and effectiveness are evaluated by a user-centered approach that has been conducted in a personalized IR setting, as it will be

explained here below. While in [10, 11] we proposed, by two distinct contributions, the formal definitions of the two operators, the aim of this paper is to give a formal unifying asset to the prioritized aggregation approach, as well as to define and present a rigourous experimental setting that makes it possible to show the usefulness, the effectiveness and the potential of the proposed aggregation schemes. In fact in [10, 11] only a few preliminary experiments were presented, based on a different approach with respect to the one presented in this paper.

To the aim of illustrating and evaluating the prioritized aggregation operators we consider a personalized IR setting. Personalized approaches to IR make use of a formal representation of the user's interests (the so called users profiles) to assess documents' relevance in relation to a user's query. Search personalization, which is particularly effective if the user often formulates queries inherent to the interests represented in his/her profile, has been introduced to overcome the "one size fits all" approach of traditional IRSs, by which a same query formulated by distinct users produces the same results, independently on the users' context and search intentions. Recently, increasing research efforts have been made to make Information Retrieval technologies exploit as much contextual factors as possible in order to tailor search results to a particular user [18, 36, 6].

An interesting aspect in considering a personalized IR setting to evaluate the prioritized aggregation (and the main reason that motived this choice) is that the priority over the considered relevance dimensions may be user dependent; as we will show in the paper, the consideration of different priority orders over the four considered relevance dimensions can identify different types of users', with distinct search intents. The main impact in making the prioritized aggregation scheme user-dependent is that for a same query and a same user different document rankings can be obtained, depending on the selected priorities over the considered relevance dimensions (each related to a distinct user type, or category). In the proposed personalized approach, we consider four relevance dimensions: *aboutness* (or topicality), *coverage*, *appropriateness* and *reliability*. It is quite important to outline that the aggregation model proposed in this paper is general, and as such it may be applied to *any* set of relevance dimensions.

An important consequence of the choice of a personalized IR setting as an instance of the proposed prioritized aggregation approach concerns the evaluation of its effectiveness. Indeed, there is not a test collection suited to evaluate with a laboratory based strategy an IRS that is user-dependent. For this reason, we propose in this paper a user-centered evaluation strategy, based on an analysis of how document rankings are modified after applying the two operators on different priorities over the four considered dimensions.

The paper is organized as follows. Section 2 presents an overview of the literature related to multidimensional relevance assessment. Section 3 presents the prioritized aggregation methods, their formal representation, and their properties. Section 4 presents an application of the proposed approach to a Personalized Information Retrieval setting, as well as the formal definitions of the considered IR context with the four relevance

dimensions we suggest to consider. Section 5 presents the experimental setting that we propose to adopt in such a context, as well as the evaluations that we have performed to demonstrate the effectiveness of the proposed approach. Finally, Section 6 concludes the paper.

2. From Topical to Multidimensional Relevance: Related Work

In the first traditional approaches to IR, relevance was modeled as "topicality", and its numeric assessment was based on the matching function related to the adopted IR model (*boolean model*, *vector space model*, *probabilistic models* or *fuzzy models*). However, relevance is, in its very nature, the result of several components or dimensions.

In the Information Retrieval literature there are many terms considered as synonymous but having sometimes different meanings. Hjørland [15] pointed out some of those which for him are the most important ones: *subject*, *aboutness*, *topicality*, *theme*, *domain*, *field*, *content*, and *information*. Some use the term *aboutness* as a kind of *topicality*. For example, *topicality relevance* or *assessment relevance* [30] means that a document is about a query if its content or its surrogate content matches with the user query content; *topic-appropriateness* [7] corresponds to Cooper's definition of topicality; *intellectual topicality* [4] refers to how an information object corresponds to the topical area required by the information need as perceived.

Others use the term *aboutness* as *topical matching* and define *aboutness* as a *topicality relation* [5]. In this case, a document topically related to a query, i.e., such that the topic of the query matches the document's topic, is said topically relevant.

Maron [19] considered aboutness by relating it to a probability of satisfaction. He characterized three types of aboutness: S-about which is a relationship between a document and the resulting experience of the user. O-about, which is a relationship between a document and a set of index terms. A document D is about a term T if user X employs T to search for D. Finally, R-about, which purports to be a generalization of O-about to a specific user community. Let I be an index term and D be a document, then "D is R-about I" is the ratio between the number of users satisfied with D when using I and the number of users satisfied by D.

An interesting experimental study about *aboutness* is that made by Fugita [13]. The basic hypothesis behind his work is that he considered that the aboutness of a subject topic consists of a "foreground" part and a "background" part and terms belong to either one of them. The motivation beyond this distinction is the metaphor of "aboutness" of visual information items: "People are clearly distinguishing foreground images from background ones when talking about *aboutness* of for example picture images" he pointed out.

The above mentioned works consider aboutness as the unique dimension of relevance. The multidimensional nature of relevance is not considered.

There are two schools of thought with two different ideas on how *aboutness* (*topical relevance*) influences *relevance* [28]. For the first, *aboutness* is considered as the basic

part of relevance [33], that is, all the other factors which can influence the relevance are topical-dependent. People from the second school of thought [3, 9, 27] think that not only aboutness is the basic component of relevance and consider that there are other factors influencing relevance beyond aboutness, and these factors do not depend on topical relevance but they are strongly based on a subjective user's perception, related to the user interests – two documents can have the same topical relevance (aboutness, subject) without having the same relevance. We adhere to the latter school of thought which considers that topical relevance is just one of the dimensions of relevance.

This idea is not completely new. Cooper [8] was a pioneer in this direction. He underlined several additional properties of documents that could be considered to assess their RSVs with respect to the users' needs. He defined relevance as *topical relevance with utility* and pointed out document properties such as novelty, informativeness, credibility, importance or weighting of components in request, clarity, involvement with the system, possible negative and positive factors, and possible esthetic factors, that can be considered when assessing relevance.

An influencial work reviewing the concept of relevance has been proposed by Saracevic in [26]. In that work, Saracevic pointed out the necessity of considering relevance also from the user's point of view, thus underlining the fact that relevance has to be considered as a dynamic and multidimensional concept.

Another pioneering and informal work defining relevance as a multidimensional concept is the one by Schamber and Eisenberg [29], in which, after a critical review of four approaches to the problem of defining relevance, namely, (i) the system oriented approach, (ii) the user-oriented approach, (iii) the multidimensional approach, and (iv) the cognitive approach; they proposed the following definition: "Relevance is a multidimensional concept based on the human judgment process; it is dependent on both internal (cognitive) and external (situational) factors; and it is intersubjective but nevertheless systematic and measurable."

Another work which discusses the advantages of defining relevance as a multidimensional concept is that by Barry [2]. She underlined the need of considering other factors beyond topicality when computing document relevance. Her research is based upon the following assumption: " Each individual does not possess a unique set of criteria by which information is judged. Motivated users evaluating information within the context of a current information need situation will base their evaluations on factors beyond the topical appropriateness of documents."

Mizzaro followed the multidimensional line of Saracevic [27]. He proposed in [21], a relevance model in which relevance is represented as a four-dimensional relationship between an information resource (surrogate, document, and information) and a representation of the user's problem (query, request, real information need and perceived information need). A further judgment is made according to the: topic, task, or context, at a particular point in time.

The relationship between time and relevance pointed out by Mizzaro has also been considered by Borlund [3] as follows: dynamic relevance refers to how the same user's

perception of relevance may change over session time. Therefore, by considering what happens during the interaction between the user and the system when the user is looking for information allows us to also consider the changes in the user's mind with respect to his previous knowledge. Harter [14] proposes that *relevance judgment*, i.e., an assignment of a value to relevance by a judge, is a psychological process in which the retrieved information objects serve as stimuli that result in cognitive changes through the time in the user's mental state.

Cosijn and Ingwersen [9] confirmed the influence of time on user relevance judgment and pointed out to both the fact that relevance can be seen as having a certain number of attributes as claimed by Saracevic, and that relevance can itself be manifested in different ways. They examined one of the possible ways to model relevance manifestations within a system of relevance attributes and shown that motivational relevance should not be viewed as part of a linear, objective-subjective scale of relevances, but rather as an attribute of relevance. Similarly, they show that the manifestation of affective relevance should not be viewed as a discrete category of relevance manifestation, but rather as an influencing factor on the other subjective relevance types.

Based on a cognitive approach, Xu and Chen [37] carried out a study in which they focused on the criteria users employ in making relevance judgment beyond topicality. The purposes of their study were (1) to identify a set of core relevance criteria using a theory-driven approach and (2) to test the validity of these factors with a rigorous psychometric approach. The result consisted in a five-factor model of relevance composed by topicality, novelty, reliability, understandability, and scope.

The work proposed by Farah and Vanderpooten [12] is aimed at dealing with *imprecision* underlying criteria design resulting from the fact that there are many formulations of the same criterion. The interpretation they give to the scores of each criterion does not consider *slight* differences which are often not meaningful. They propose a multi-criteria framework using an aggregation mechanism based on decision rules identifying positive and negative reasons for judging whether a document should get a better ranking than another.

More recently, Taylor and colleagues [35] presented a statistical study which extends the research on the relationship between multidimensional user relevance assignments and the stage in the process of completing a task. The obtained results suggest that users consistently identify relevance criteria beyond topical relevance.

