Measures of Relative Relevance and Ranked Half-Life: Performance Indicators for Interactive IR

Pia Borlund
Dept. of IR Theory
Royal School of Library and Information Science
Email: {pbj@db.dk}

Peter Ingwersen
Dept. of IR Theory
Royal School of Library and Information Science
Email: {pi@db.dk}

Abstract This paper introduces the concepts of the relative relevance (RR) measure and a new performance indicator of the positional strength of the retrieved and ranked documents. The former is seen as a measure of associative performance computed by the application of the Jaccard formula. The latter is named the Ranked Half-Life (RHL) indicator and denotes the degree to which relevant documents are located on the top of a ranked retrieval result. The measures are proposed to be applied in addition to the traditional performance parameters such as precision and/or recall in connection with evaluation of interactive IR systems. The RR measure describes the degree of agreement between the types of relevance applied in evaluation of information retrieval (IR) systems in a non-binary assessment context. It is shown that the measure has potential to bridge the gap between subjective and objective relevance, as it makes it possible to understand and interpret the relation between these two main classes of relevance used in interactive IR experiments. The relevance concepts are defined, and the application of the measures is demonstrated by interrelating three types of relevance assessments: algorithmic; intellectual topicality and; situational assessments. Further, the paper shows that for a given set of queries at given precision levels the RHL indicator adds to the understanding of comparisons of IR performance.

1 Introduction

Evaluation of interactive information retrieval (IIR) systems presents a notable challenge, as the dynamic nature of an information need has been acknowledged [1, 2, 16 and 22]. Similarly, the awareness of the multi-dimensionality and variability of relevance has changed the understanding of how evaluation of IIR systems can be carried out, and has recently lead to changes of the test settings in the direction of an interactive user-centred approach [3, 4 and 18].

From the traditional system-driven point of view a user-centred approach causes problems in terms of lack of control of variables as well of the extent to which subjectivity is involved -- a subjectivity which modifies the overall test results as the end-users interpret and subjectively assess the relevance of the retrieved objects. The argument from the researchers representing the user-centred approach is that this dimension is exactly the rationale which makes their results trustworthy and reliable. From the traditional viewpoint relevance can only be measured in terms of so-called objective topical relevance assessments. The core of the discussion between the two major approaches to the evaluation of IR systems, i.e. the system-driven versus the interactive user-centred approach, has been clearly summarised by Robertson and Beaulieu:

> "The conflict between laboratory and operational experiments is essentially a conflict between, on the one hand, control over experimental variables, observability, and repeatability, and on the other hand, realism." [18, p. 460]

A specific issue in the current IR evaluation setting is constituted by the role and intellectual performance of the human assessor in, for instance, the TREC experiments [8-12]. Basically, such assessments are interpretations of the topics and the retrieved objects. Consequently, they cannot, scientifically speaking, be strictly objective. In a cognitive sense assessors are like users, that is, subjective in their assessments; in particular when they originally have generated the experimental query or topic. Although an assessor is supposed to act like an algorithmic entity in order to produce a strictly objective performance baseline, he or she will, to a degree, involuntarily become dependent of interpretations and subjectivity.

This paper proposes a pragmatic solution of how to bridge the gap between subjective and objective relevance -- the two main classes of relevance applied to performance evaluation of IR systems, in particular IIR systems. The goal is to allow for an improved understanding and interpretation of the more or less objective and clearly subjective relevance judgements made in experimental IIR, for instance, algorithmic, topical, and situational relevance assessments based on identical information needs. One consequence of the multi-dimensional relevance scenario is the extent to which different types of objective and subjective relevance assessments are *associated* across several users and retrieval engines. Also, a question is what such associations between relevance types do *signify* in terms of IIR performance. Another consequence is the fact that algorithmically ranked retrieval results become interpreted and assessed by users during session time. The judgements are then in accordance with the users' dynamic and situational perceptions of a real or simulated information retrieval task. In addition, the assessments may incorporate

non-binary values. Aside from traditional performance measures like precision and/or recall the user-generated assessments necessarily point to a *positional measure* of the *quality* of the ranked output: for a given type of relevance, say topicality, how *far up* the ranked list does an engine actually place the relevant objects. Seen from the user's perspective the higher the engine can place relevant objects the better the system.