While the works reported in this section have given a contribution to clarify the important notion of relevance, no sufficient attention has been devoted in the literature to the important problem of aggregation, naturally raised by a multidimensional relevance modeling. The simplest way to obtain a unique ranking score based on the computation of several relevance scores associated with a given document for a specific query is to aggregate the relevance scores by means of an appropriate mathematical operator (usually called an aggregation operator). The usual choice in the IR context has been to apply the average operator. However, as also outlined in the Introduction, the existence of several distinct aggregation schemes may be usefully exploited so as to make

the relevance criteria interplay in distinct ways in the process of relevance assessment. To clarify this assertion let us think that the aggregation by a min or a max operator for example, implies to consider the criteria totally independent and interchangeable, in the sense that no compensation is applied to their satisfaction degrees. If one wants to define an interdependence over the criteria (the satisfaction of a criterion may influence the satisfaction of a less prioritized criterion), a different aggregation scheme has to be applied, as we will see in the next section. It is important to outline that the approach proposed in this paper has been inspired by the scientific context of Decision Theory, in particular by Multi-Criteria Decision Making (MCDM), where aggregation plays an important role. More precisely, we have been inspired by the fuzzy extension of MCDM proposed by Yager in [38, 39] in which the author provided some models that allow for the formalization of prioritized MCDM problems using both the Bellman–Zadeh paradigm for MCDM multi-criteria decision making and the OWA operator method.

The concept of *dimension* we use is somehow different from that used by Mizzaro and Saracevic. They defined several kinds of relevance and call them *dimensions of relevance* while we define relevance as a *concept of concepts*, i.e., as a concept defined thanks to other dimensions. The RSV of a document is then the result of a particular aggregation of the considered dimensions as explained in the following section.

3. Prioritized Multi-criteria Aggregation

In a multi-criteria decision making setting, the problem of prioritized aggregation is typical when one wants to model a relationship between the considered criteria. In such a case, the lack of satisfaction by an higher priority criterion cannot be compensated with the satisfaction by a lower priority criterion. To illustrate this assertion we present here below a simple decision making problem in a real life situation.

Introductory Example. Let us suppose that John is looking for a bicycle for his little son. His choice is based on the consideration of two selection criteria. The first one is "safety" and the second one is "inexpensiveness". For John, it is more important that the bicycle is safe than inexpensive. Therefore, he would like to buy a bicycle which is first of all safe and then, if possible, inexpensive. We can consider two situations.

1. John can buy a bicycle that is safe but expensive if he does not find any safe bicycle that is inexpensive (so the under-satisfaction of the inexpensiveness criterion may be in some way compensated by the satisfaction of the safety criterion).
2. John has little money. He still considers that safety is more important than cost, however, he can not afford big expenses. In this case, John would look for a bicycle that is first of all, safe but also ("and") inexpensive (relative to what he can afford). So in this case the under-satisfaction of the inexpensiveness criterion cannot be compensated by the satisfaction of the safety criterion.

We can notice that in both cases we need a prioritized aggregation operator. However, the suitable operator for the first case is different from the suitable operator for

the second case. The difference with respect to the first case is that in the second case, a bicycle that is safe but not inexpensive enough would not be considered by John. Instead, the similarity with the first case is that a bicycle that is inexpensive but not safe would not be considered because safety is the most important criterion for John. An illustrative example that shows how the two RSV values change depending on case 1 and case 2 will be given at the end of subsection 3.3, after the presentation of the two aggregation models.

In this section, an approach to the problem of aggregating distinct documents' relevance assessments in such situations is defined. As previously outlined, we consider this problem as a multi-criteria decision making problem. By making a shift from real life examples to the Information Retrieval context, the considered criteria are the various relevance dimensions, and the possible alternatives are constituted by the documents.

The two prioritized aggregation operators that are described in this section (the *"scoring" operator* and the *"and" operator*), have been first proposed in [10, 11]. As it will be seen, the proposed operators apply a weighted aggregation, where the weights associated with the considered criteria (relevance dimensions) are computed on the basis of the specified priority order. This means that the weight associated with a criterion depends on the satisfaction of the higher-priority criteria.

3.1. Problem Representation

The presented multi-criteria decision making approaches have the following components:

- the set C of the n considered criteria: $C = \{C_1, \ldots, C_n\}$, for the sake of notational clarity C_i will also denote the function evaluating the ith criterion;
- the collection of documents D;
- an aggregation function F to calculate for each document $d \in D$ a score $F(C_1(d), \ldots, C_n(d)) = RSV(d)$ on the basis of the evaluation scores of the considered criteria.

We assume that each user may express his/hes *total* preference (or priority) order, denoted by the symbol $\succ$, on the criteria C_i. Such an order is used for exploiting the relevance assessment of a document, i.e. to compute its RSV with respect to the considered query. This means that by considering the same set of criteria, the relevance assessment of a given document to a query may produce different scores (different RSVs) for different users (and also for a same user in different search sessions with distinct search intents). This is due to the fact that the preference order over the criteria is user-dependent, and, as such, by applying the method that we will explain here below, it induces a different importance weight associated with each criterion, which is taken into account in the RSV computation.

Let us assume that the $C_1, C_2, ..., C_n$ considered criteria have been ordered by the user in the sequence $C_1', C_2', ..., C_n'$, where C_1' is the most important criterion and C_n' the

less important one for the user. The method we define to compute the numeric weight associated with each criterion C'_j is both document and user-dependent. It depends in fact first on the preference order expressed by the user over the considered criteria, and also on both the satisfaction degree of the document with respect to criterion C'_{j-1} (in the preference order defined by the user) and the weight associated with criterion C'_{j-1} (except the case of criterion C'_1, to which the importance weight 1 is assigned)[1].

We denote by $\lambda_i \in [0,1]$ the importance weight associated with criterion C_i and computed for a given document d. Notice that different users can have a different preference order over the criteria and, therefore, it is possible to obtain different importance weights for the same document for different users.

To simplify the adopted notation, in the following we assume that after the user preference reordering of the n considered criteria we denote by C_1 the most preferred criterion, by C_n the least preferred criterion (i.e., the last in the user preference list), and we assume that C_i is preferred to $C_j (C_i \succ C_j)$ if and only if $i < j$.

The above intuition may be formalized as follows:

- for each document d, the weight of the most important criterion C_1 is set to 1, i.e., by definition we have: $\forall\, d\ \lambda_1 = 1$;
- the weights of the other criteria C_i, $i \in [2,n]$, are calculated as follows:

$$\lambda_i = \lambda_{i-1} \cdot C_{i-1}(d), \tag{1}$$

  where $C_{i-1}(d)$ is the degree of satisfaction of criterion C_{i-1} by document d, and λ_{i-1} is the importance weight of criterion C_{i-1}.

3.2. The Prioritized Scoring Model

In this section, we present the formalization and the properties of the "scoring" prioritized aggregation operator, F_s. This operator allows to calculate the overall score from several criteria evaluations, where the weight of each criterion depends both on the weights and on the satisfaction degrees of the most important criteria — the higher the satisfaction degree of a more important criterion, the more the satisfaction degree of a less important criterion influences the overall score.

Operator F_s is defined as follow: $F_s : [0,1]^n \to [0,n]$ and it is such that, for a given document d,

$$F_s(C_1(d), \ldots, C_n(d)) = \sum_{i=1}^{n} \lambda_i \cdot C_i(d). \tag{2}$$

The RSV_s of the document d is then given by:

$$RSV_s(d) = F_s(C_1(d), \ldots, C_n(d)), \tag{3}$$

[1]If there are more than one criterion with the same priority order, the average weight and the average satisfaction degree are considered.

where C_i's represent the considered relevance dimensions.

Let us come back to the introductory example. $C_1 = Safety$ and $C_2 = inexpensiveness$, with $C_1 \succ C_2$. A bicycle with a "safety" degree of 0.6 and an "inexpensiveness" degree of 0 would have an RSV_s of 0.6. Indeed, we have $\lambda_1 = 1$, $\lambda_2 = 0.6$ and then the RSV_s is $(1 \cdot 0.6) + (0.6 \cdot 0) = 0.6$. Instead, if the "safety" degree is 0 and the "inexpensiveness" degree is 1, we would have an RSV_s of $(1 \cdot 0) + (0 \cdot 1) = 0$.

Let us consider two further simple examples with four criteria (i.e., $n = 4$): C_1, C_2, C_3 and C_4, where C_1 is the more important criterion and C_4 is the less important criterion. If document d_i is such that $C_1(d_i) = 0.6$, $C_2(d_i) = 0.8$, $C_3(d_i) = 0.9$, and $C_4(d_i) = 1$, we have:

- $\lambda_1 = 1$;
- $\lambda_2 = \lambda_1 \cdot C_1(d_i) = 0.6$;
- $\lambda_3 = \lambda_2 \cdot C_2(d_i) = 0.48$;
- $\lambda_4 = \lambda_3 \cdot C_3(d_i) = 0.432$;

and the RSV_s is then:

$$RSV_s(d_i) = (1 \cdot 0.6) + (0.6 \cdot 0.8) + (0.48 \cdot 0.9) + (0.432 \cdot 1) = 1.944.$$

Instead, if we exchange the satisfaction degrees of C_2 and C_3 such that $C_2(d_i) = 0.9$ and $C_3(d_i) = 0.8$. We would have:

- $\lambda_1 = 1$;
- $\lambda_2 = \lambda_1 \cdot C_1(d_i) = 0.6$;
- $\lambda_3 = \lambda_2 \cdot C_2(d_i) = 0.54$;
- $\lambda_4 = \lambda_3 \cdot C_3(d_i) = 0.432$;

and the RSV_s is then:

$$RSV_s(d_i) = (1 \cdot 0.6) + (0.6 \cdot 0.9) + (0.54 \cdot 0.8) + (0.432 \cdot 1) = 2.004.$$

This result is justified by the fact that the second more important criterion is better satisfied by the document than the second more important criterion is in the first example.

Properties of the Prioritized "Scoring" Operator

Here, we present some mathematical properties of the prioritized "scoring" operator.

Continuity. The proposed aggregation is a polynomial; therefore, the property of continuity holds.

Boundary Conditions. The "scoring" operator satisfies:

- if $\forall\, i\, C_i(d) = 0$ then $\sum_{i=1}^{n} \lambda_i \cdot C_i(d) = 0$;
- if $\forall\, i\, C_i(d) = 1$ then $\sum_{i=1}^{n} \lambda_i \cdot C_i(d) = n$;
- $\forall\, i \in \{1, \ldots, n\}$ we have $0 \leq \sum_{i=1}^{n} \lambda_i \cdot C_i(d) \leq n$

Monotonicity (non decreasing). The "scoring" operator is *monotonous*. If the satisfaction degree of one of the criteria increases, then the final aggregation increases (or at least it does not decrease, remaining equal). Indeed, let d be a document and let us consider the two sets of criteria $\{(C_1(d),\ C_2(d),\ \ldots,\ C_n(d)\}$ and $\{(C'_1(d),\ C'_2(d),\ \ldots,\ C'_n(d)\}$ so that:

- $C'_i(d) = C_i(d),\ \forall i \neq j$,
- $C'_j(d) \geq C_j(d)$.

We have than

$$\sum_{i=1}^{n} \lambda'_i \cdot C'_i(d) \geq \sum_{i=1}^{n} \lambda_i \cdot C_i(d).$$

Indeed, $\sum_{i=1}^{n} \lambda'_i \cdot C'_i(d) - \sum_{i=1}^{n} \lambda_i \cdot C_i(d) = \sum_{k=j}^{n} \lambda'_k \cdot C'_k(d) - \lambda_k \cdot C_k(d)$. This difference is positive because:

- $\lambda'_j = C'_1(d) \cdot C'_2(d) \ldots C'_{j-1} = \lambda_j$ (see Equation 1), and $C'_j(d) \geq C_j(d)$ and then $\lambda'_j \cdot C'_j(d) \geq \lambda_j \cdot C_j(d)$,
- $\forall k > j \quad \lambda'_k \geq \lambda_k$ (by construction, see Equation 1) and then $\lambda'_k \cdot C'_k(d) \geq \lambda_k \cdot C_k(d)$.

Absorbing Element. The "scoring" operator has an *absorbing element*: $C_1(d) = 0$. If the most important criterion is not satisfied at all, we have $\sum_{i=1}^{n} \lambda_i \cdot C_i(d) = 0$ independently on the satisfaction degrees of the other criteria. Indeed,

$$\sum_{i=1}^{n} \lambda_i \cdot C_i(d) = \lambda_1 \cdot C_1(d) + \sum_{i=2}^{n} \lambda_i \cdot C_i(d)$$

with $\lambda_i = C_1(d) \cdot C_2(d) \ldots C_{i-1}(d) = 0, \quad \forall\, i > 2$.

This property can be extended to other levels of importance. Indeed, if the kth most important criterion is not satisfied at all, i.e. $C_k(d) = 0$, then we have

$$\sum_{i=1}^{n} \lambda_i \cdot C_i(d) = \sum_{i=1}^{k-1} \lambda_i \cdot C_i(d).$$

3.3. The Prioritized "and" Operator

In this section the prioritized "and" operator is introduced [11]. This operator allows to model a situation where the overall satisfaction degree is strongly dependent on the degree of satisfaction of the least satisfied criterion. The peculiarity of such an operator, which also distinguishes it from the traditional "and" operator, is that the extent to which the least satisfied criterion is considered depends on its importance for the user. If it is not important at all, its satisfaction degree should not be considered in the aggregation process, while if it is the most important criterion for the user, only its satisfaction degree is considered. This way, if we consider a document d, for which the least satisfied criterion C_k is also the least important one, the overall satisfaction degree will be greater than $C_k(d)$; it will not be C_k as it would be the case with the traditional "and" operator — the less important is the criterion, the lower its chances to represent the overall satisfaction degree.

The aggregation operator F_a is defined as follows. $F_a : [0,1]^n \rightarrow [0,1]$ is such that, for all documents d,

$$F_a(C_1(d), \ldots, C_n(d)) = \min_{i=1,n}(\{C_i(d)\}^{\lambda_i}). \tag{4}$$

The RSV_a of the document d is then given by: $RSV_a(d) = F_a(C_1(d), \ldots, C_n(d))$.

$$RSV_a(d) = F_a(C_1(d), \ldots, C_n(d)), \tag{5}$$

where the C_i's represent the satisfaction degrees of the considered relevance dimensions.

Let us come back again to the introductory example. $C_1 = Safety$ and $C_2 = inexpensiveness$, with $C_1 \succ C_2$. Here, a bicycle with a "safety" degree of 1 and a "inexpensiveness" degree of 0 would have an RSV_a of 0. Indeed, we have $\lambda_1 = 1$, $\lambda_2 = 1$ and then the RSV_a is $min(1^1, 0^1) = 0$. Instead, if the "safety" degree is 0.9 and the "inexpensiveness" degree is 0.8, we would have an RSV_a of $min(0.9^1, 0.8^{0.9}) = 0.818$.

Let us consider two further examples with four criteria C_i with $i \in \{1, \ldots, 4\}$, where C_1 is the most important criterion and C_4 is the less important criterion. If document d_i is such that $C_1(d_i) = 0.9$, $C_2(d_i) = 0.7$, $C_3(d_i) = 0.9$, and $C_4(d_i) = 0.6$, we have:

- $\lambda_1 = 1$;
- $\lambda_2 = \lambda_1 \cdot C_1(d_i) = 0.9$;
- $\lambda_3 = \lambda_2 \cdot C_2(d_i) = 0.63$;
- $\lambda_4 = \lambda_3 \cdot C_3(d_i) = 0.567$;

and the overall score would then be:

$$\begin{aligned} F_a(C_1(d_i), \ldots, C_4(d_i)) &= \min(0.9^1, 0.7^{0.9}, 0.9^{0.63}, 0.6^{0.56}) \\ &= \min(0.9, 0.72, 0.93, 0.74) = 0.72. \end{aligned}$$

Instead, if

$C_2(d_i) = 0.9$, $C_3(d_i) = 0.7$, we would have:

- $\lambda_1 = 1$;
- $\lambda_2 = \lambda_1 \cdot C_1(d_i) = 0.9$;
- $\lambda_3 = \lambda_2 \cdot C_2(d_i) = 0.81$;
- $\lambda_4 = \lambda_3 \cdot C_3(d_i) = 0.567$;

and the overall score

$$\begin{aligned} F_a(C_1(d_i), \ldots, C_4(d_i)) &= \min(0.9^1, 0.9^{0.9}, 0.7^{0.81}, 0.6^{0.567}) \\ &= \min(0.9, 0.909, 0.749, 0.748) \\ &= 0.748. \end{aligned}$$

This result is justified by the fact that the second more important criterion is better satisfied by the document in the latter case than the second more important criterion in the former case.

Properties of the Prioritized "and" Aggregation Operator

Here, we present some mathematical properties of the prioritized "and" operator.

Continuity. The proposed aggregation is a polynomial, as it can be written as

$$F_a(C_1(d), \ldots, C_n(d)) = min(C_1(d), \min_{i=2,n}(\{C_i(d)\}^{\lambda_i})),$$

which is continuous, because, $\forall\, x, y \in \mathbb{R}$, $\min(x, y) = \frac{x+y}{2} - \frac{|x-y|}{2}$. Therefore, the property of continuity holds for F_a.

Boundary Conditions. The proposed "and" operator satisfies:

- if $\forall\, i\, C_i(d) = 0$, then $\min_{i=1}^{n} \lambda_i \cdot C_i(d) = 0$;
- if $\forall\, i\, C_i(d) = 1$, then $\min_{i=1}^{n} \lambda_i \cdot C_i(d) = 1$;
- $\forall\, i \in \{1, \ldots, n\}$, we have $0 \leq \min_{i=1}^{n} \lambda_i \cdot C_i(d) \leq 1$.

Monotonicity. The proposed "and" operator is not *monotonous*; indeed, if the satisfaction degree of one of the criteria increases, the overall final aggregation value can decrease.

For example, if we have $C_1(d) = 0.7, C_2(d) = 0.1, C_3(d) = 0.3$, from Equation 1 we obtain that $\min(0.7^1, 0.1^{0.7}, 0.3^{0.07}) = 0.1995$. If we replace $C_1(d)$ by a greater value $C_1'(d) = 0.9$, we obtain that $\min(0.9^1, 0.1^{0.9}, 0.3^{0.09}) = 0.125$, which is less than the previous value.

Absorbing Element. The proposed operator has an *absorbing element*, $C_i(d) = 0$. Indeed, if there is a criterion C_i such that $C_i(d) = 0$, we have

$$F_a(C_1(d), \ldots, C_i(d), \ldots, C_n(d)) = 0.$$

Another consequence of the *absorbing* property is that if a document d slightly satisfies the most important criterion C_1, and satisfies, to some extent, all the other criteria, the score obtained for all the possible aggregations in which C_1 is the most important criterion goes to $C_1(d)$, independently of the satisfaction degrees of the other criteria. This is due to Equations 1 and 4 and to the fact that $\lim_{k\to 0} x^k = 1$.