For the associative relations the suggestion is to compute a *relative relevance (RR) measure* between the types of relevance assessments by use of the symmetrical and associative Jaccard formula [26]. Similar ideas of linking and describing relevance-relations by use of association measures have been suggested in 1964 by Hillman [14] though the purpose was different. The assumption behind this paper is that an associative interrelationship exists between the various types of relevance which may indeed be expressed by associative relations. The RR measure is thus supposed to yield quantitative information about the performance during IIR in addition to the traditional recall and precision measures. In the present case we make use of relevance data collected during IIR experiments involving actual users, an assessment panel, and ranked output [3].

For the positional measure of ranked retrieval results we propose a *Ranked Half-Life (RHL) indicator*. In principle it functions as the Median value of grouped data like the "cited half-life" indicator known from bibliometrics or as the "half-life" concept used in Nuclear Physics.

The paper is organised as follows: Section 2 presents a brief introduction to the concepts of relevance by defining their two existing main classes and sub-categories. Section 3 presents the Jaccard formula and demonstrates the application of the RR measure on subjective and objective relevance assessments. The 4th section draws attention to the issues of comparison of ranked output by presenting and applying the Ranked Half-Life indicator. The concluding section discusses the findings, also in relation to TREC, and summarises the major points presented in this paper.

2 The concepts of relevance

From the development of the first information retrieval systems in the 1950's the main objective has been the retrieval of *relevant* information [21]. Relevance has remained the primary evaluation criterion for the majority of the studies at the information processing level [19]. Since the Cranfield experiments [5] the debate concerning the concept of relevance has formed an important part of the discussions in the field of Information Science [20]. In this discussion both Saracevic [21] as well as Harter [13] divide the concepts of relevance into two main classes:

- objective *or* system-based relevance
- subjective *or* human (user)-based relevance

Since the two main classes of relevance are quite *different in nature* and by default imply different degrees of intellectual involvement, we do not refer to their interrelationships as similarities. We regard their relation as a *relative relevance association* or *nearness*, in spite of the application of the Jaccard measure with which one commonly computes similarities.

2.1 Objective relevance

The objective relevance is also known as "algorithmic" [21] or "logical" relevance [6, 7] and can be defined as a *topicality measure*, in the sense of:

> "...how well the topic of the information retrieved matches the topic of the request. A document is objectively relevant to a request if it deals with the topic of the request." [13, p. 602]

Topicality (Figure 1) seems the most common and clearest definition of relevance, and is the measure applied in traditional evaluation of IR systems. But the prerequisite for the definition is to understand the concept of *topic*. The objective-oriented type of topicality relevance is restricted to deal only with the degree to which the query representation matches the contents of the retrieved information objects, i.e. topic equals *contents*. This type of relevance is context free and is, for instance, applied when the relevance of an information object is computed as a function of the number of features in common between the query representation and the information objects, usually resulting in values between 0 and 1. The values are then used to rank the retrieved information objects by listing the most topically relevant information object first.

2.2 Subjective relevance

The subjective or user-based relevance is concerned with the aboutness and appropriateness of a retrieved information object and refers to the various *degrees of intellectual interpretations* carried out by human observers -- whether assessors or users. The subjective type of relevance may, as a generic concept, refer to the usefulness, usability, or utility of information objects in relation to the fulfilment of goals, interests, work tasks, or problematic situations intrinsic to the user. It is context-dependent.

According to Saracevic [21] and Ingwersen [15] four major categories of subjective relevance exist: a *topicality-like* type, associated with aboutness; *pertinence*, related to the information need as perceived by the user; *situational*, depending on the task interpretation, and; *motivational and affective*, which is goal-oriented. In the subjective topicality-like relevance category the concept of topic is understood as *aboutness*, not contents, i.e. an intellectual assessment of how an information object corresponds to the topical area required by the information need as perceived. This relevance measure is consequently not based on the relationship between a query representation and a retrieved information object. The judgement is made by an observer, either an assessor or a user. It is this kind of subjective relevance assessment we ascribe to individual assessors who participate in common IR experiments like TREC, although their judgements traditionally are intended to be of an objective nature. In the remaining of the paper we name this type of relevance as *intellectual topicality* (Figure 1).

Pertinence represents the intellectual relation between the intrinsic human information need and the information objects as currently interpreted or perceived by the cognitive state of an assessor or user. This allows for the existence of a *dynamic* information need. Like for human test persons or real users we are aware that pertinence may indeed be

involved in assessors' relevance judgements and, depending on the experimental scenario, strongly related to the more narrow intellectual topicality type.