Neutral Element. The proposed operator has a *neutral element*, $C_i(d) = 1$. Indeed, if there is a criterion C_i such that $C_i(d) = 1$, we have

$$F_a(\ldots, C_i(d), \ldots) = F(\ldots, C_{i-1}(d), C_{i+1}(d) \ldots).$$

Idempotence. The proposed operator is *idempotent*. Indeed, if all the criteria have the same satisfaction degree, $x \in [0, 1]$, we have

$$F_a(C_1(d), \ldots, C_i(d), \ldots C_n(d)) = F_a(x, \ldots, x) = x.$$

If $k \in [0, 1]$, we have $x \leq x^k \leq 1$ for all $x \in [0, 1]$.

We reconsider here the Introductory Example presented at the start of Section 3 to show how the application of the two aggregation oprators produces distinct and intuitive results.

Let us suppose that $C_1(d) = 0.9$ and $C_2(d) = 0.2$, with C_1 corresponding to the "safety" criterion and C_2 to the "inexpensiveness" criterion. The different scores associated with John 1 and John 2 are computed as follows. $\lambda_1 = 1$; $\lambda_2 = 0.9$:

- John 1: $RSV_s = \lambda_1 C_1(d) + \lambda_2 C_2(d) = 0.9 + 0.18 = 1.08$;
- John 2: $RSV_a = \min(0.9^1, 0.18^{0.9}) = 0.213$.

The value of RSV_s is better (greater than the midpoint of the definition interval—indeed, $RSV_s \in [0, 2]$) than the score of RSV_a which, instead, belongs to the $[0, 1]$ interval. In the first case, there is an evident compensatory effect, which is not the case (not so evident anyway) for John 2.

4. A Case Study: Prioritized Aggregation Operators in a Personalized IR Setting

As outlined in Section 2, several relevance dimensions have been proposed in the literature. In this paper, in order to illustrate how the prioritized operators can be applied to aggregate several relevance dimensions, we have considered an IR context that

can be personalized according to the priorities that distinct users, or the same user can give to the following four relevance dimensions: aboutness, coverage, appropriateness, reliability (appropriateness was originally proposed in [10]). In particular, the coverage, and the appropriateness dimensions are explicitly related to the formal representation of a user profile represented as a bag of words. Moreover, personalization here is also related to the search task of the users: if, e.g., the user wants to privilege coverage with respect to reliability and to the other two dimensions, a coherent priority order may be specified. This way, the evaluation of a same query for the same user (or for different users) can produce distinct rankings of the retrieved documents.

As we have pointed out in Section 1, the role of *aboutness* (*topicality*) in Information Retrieval has been largely studied — it is the most studied among the existing dimensions of relevance and it is also considered as being the basic one. *Coverage*, *appropriateness* and *reliability* have been studied more recently, and express relevance dimensions which are related to the personal user context.

In this section we present a description and a formal definition for each of these relevance dimensions.

We assume that the user interests are represented by a user profile, formally expressed as a vector of terms $c = [w_{1c}, \ldots, w_{|T|c}]$. In the literature, several research works have addressed the problem of users' profiles definition, both based on explicit or implicit user indications. The problem of how to define such a profile is out of the scope of this paper. See [36] for further details.

4.1. The Four Considered Criteria

4.1.1. Aboutness

In this paper, we use the term *aboutness* as a synonym of *topical relevance*. The aboutness is generally measured based on an Information Retrieval Model. One of the most used models is the Vector Space Model where both queries and documents are represented by vectors of terms; in our experiments we adopt a system based on this model.

Formally, we have $d = [w_{1d}, \ldots, w_{|T|d}]$ and $q = [w_{1q}, \ldots, w_{|T|q}]$ representing document d and query q respectively, with $|T|$ representing the size of the term vocabulary used. The measure of aboutness is then calculated by the standard cosine-similarity [24]:

$$\text{Aboutness}(q, d) = \frac{\sum_{i=1}^{|T|} w_{iq} \cdot w_{id}}{\sqrt{\sum_{i=1}^{|T|} w_{iq}^2 \cdot \sum_{i=1}^{|T|} w_{id}^2}}. \tag{6}$$

4.1.2. Coverage

The *coverage* criterion is assessed on the document representation and on the user profile representation. This criterion has been recently introduced and formalized in [22], and it measures how strongly the user interests are included in a document. The coverage between the user interest and a considered document may be defined as a

fuzzy inclusion [20] based on the cardinalities of the fuzzy subsets representing the user interests and the considered document respectively:

$$\text{Coverage}(c, d) = \frac{\sum_{i=1}^{|T|} \min(w_{ic}, w_{id})}{\sum_{i=1}^{|T|} w_{ic}}. \tag{7}$$

This function produces the maximum value 1 when all the indexes of c are also indexes of d. It produces the value zero in case of no common index terms; the value of the function increases with the increase of the number of common indexes. To be fully included in d, the index term weights of c must be smaller than the index term weights of d.

4.1.3. Appropriateness

This dimension, originally proposed in [10], allows to measure how appropriate or how seemly a document is with respect to the user interests. *Appropriateness*, which can be viewed as a measure of similarity between the vector of document d and the vector of interest c, can be formally defined as follows:

$$\text{Appropriateness}(d, c) = 1 - \frac{\sum_{i=1}^{|T|} |w_{ic} - w_{id}|}{\sum_{i=1}^{|T|} w_{id}}, \tag{8}$$

where, like in the coverage definition, both d and c are vectors of terms representing, respectively, the document and the user interest, and $|T|$ is the size of the term considered vocabulary.

Let us consider a simple example. Let us consider the following document vector $d = [1, 1, 1, 1]$, where the four dimensions correspond to the terms *geography*, *Europe*, *economy* and *politics*; let us suppose that the user is just interested in politics with $c = [0, 0, 0, 1]$. In this case, while the coverage score is the maximum one, i.e., 1, the appropriateness score is 0.25. Indeed, we may note that document d contains information about other topics than just politics, and the appropriateness criterion takes this fact into account. Consider instead another document d' which is only related to politics, and which is represented by $d' = [0, 0, 0, 1]$. The coverage of c in d' is the same as in d. However, the appropriateness of d' with respect to c is 1, because d' is only concerned with politics, and it then represents more faithfully the user interest. This demonstrates that appropriateness is more specific than coverage.

4.1.4. Reliability

The two most common ways to determine the reliability of a document are through (i) *sets of rules* allowing to classify a document as reliable by considering some document's properties; and (ii) *trust* that the user has for the document source; this property could be assessed based on the history of interactions made or past observations [1].

The *trust* a user has about a source can then be measured only in the cases in which the user has seen other documents from that source in the past. The result of such experience allows the user to judge a new incoming document as trustworthy or

not. We assume that the user can insert a list of her/his favorite sources with their respective degrees of preference. As a consequence, the reliability degree of a document d for user i, noted $\mathcal{T}_i(d)$, may be evaluated on the basis of the degree to which the user trusts the source from which the document d comes, i.e.,

$$\mathcal{T}_i(d) = T_i(s(d)), \tag{9}$$

where $s(d)$ represents the source of document d, and $T_i(s(d))$ represents the trust degree of the source for user i. In this paper we do not consider the properties of the documents for determining their reliability.

5. Evaluation of the proposed approach

In this section, the effectiveness of the proposed prioritized aggregation is evaluated in the personalized IR setting described in section 4.

The traditional approach to evaluate the effectiveness of Information Retrieval algorithms is based on the Cranfield paradigm that allows the so called laboratory-based evaluations, which make use of pre-defined test collections composed by a document collection, a set of queries (topics), and a set of relevance judgments. However, as it has been largely outlined in the literature, when interactive and/or personalized IR approaches are considered, the user-centered evaluation approach is usually applied [17] [32] [23]. The unavailability of a standard test collection holds also for the application scenario of the prioritized aggregation proposed in this paper. In fact, when applying the prioritized aggregation strategy, a same document evaluated with respect to a same query can produce distinct assessment scores, depending on the adopted prioritized scheme, which is user-dependent.

We propose then a user-centered evaluation, as it will be explained in Section 5.1.

The aim of the experiments presented in this section is twofold: (i) to verify that when a user performs queries related to his/her interests, by applying a prioritized aggregation operator (prioritized "scoring" or prioritized "and" operators), the system produces an improved ranking with respect to the one produced by its correspondent non-prioritized aggregation operator (the average operator), and (ii) to verify that when a user performs queries that are not related to his/her interests, by applying a prioritized aggregation operator, the quality of the produced rank does not decrease with respect to the one of the rank produced by the average operator.

As outlined in Section 4, the priority order of the relevance criteria depends on the user's search intent. To the aim of performing a meaningful evaluation, based on the semantics of the relevance dimensions defined in section 4.1, we identify three users' categories corresponding to distinct search intents, which induce three different priority orders over the considered relevance dimensions, as described here below. The identified categories constitute the evaluation scenarios we have considered to the aim of the evaluation.

As the first evaluation scenario we consider the case when a user formulates a query focused only on his/her interests; in this case, we assume that the user aims at locating documents that are first of all related to his/her interests, while at the same time requiring that the searched documents do not focus on additional topics other than those expressed by the query. Let us suppose that a user is looking for documents about "gold" and he/she is interested in chemistry and not in economics. This means that the user is looking for documents about "gold" as a chemical element, not about, e.g., "gold" as a store of value. According to this search scenario, we identify a first user category that we call "coverage seeker". With this user category we associate the following priority order over the four considered relevance dimensions:

CA_pAR: $coverage \succ appropriateness \succ aboutness \succ reliability$.