Situational relevance is understood as the utility or *usefulness* of the viewed information objects by pointing to the relationship between such objects and the work task underlying the information need development and the current cognitive state as perceived by the observer (Figure 1).

Motivational and affective-oriented relevance is goal-oriented and associates to the overall intentionality pertaining to the user or test person. All subjective assessments are fundamentally individual. Any one assessor's or user's judgements are as valid as the assessments made by any other observer. The assessments depend on the actual cognitive context.

In this paper we apply the following three kinds of relevance as the criteria for evaluation of IIR systems and the demonstration of the RR and RHL measures: the *algorithmic topicality* type, carried out directly at the objective *processing* level of the system i.e. the ranked list of documents itself generated by the system; the *intellectual topicality*, related to the *output* level; and the *situational* type of relevance, associated with the *use* and *user* level [19].

3 Methodology and data sets

3.1 The Jaccard measure

The Jaccard measure [26] serves the purpose of quantifying a given relation between two sets of entities, in this case between the output of two types of relevance assessments (R_1, R_2). R_1 and R_2, respectively, are constituted by the assessment values as attributed by an assessor, a user, or an engine to the retrieved objects.

$$\text{Association}(R_1,R_2) = \frac{|R_1 \cap R_2|}{|R_1 \cup R_2|}$$

The Jaccard measure expresses the association by the intersection of the assessment values from the two types of relevance (R_1, R_2) relative to the union of the total number of assessments for a given retrieval situation. Provided that non-negative assessments are used the computed value of the association measure is between 0 and 1. The value 1 implies an identical match or total correspondence between the two types of relevance judgements. The Jaccard association coefficient is preferred to the cosine measure since we are dealing with pair of sets not vectors.

3.2 The data set and types of relevance

The data set used to demonstrate the application of the RR measure and the RHL indicator is taken from a test concerned with the methodological aspects of the evaluation of IIR [3]. The aim of that test was to gain indicative results of the workability and functionality of a proposed method and its sub-components, and not directly to compare the IR systems selected for the test purpose. The latter simply functioned as the instrumental apparatus for the former. The test setting was an operational online system which made it possible to evaluate two types of IR techniques within the same database environment in full-text (Knight Ridder Information, file 15, ABI/Inform on management information). The IR techniques were the ranked output facility Target [25] versus the Boolean-based Quorum technique. The collected sets of data consist of fifty-four search sessions deriving from the execution of the first retrieval run. The search sessions were performed by five persons: three test persons and two panel members. Each test person applied three simulated needs to the Target engine and assessed the outcome by use of *situational relevance* in the form of *usefulness*. A panel of two information professionals then performed the Boolean Quorum searches based on a direct transformation of each test person's three query formulations, and assessed the outcome by *intellectual topicality* and *situational relevance* (Table 1). Up to fifteen documents were relevance assessed per search session, distributed with reference to the types of relevance assessments in the following way:

Target engine:	**134**	(situational relevance);
	134	(intellectual topicality)
Quorum engine:	**246**	(situational relevance);
	246	(intellectual topicality)

The total number of human relevance judgements were 760. For each subjective relevance type the assessments were made in a non-binary way according to the three categories: highly relevant; partially relevant and; not relevant, as applied, for instance, by Pao [24] and Saracevic and colleagues [23]. In addition, the *algorithmic topicality* produced 270 assessments distributed over the two engines (see Figure 1).

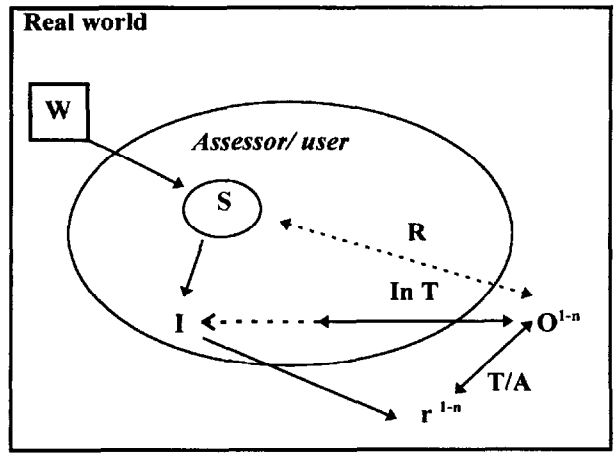

Figure 1. Illustration of the three types of relevance applied. Modified version of Borlund & Ingwersen [3].