It is very important to outline that slightly different priority orders could be associated with this user category, as well as with the other two users' categories specified in the following.

As a second situation, we consider the case when the user's intent is to find documents which perfectly fit his/her interests, as specified in the user's profile; we name this second user's category "appropriateness seeker", and we associate with it the following priority order over the four considered relevance dimensions:

A_pACR: $appropriateness \succ aboutness \succ coverage \succ reliability$;

The third user category we introduce refers to users who give a priority to the reliability of the information source of the retrieved documents; we call the users belonging to this category *cautious*; these users give a greater priority to the reliability criterion than to the other criteria.
With this user category we associate the following priority order over the four considered relevance dimensions:

RAA_pC: $reliability \succ aboutness \succ appropriateness \succ coverage$;

In the next sub-sections, after presenting the comparative evaluations of the rankings produced by the two priority based aggregation operators with the ranking produced by the average operator (considered as the baseline operator) for each user category and on queries related to the interests expressed in the users' profiles, we also present the comparative evaluation of the various aggregation schemes on queries not related to the users' interests. In fact, a situation which may often appear in a personalized IR setting is when a user formulates a query which has no intersection with his/her interests. The aim of this last comparative evaluation is to show that the quality of the ranking produced by the priority based operators does not decrease w.r.t. the ranking produced by applying the classical average operator.

In Section 5.1 we present the experimental settings, while in Section 5.2 we discuss the results produced by the performed evaluation.

5.1. Experimental Settings

The four relevance criteria as well as the considered aggregation schemes have been implemented on top of the Apache Lucene open-source API [2].

The method we have used to generate both queries and user's profiles is inspired to the approach proposed by Sanderson in [25]. In this work, the author presents a method to perform IR evaluations by using the Reuters collection that does not have queries nor relevance judgments, but has one or more subject codes associated with each document. As test collection we have used the Reuters RCV1 Collection (806,791 documents), which we have split into two subsets: a set **"Q"**, which we have used to generate the user profiles, as it will be explained here below, and a test set **"T"**.

Document Collection. The set **"T"** is used as the document collection, and it has been indexed with Lucene. This set is composed by 403,395 documents.

Users' Profiles. As previously said, the set **"Q"** has been used to build the simulated user's profiles. Documents in the Reuters collection have an associated subject code; based on this subject code, we have defined 30 users' profiles by applying the following procedure: first we have generated 30 disjoint subsets of the set **"Q"**, based on the subject codes of each document (e.g. "sport", "science", "economy", etc.). In other words, for each of the 30 considered subject codes we have generated a subset of **"Q"** by grouping all documents with that subject code. As a second step we have generated a user profile for each of these subsets; each profile is formally represented as the weighted vector of terms obtained as the average of the document vectors belonging to the considered subset. An example of user's profile is shown in the first column of Table 2.

We have finally associated each user profile with one of the three user's categories previously defined (coverage seeker, appropriateness seeker, and cautious). The names (corresponding to the associated document category) of the generated user profiles are listed in Table 1, where also the associated user's category is specified.

For each user profile, the reliability criterion has been implemented by exploiting the agency information included in each document of the collection: for each agency the user has set a trust value that has been used for computing the final document score.

Users' Queries. for each user profile, 10 users' queries have been manually defined by analyzing the documents belonging to the subset of **"Q"** from which the profiles have been generated. From the selected documents some meaningful and related terms have been selected to generate a set of 10 queries for each user profile. All the 10 queries are topically related to the associated user's profile, and they are composed by a few terms, in order to simulate real users' behavior. In the second column of Table 2 the user's queries associated with the profile reported in the first column are listed.
Moreover, with each user profile, a set of 10 queries has also been associated with terms not related to the topical interests represented in the profile. To do so, we have simply

[2]See URL `http://lucene.apache.org/`.

taken the queries associated with a distinct profile, and used them in the context of another user profile.
A total number of 10*30 = 300 queries has then been generated.

Evaluation metrics. To evaluate the quality of the rankings produced by the various aggregation operators the following measures are used: the normalized discounted cumulated gain (nDCG) [16], and the Precision at n. The nDCG is a measure defined to compare IR methods with respect to their ability to favor relevant search results in top positions of the ranked list of results.

PROFILE	USER INTENT
ARTS, CULTURE, ENTERTAINMENT FASHION MILLENNIUM ISSUES OBITUARIES RESEARCH, DEVELOPMENT SCIENCE, TECHNOLOGY SPORTS TRAVEL, TOURISM WAR, CIVIL WAR WEATHER	Coverage Seeker Users
BIOGRAPHIES, PERSONALITIES, PEOPLE CRIME, LAW ENFORCEMENT DEFENCE DISASTERS, ACCIDENTS DOMESTIC POLITICS ENVIRONMENT, NATURAL WORLD INDUSTRIAL PRODUCTION INTERNAL POLITICS INTERNATIONAL RELATIONS LABOUR ISSUES	Appropriateness Seeker Users
ECONOMY, FINANCE EDUCATION ELECTIONS EUROPEAN COMMUNITY GOVERNMENT, SOCIAL HEALTH HUMAN INTEREST MARKETS RELIGION WELFARE, SOCIAL SERVICES	Cautious Users

Table 1: The 30 profiles used in the experiments and the correspondent user intent for each profile.

5.2. Discussion of the Results

All queries related to the 30 user's profiles have been first separately evaluated with respect to each of the four relevance dimensions (aboutness, coverage, appropriateness and reliability). For a given document, the matching between the query vector and the document vector is first computed, thus obtaining the aboutness score; then, the coverage and the appropriateness criteria are evaluated by comparing the document vector with the user profile vector. Finally, the value of the reliability criterion, which corresponds to the degree to which the user trusts the source from which the document comes, is taken into account.

SPORT PROFILE		PERFORMED QUERIES	
match	1.000	Q_1	premiership league score
cup	0.931	Q_2	european cup match
game	0.881	Q_3	premiership players strike
team	0.769	Q_4	australia england rugby
league	0.644	Q_5	cricket match score
world	0.614	Q_6	tennis wimbledon players
season	0.562	Q_7	american football playoffs
final	0.539	Q_8	tennis players injuries
club	0.529	Q_9	basketball players contract
champion	0.485	Q_{10}	sydney olimpic results
player	0.483		
championship	0.454		
goal	0.448		
score	0.424		
coach	0.393		

Table 2: Example of a user's profile (showing only the first 15 terms), and the performed queries.

The obtained scores are then aggregated by the following three aggregation schemes, according to the order induced by the related user's category, thus producing for each query three different rankings:

- averaged: the score of each document is computed by averaging the satisfaction degrees of each criterion;

- prioritized "scoring" aggregation: the score of each document is computed by aggregating the satisfaction degree of each criterion by using Equation 2.

- prioritized "and" aggregation: the score of each document is computed by aggregating the satisfaction degree of each criterion by using Equation 4.

In order to evaluate the produced ranking we have selected 30 assessors, one per each user profile; the assessors are people working in our Universities (PhD students and researchers), who have been instructed to learn the topics expressed in the profile, and to evaluate the rankings based on the search intent of their associated user's category.

For each ranking the assessors have been asked to analyze the top 15 documents, and to determine for each of them the document relevance (in a binary way) with respect to both the search intent of the associated user's category, and the considered query. The rationale behind the decision of evaluating only the top 15 documents is that the majority of search results click activity (89.8%) happens on the first page of search results [34]. In fact, generally users only look at the first 10 to 20 documents in the ranked results list.

We have then averaged the evaluation scores over all users belonging to each of the three considered categories. In Figures 1, 2, 3, and 4 we show the *Precision* graphs computed respectively by considering the coverage seeker users, the appropriateness seeker users, the cautious users, and by considering all users. In Figures 5, 6, 7, and 8 we show the corresponding $nDCG$ graphs.

In each graph we show three curves: the two curves obtained by applying the two operators presented in this paper, and the curve related to the application of the baseline operator, i.e. the average aggregation operator.

By considering the graphs it is interesting to notice that the rankings produced by the two prioritized aggregation operators are always better than the ranking produced by the average operator; moreover, the two operators present a different behavior with respect to the considered users categories. In fact, as it may be observed in Figure 1 and in Figure 5, the "scoring" operator performs better for coverage seeker users, while for the appropriateness seeker users the "and" operator outperforms the "scoring" one (see Figures 2 and 6).

This different behavior is not surprising, as it is related to the different specificity of the coverage and appropriateness criteria. As it may be inferred from Equation 8, generally, the appropriateness score of document is a low value because it is easier to find a document focused on a higher number of topics than the number topics represented in a user profile, instead of finding a document that perfectly matches a user profile. It is much easier to find a document having a high coverage score because the computation of the coverage criterion takes into account only how much user interests are contained in a document (Section 4.1.2). This way, when we compute the overall document score by applying the "scoring" operator, the difference of the contribution of the low priority criteria is more significant when we consider the coverage seeker user's category with respect to the appropriateness seeker one, because in this last case the appropriateness criterion is the highest priority criterion.

The opposite behavior of the prioritized "and" operator with respect to the prioritized "scoring" one is caused by the fact that the overall document score strongly depends by the least satisfied criterion. This way, it may happen that sometimes there are documents that, even if the score of the most important criterion is high (for example the coverage criterion in the coverage seeker aggregations), the overall document score is decreased by one of the low priority criteria. In this particular study, we have noticed that the performance of the "and" operator in the coverage seeker aggregations is influenced by the reliability criterion. In fact, if we consider a document in which the scores of the coverage and aboutness criteria are high, while the score of the reliability criterion is low, the overall document score is lower than the score of a document that, for example, has medium scores for all criteria. Therefore, the probability that documents that are not considered relevant by users appear in the ranks, increased. Instead, when we consider the appropriateness seeker users, the "and" operator is more effective than the "scoring" due to the highest priority of the appropriateness criterion. In this scenario the low score values computed on the appropriateness criterion strongly influence the overall document score; therefore, the probability that documents that are strongly related to the user profile appear in the top positions of the ranking increases.