Legend:

W : Work task
S : Cognitive perception of W
I : Information need version(s)
O : Information object(s)
r : Request version(s)
R : Situational relevance
T/A : Topical / Algirithmic relevance
In T : Intellectual topical relevance
 : Relevance assessment(s) or interpretation(s)
 : Transformation
 : Assessor / user's cognitive space

Figure 1 illustrates the three different types of relevance applied. The traditional and most commonly used type of relevance is the topical or algorithmic relevance (T/A) which expresses the degree of match between the request/query version (r^{1-n}) and the retrieved information object (O^{1-n}). Situational relevance (R), which has its main potential in connection with evaluation of interactive IR systems, is a measure of the relationship between the retrieved object (O^{1-n}) and the information need (I) for a given cognitive work task situation (S) originated from (W). The topical founded, though intellectually influenced type of relevance, intellectual topicality (In T), is signified by the topical nearness between the retrieved objects (O^{1-n}) and the information need (I).

The test made use of semantically open so-called "simulated work task situations" (W), which were employed by the individual panel members and test persons as a platform for request/query formulations and the situational relevance judgements. A simulated work task situation describes the information need scenario by providing information as to: (1) the source of need; (2) the environment of the situation; (3) the problem which has to be solved; and (4) make the test person understand the objective of the search [3]. Simultaneously, the simulated work task situation assures that the evaluation is controllable and that the relevance assessments are comparable.

3.3 The RR measure

The associative RR measure is applicable for a pair of different types of relevance assessments based on the same simulated work task situation and collection of objects. The applied simulated work task situations are referred to as A; B; and C. The various versions of the assessments of these work task situations are differentiated by the following number A1-3; B1-3 and C1-3.

A1: assessment values of simulated work task situation A, outcome version 1.

	Target					**Quorum**					
	Topicality	**Situational**	**Intellectual top.**			**Situational**			**Intellectual top.**		
Rank order	Algorithmic rank output:	Test person (nr. 1)	Panel a	Panel b	Panel a+b/2	Panel a	Panel b	Panel a+b/2	Panel a	Panel b	Panel a+b/2
1	0.99	0.5	1.0	0.5	0.75	1.0	1.0	1.0	1.0	1.0	1.0
2	0.87	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.5	0.75
3	0.86	1.0	1.0	1.0	1.0	0.5	0.5	0.5	0.5	0.5	0.5
4	0.86	0.0	0.5	0.0	0.25	0.0	0.0	0.0	0.0	0.0	0.0
5	0.86	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5	0.0	0.25
6	0.72	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	0.5	0.75
7	0.71	0.0	0.0	0.0	0.0	1.0	1.0	1.0	0.5	0.5	0.5
8	0.71	0.0	0.0	0.0	0.0	0.5	0.5	0.5	0.5	0.0	0.25
9	0.57	1.0	0.0	0.5	0.25	0.0	0.0	0.0	0.0	0.0	0.0
10	0.57	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	0.57	0.0	0.0	0.5	0.25	1.0	1.0	1.0	0.5	1.0	0.75
12	0.57	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	0.57	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	0.57	0.5	1.0	0.5	0.75	0.0	0.0	0.0	0.0	0.0	0.0
15	0.57	0.0	0.5	0.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0
SUM	*10.575*	*3*	*5*	*4.5*	*4.75*	*6*	*6*	*6*	*5.5*	*4*	*4.75*
Precision	**0.71**	**0.2**	**0.33**	**0.3**	**0.32**	**0.4**	**0.4**	**0.4**	**0.37**	**0.27**	**0.32**
RHL Indicator	6.18	3	2.5	2.75	2.63	5.5	5.5	5.5	4.5	3	4.52
RHL Index	8.7	15	7.58	9.17	8.22	13.75	13.75	13.75	12.16	11.11	14.13

value: 1.0 (Highly relevant); value: 0.5 (Partial relevant) and; value: 0.0 (Non-relevant)

Table 1. The distribution of the percentage values of the three types of relevance assessments for version A-1, for the Target and Quorum engines. In this particular case the Quorum engine's output, different from that from Target, takes the value 1.0 for all the first fifteen documents (not shown). The precision and Ranked Half-Life values, and the RHL index for normalised precision, are shown for each list of assessments.