Regarding the scenario related to cautious users, the reliability criterion is considered as the highest priority criterion in the aggregation. This criterion represents the reliability of a document source for a user (see Section 4.1.4). By observing Figures 3

and 7), we may conclude that by documents coming from with highly reliable sources might be ranked in top positions even though they weakly satisfy the aboutness, coverage or appropriateness criteria. This is witness by the relative positions of the curves related to the prioritized aggregation operators with respect to the baseline curve.

The differences between the cautious users scenario and the other two scenarios may be also inferred by observing how the curves decrease. On the coverage seeker and appropriateness seeker graphs, the $nDCG$ values are always over the value of 0.5 with peaks of 0.738 for the "scoring" operator on the coverage seeker graph, and of 0.801 for the "and" operator on the appropriateness seeker graph. Instead, by observing the cautious graph, we may notice that the curves decrease at values around 0.4.

The overall curves showed in Figure 4 and 8 clearly demonstrate the better behavior of the two prioritized operators with respect to the baseline operator.

In Table 3 we have the t-Test to analyze the performance of our operator by comparing the values of the MAP@15. The results show that the improvement obtained by the presented operators with respect to the baseline are statistically significant.

Precision@X	Baseline	AND	Scoring
P@1	0.727	0.903	0.917
P@2	0.712	0.878	0.907
P@3	0.687	0.861	0.881
P@4	0.673	0.831	0.868
P@5	0.657	0.821	0.845
P@6	0.639	0.796	0.823
P@7	0.620	0.772	0.801
P@8	0.606	0.755	0.775
P@9	0.587	0.737	0.756
P@10	0.570	0.717	0.734
P@11	0.553	0.699	0.714
P@12	0.540	0.679	0.696
P@13	0.527	0.660	0.675
P@14	0.511	0.641	0.657
P@15	0.493	0.622	0.640
MAP@15	**0.363**	**0.541**	**0.558**
t-Test w.r.t. Baseline	-	99.356%	99.962%

Table 3: Precision@ and t-Test significance results.

Finally, in Figures 9 and 10 we present a summary of the results obtained by the assessors in the case of queries that are not related to their associated profiles. Also in this case each user performed 10 short queries that contained terms that was not present in their profile (we recall that these queries have been selected among those associated with the other user's profiles). (or present with a very low interest degree). In this case, we expected that the rankings produced by the three aggregations are

comparable. As it may be seen in Figures 9 and 10, this fact is verified only for the "scoring" operator. In fact, its curve remains close to the baseline for all graph points; this is given because when the user-dependent criteria are weakly satisfied, the aboutness plays the major role in determining the document overall score. This way, the effectiveness of the produced ranks is similar. Instead, concerning the results obtained by the prioritized "and" operator, we can notice that such operator is not suitable when the user formulates a query that is not related with his profile. Indeed, if the query is not related with the user profile, criteria like coverage and appropriateness are weakly satisfied and although the overall score depends also on the importance degree of the least satisfied criterion, the gap with the baseline can be higher.

6. Conclusions

In this paper two prioritized aggregation operators have been proposed to the aim of offering a new aggregation scheme for multiple relevance assessments.

The effectiveness of the two operators has been evaluated in a personalized IR scenario, by identifying three user's categories, related to distinct search intents that induce different priority orders over the considered relevance dimensions.

The prioritized "scoring" operator models a situation where the weight of a less important criterion is proportional to the satisfaction degree of more important criteria. The performed evaluations have shown that the prioritized 'scoring" operator allows to improve the ranking of the documents which are related to the user interests, when the user formulates an interest-related query, and when a user has no interests or formulates a query which is not related to his interests, the ranking of the documents is similar to the ranking obtaining by using the average operator.

The second is the prioritized "and" operator. The peculiarity of this operator, which also distinguishes it from the traditional "and" operator, is that the extent to which the least satisfied criterion is considered in the overall satisfaction degree depends both on its satisfaction degree and on its importance for the user. This model is suited to improving document ranking when every requirement (criterion) is essential and no requirement can be dropped without defeating the purpose of the user interests, and when we dispose of a user *preference order* on these requirements. The performed evaluations show that the proposed operator improves the ranking of the documents which are related to the user interests, when the user formulates an interest-related query, otherwise this operator is not suitable.

An advantage of the proposed operators is that they allow to calculate the weights of the criteria in a simple way and without requiring any learning method.

References

[1] D. Artz and Y. Gil. A survey of trust in computer science and the semantic web. *Web Semant.*, 5(2):58–71, 2007.

[2] C. L. Barry. User-defined relevance criteria: an exploratory study. *J. Am. Soc. Inf. Sci.*, 45(3):149–159, 1994.

[3] P. Borlund. The concept of relevance in ir. *J. Am. Soc. Inf. Sci. Technol.*, 54(10):913–925, 2003.

[4] P. Borlund and P. Ingwersen. Measures of relative relevance and ranked half-life: Performance indicators for interactive ir. In *SIGIR*, pages 324–331, 1998.

[5] P. Bruza, D. Song, and K. Wong. Aboutness from a commonsense perspective. *Journal of the American Society for Information Science*, 51:1090–1105, 2000.

[6] P. A. Chirita, C. S. Firan, and W. Nejdl. Personalized query expansion for the web. In *SIGIR '07*, pages 7–14. ACM, 2007.

[7] W. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7(1):19–37, 1971.

[8] W. S. Cooper. On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*, 24(2):87–100, 1973.

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

[10] C. da Costa Pereira, M. Dragoni, and G. Pasi. Multidimensional relevance: A new aggregation criterion. In M. Boughanem, C. Berrut, J. Mothe, and C. Soulé-Dupuy, editors, *ECIR*, volume 5478 of *Lecture Notes in Computer Science*, pages 264–275. Springer, 2009.

[11] C. da Costa Pereira, M. Dragoni, and G. Pasi. A prioritized "and" aggregation operator for multidimensional relevance assessment. In R. Serra and R. Cucchiara, editors, *AI*IA*, volume 5883 of *Lecture Notes in Computer Science*, pages 72–81. Springer, 2009.

[12] M. Farah and D. Vanderpooten. A multiple criteria approach for information retrieval. In *SPIRE*, pages 242–254, 2006.

[13] S. Fujita. Reflections on "aboutness" trec-9 evaluation experiments at justsystem. In *TREC*, 2000.

[14] S. P. Harter. Psychological relevance and information science. *Journal of American Society for Information Science*, 43(9):602–615, 1992.

[15] B. Hjørland. Towards a theory of aboutness, subject, topicality, theme, domain, field, content . . .and relevance. *J. Am. Soc. Inf. Sci.*, 52(9):775, 2001.

[16] K. Järvelin and K. J. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[17] D. Kelly, X. Fu, and C. Shah. Effects of position and number of relevant documents retrieved on users' evaluations of system performance. *ACM Trans. Inf. Syst.*, 28(2), 2010.

[18] F. Liu, C. Yu, and W. Meng. Personalized web search by mapping user queries to categories. In *CIKM '02*, pages 558–565. ACM, 2002.

[19] M. Maron. On indexing, retrieval and the meaning of "about". *Journal of the American Society for Information Science*, 28:38–43, 1977.

[20] S. Miyamoto. Information clustering based on fuzzy multisets. *Inf. Process. Manage.*, 39(2):195–213, 2003.

[21] S. Mizzaro. How many relevances in information retrieval? *Interacting With Computers*, 10(3):305–322, 1998.

[22] G. Pasi, G. Bordogna, and R. Villa. A multi-criteria content-based filtering system. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 775–776, 2007.

[23] D. Petrelli. On the role of user-centred evaluation in the advancement of interactive information retrieval. *Inf. Process. Manage.*, 44(1):22–38, 2008.

[24] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, 1984.

[25] M. Sanderson. The reuters collection. In *Proceedings of the 16th BCS IRSG Colloquium*, 1994.

[26] T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343, 1975.

[27] T. Saracevic. The stratified model of information retrieval interaction: Extension and applications. *Proceedings of American Society for Information Science*, 34:313–327, 1997.

[28] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. part ii: nature and manifestations of relevance. *J. Am. Soc. Inf. Sci. Technol.*, 58(13):1915–1933, 2007.

[29] L. Schamber and M. Eisenberg. Relevance: The search for a definition. In *Proc. 51st Annual Meeting of the American Society for Information Science*, 1988.

[30] L. Schamber, M. Eisenberg, and M. Nilan. A re-examination of relevance: toward a dynamic, situational definition*. *Information Processing & Management*, 26(6):755–776, 1990.

[31] J. A. Shaw and E. A. Fox. Combination of multiple searches. In *Text REtrieval Conference*, pages 243–252, 1994.

[32] A. Sieg, B. Mobasher, and R. Burke. Web search personalization with ontological user profiles. In M. Silva, A. Laender, R. Baeza-Yates, D. McGuinness, B.Olstad, Ø. Olsen, and A. Falcão, editors, *CIKM*, pages 525–534. ACM, 2007.

[33] D. Soergel. Indexing and retrieval performance: the logical evidence. *J. Am. Soc. Inf. Sci.*, 45(8):589–599, 1994.