The computation of the associative RR measure is done by matching the outcome of each of the two relevance judgements per assessed document. The relevance assessments are assigned with the value according to the applied categories of relevance: Highly relevant (1.0), Partial relevant (0.5) and, Non-relevant (0.0). The products of the values of the judgements for each pair of assessed documents defines the nominator of the Jaccard formula which is divided by the union of the values of the viewed and assessed information objects. An example of the basic values applied for the computation of the RR measure by use of the Jaccard formula is shown in Table 1. The overall statistically validated result of the test reported in [3] showed that the performance of the two engines was quite similar, measured by precision. The example, Table 1, demonstrates the case. Seen from the system's point of view the algorithmic topicality precision for the Target engine is high (0.71). But from the panel members' perspective its intellectual topical relevance is quite low (0.32). The Quorum engine's own performance at processing level is not shown on the table, since the Boolean Quorum technique did only produce an unranked output where each of the first fifteen documents algorithmically possesses the identical value of 1.0. However, the intellectual topicality assessed by the panel members resulted in a precision value of 0.32 identical to that of Target.

The overall test results also showed a substantial deviation in terms of intellectual topicality assessments the panel members in between. Table 1 demonstrates an example, for instance, between panel member a and b concerning the Quorum engine (0.37 versus 0.27). Since intellectual topicality is the preferred traditional baseline relevance measure, panel member a's assessments used as the single baseline would increase the performance of the Quorum engine (as well as of the Target engine).

In such cases of similar performance or differences between test persons' and panel members' assessments the RR measure is intended to produce valuable information, *supplementary* to the simplistic precision measures in IIR.

Table 2 shows the computed RR measures for the three types of relevance assessments.

The application of the RR measure to 3 types of relevance:

	Target			**Quorum**		
situation no:	Situational vs. Algorithmic	Intellectual topicality* vs. Algorithmic	Intellectual topicality* vs. Situational	Situational vs. Algorithmic	Intellectual topicality* vs. Algorithmic	Intellectual topicality* vs. Situational
A-1	0.19	0.32	0.35	0.4	0.32	0.62
A-2	0.32	0.19	0.37	0.15	0.11	0.18
A-3	0.42	0.38	0.63	0.35	0.57	0.48
mean:	0.31	0.30	0.45	0.3	0.33	0.43
B-1	0.46	0.44	0.48	0.61	0.68	0.67
B-2	0.32	0.39	0.24	0.58	0.65	0.69
B-3	0.57	0.17	0.21	0.68	0.78	0.64
mean:	0.45	0.31	0.31	0.62	0.70	0.67
C-1	0.40	0.29	0.47	0.13	0.08	0.24
C-2	0.33	0.38	0.30	0.33	0.33	0.55
C-3	0.56	0.23	0.31	0.55	0.7	0.48
mean:	0.43	0.30	0.36	0.34	0.37	0.42
mean of mean:	0.40	0.31	0.37	0.42	0.47	0.50

* The applied values of the intellectual topicality assessments are based on the average values of the two panel members.

Table 2. The relative relevance measure applied to three types of relevance, for the Target and Quorum engines.

For all of the retrieved test situations (A, B and C) we have the following associative RR measures:

		Target:	Quorum:
A	Situational relevance vs. Algorithmic relevance:	0.40	0.42
B	Intellectual topicality vs. Algorithmic relevance:	0.31	0.47
C	Intellectual topicality vs. Situational relevance:	0.37	0.50

Analysing the three RR values for the Target engine we obtain a more tangible understanding of the value and nature of subjective relevance assessments, compared to the performance of a system. With the algorithmic results as baseline the RR measure indicates that Target is slightly better for the retrieval of documents useful for a task than for finding topically relevant documents. The RR measures for the Quorum engine show a different pattern. Generally speaking the Quorum engine performs better than Target on all types of relevance. Further, the Quorum engine seems better at satisfying aboutness-related demands.

Ad A) The associative relation between *situational relevance* (subjective user/assessor judgements) and *algorithmic relevance* (objective) uncovers values which lead to:

(1) an understanding of how well the perceived work task situation, assessed through situational relevance, is satisfied by the ranked output retrieved by the systems. In this case we observe that the Quorum engine performs better than the Target technique.

(2) the degree to which the Highly and Partially relevant assessments relate to the baseline measures. The lower the value, the less correspondence exists between the systems' prediction of relevance and the observers' interpretation of the documents as useful to a given task.

Ad. B) In this case, a similar situation can be shown for the RR measure between *intellectual topicality* (subjective assessment) and *algorithmic relevance*. We are informed about to what extent the two types of topical relevance assessments match each other -- in the Target engine by a RR measure of 0.31, and Quorum by measure of 0.47. This tells us partly something about how well a system is capable of retrieving *predictable* topically relevant information objects and partly how well the same objects actually *are* topically relevant in the intellectual sense.