[34] A. Spink, B. Jansen, C. Blakely, and S. Koshman. A study of results overlap and uniqueness among major web search engines. *Inf. Process. Manage.*, 42(5):1379–1391, 2006.

[35] A. R. Taylor, C. Cool, N. J. Belkin, and W. J. Amadio. Relationships between categories of relevance criteria and stage in task completion. *Inf. Process. Manage.*, 43(4):1071–1084, 2007.

[36] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *SIGIR '05*, pages 449–456. ACM, 2005.

[37] Y. C. Xu and Z. Chen. Relevance judgment: What do information users consider beyond topicality? *J. Am. Soc. Inf. Sci. Technol.*, 57(7):961–973, 2006.

[38] R. Yager. Modeling prioritized multicriteria decision making. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 34(6):2396–2404, 2004.

[39] R. R. Yager. Prioritized aggregation operators. *Int. J. Approx. Reasoning*, 48(1):263–274, 2008.

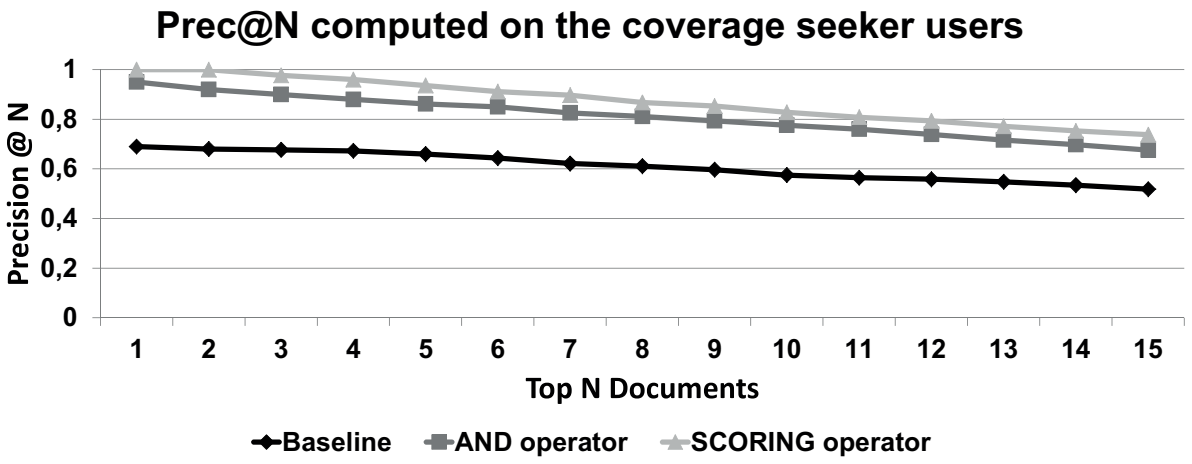

Figure 1: Precision@N computed on the ranks produced by considering coverage seeker users.

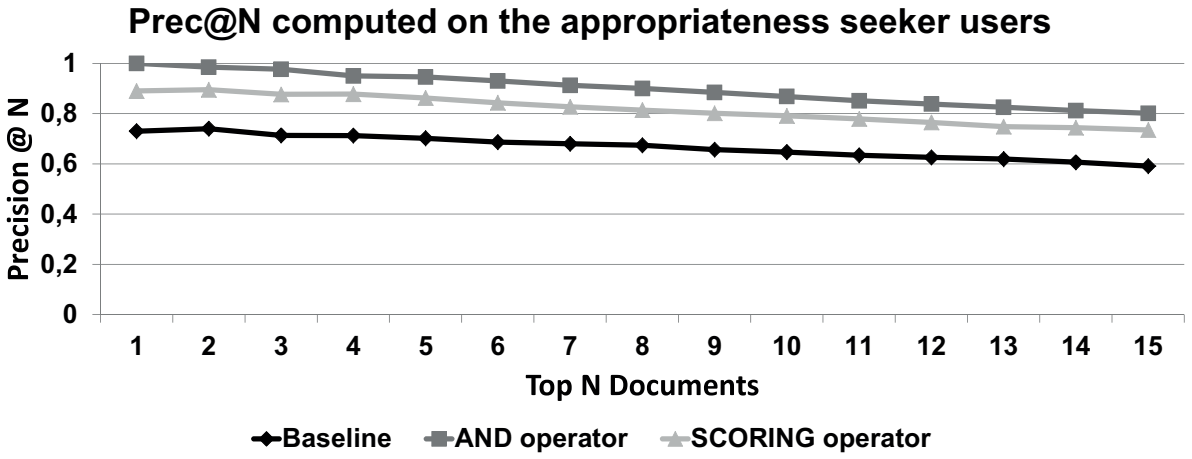

Figure 2: Precision@N computed on the ranks produced by considering appropriateness seeker users.

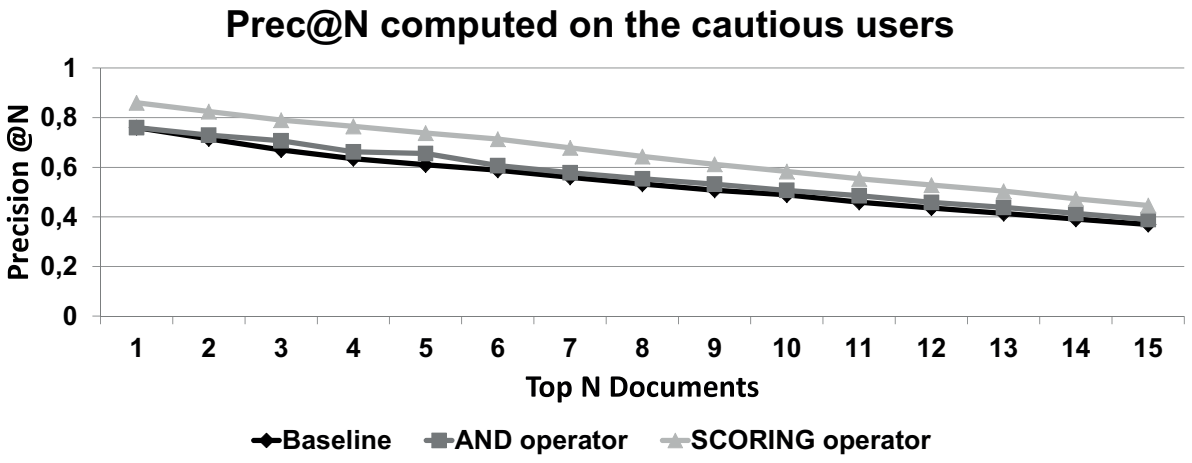

Figure 3: Precision@N computed on the ranks produced by considering cautious users.

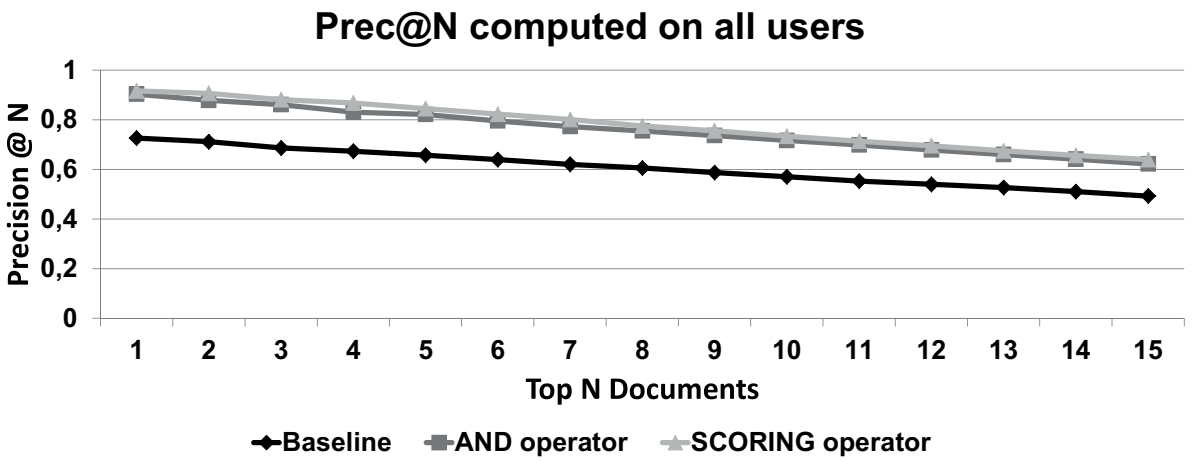

Figure 4: Precision@N computed on the ranks produced by considering all users.

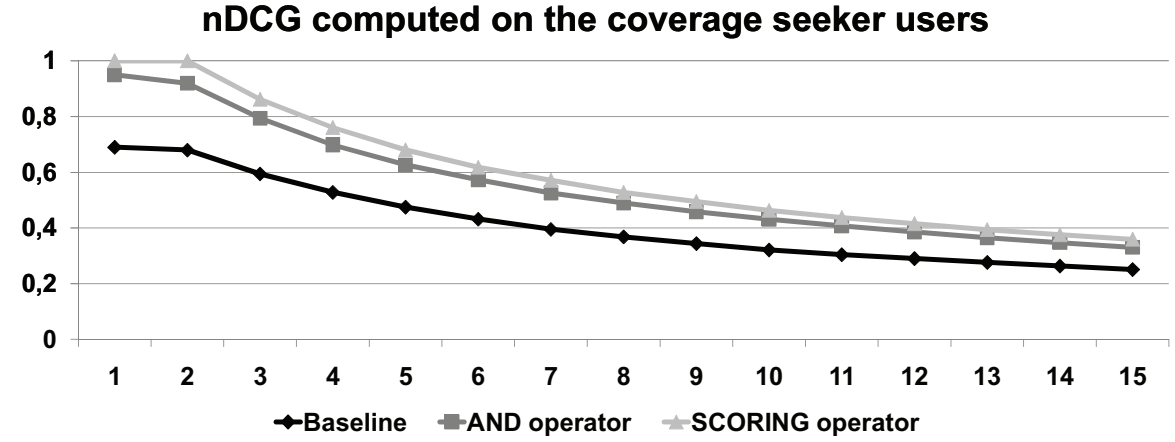

Figure 5: nDCG rank computed on the ranks produced by considering coverage seeker users.

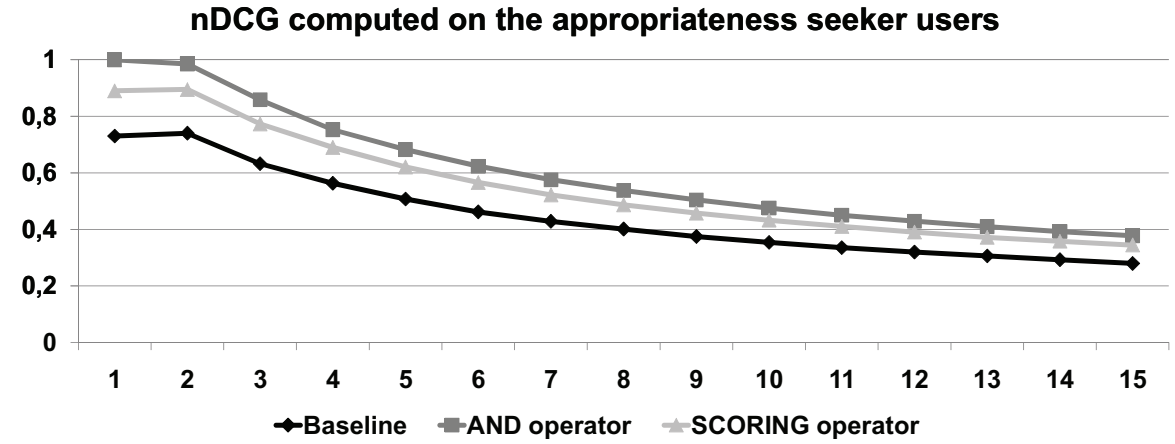

Figure 6: nDCG rank computed on the ranks produced by considering appropriateness seeker users.

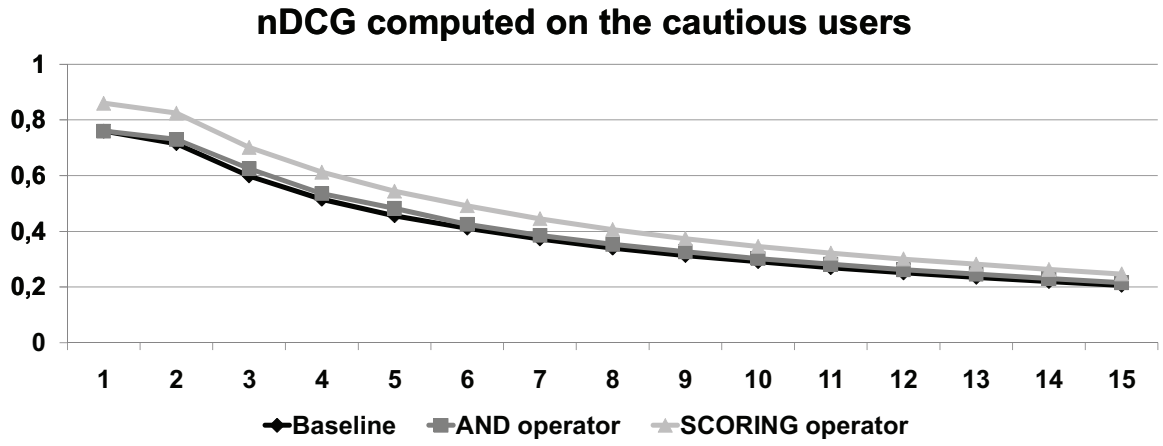

Figure 7: nDCG rank computed on the ranks produced by considering cautious users.

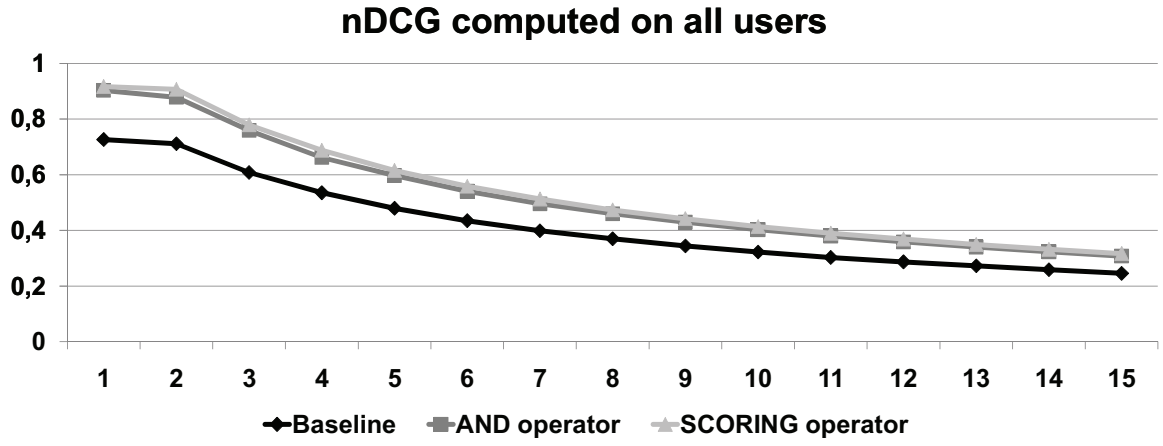

Figure 8: nDCG rank computed on the ranks produced by considering all users.

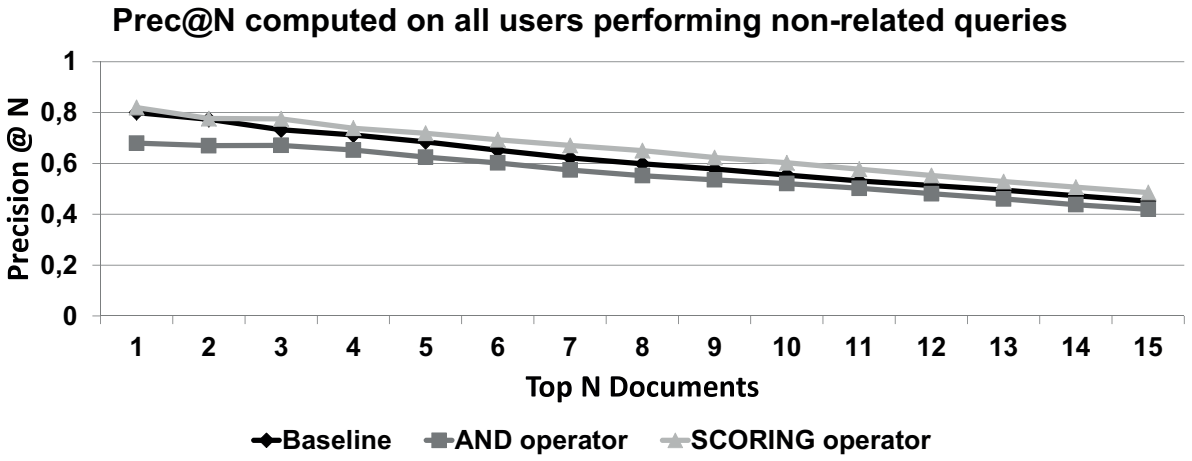

Figure 9: Precision@N computed on the ranks produced by considering all users performing non-related queries.

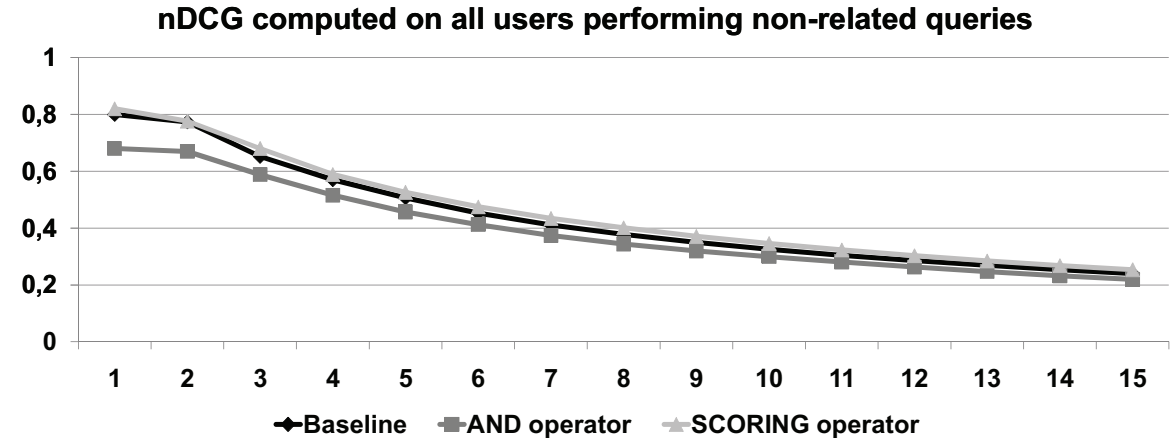

Figure 10: nDCG rank computed on the ranks produced by considering all users performing non-related queries.