Ad. C) The third revealed value, the relation between *intellectual topicality* (subjective assessor judgements) and *situational relevance* (subjective assessor/user judgements), tells us about the *nearness* between the two subjective-oriented relevance measures. With an associative measure of 0.37 for Target we may conceive that although there is a high degree of equivalence in the match between the algorithmic and intellectual assessed topicality, this fact is *no guarantee* that the aboutness of the information objects also match the underlying purpose and work task against which the situational relevance is assessed. In the case of the Quorum engine which produced equivalent precision values (Table 1), we find a higher degree of association (0.50) between perceived aboutness and situational relevance among the observers.

The values of the RR measure generate a more comprehensive understanding of the *characteristics* of the performance of single or several retrieval engines and algorithms in between, in particular when confronted with users.

4 The Ranked Half-Life indicator

Performance presentation and comparison have been considered since the mid 1960's, and as no single best method has been developed, researchers continue to present data in similar ways [27, p. 491]. The research and discussions carried out on this issue is mostly done with reference to controlled laboratory experiments and tends to focus on recall oriented aspects [e.g. 17; 27]. The research is usually based on the assumption of the objectivity of relevance judgements made by single assessors. We adhere to the assumption that an assessor introduces a degree of subjectivity which *per se* adds to the experimental outcome and should be explored. Interesting though, little attention has been drawn to the *issue of rank performance* in this area of research.

Wallis and Thom [27] suggest in their paper a system-oriented way of how to handle and compare relevance assessments on a rather detailed level, computed as precision in a binary mode. The procedure they suggest takes into account the ranked position of the evaluated object by favouring higher positioned relevant objects. However, as illustrated in Table 1, the case can very well be that a lower positioned object is the one of real interest to the users for various cognitive reasons. Among the algorithmically ranked Target documents, number 1, 3, 9 and 14 are regarded highly or partially *useful* by the actual test person. The panel perceives the documents 1-3 topically relevant in the *intellectual sense*, but find also objects 14 and 15 quite appropriate.

As a consequence of two or more types of relevance assessments in IIR the issue of *comparisons* of computed retrieval rankings become critical. By taking into account the algorithmic *rank position* and the various assigned relevance values of the retrieved objects one takes advantage of two parameters: 1) the *algorithmically ranked order* which represents a list of decreasing degrees of *predicted* objective relevance to the user's information need, traditionally represented as a query; *and* 2) the applied subjective types and values of the relevance assessments representing the assessor's or user's interpretations of the ranked documents.

The proposed *Ranked Half-Life indicator* (RHL indicator) makes direct use of both parameters. The statistical method applied to calculate the RHL value corresponds to the computation of the Median of grouped continuous data. The RHL value is the median "case", i.e. the point which divides the continuous data area exactly into two parts. In Nuclear Physics the "half-life" of a specific radioactive material is the time taken for half the atoms to disintegrate. In Bibliometrics cited half-life is the time taken for half the citations to be given to a particular document. For the RHL indicator the time dimension is substituted by the continuous ranking of documents produced algorithmically by a retrieval engine. Each listed document represents a class of grouped data in which the frequency corresponds to the relevance value(s) assigned the document.

The idea behind the application of the median point of grouped data is the fact that if top-listed documents obtain high relevance scores the median ranking for a given document cut-off and a given precision value will rise. With scattered or low placed highly relevant documents the median "case" will drop downwards on the original output list of documents. Precision simply signifies the mean of the

cumulated frequency, also used for the median calculation; but it does not inform about ranked positions as illustrated above (Table 1).

The formula used for calculating the Ranked Half-Life indicator is the common formula for the median of grouped data:

$$M_g = L_m + \left(\frac{n/2 - \sum f2}{F(med)} \times CI \right)$$

where L_m = lower real limit of the median class, i.e. the lowest positioned document above the median class;

n = number of observations, i.e. the total frequency of the assigned relevance values;

$\sum f2$ = cumulative frequency (relevance values) up to and including the class preceding the median class;

$F(med)$ = the frequency (relevance value) of the median class;

CI = class interval (upper real limit minus lower real limit), commonly in IR = 1.

The cut-off in IIR is the number of documents retrieved and assessed by a person. In the present analysis the cut-off was reasonably set to 15 documents. Table 1 illustrates the different RHL indicator values for a particular simulated information need, according to two relevance types. The Target engine achieves a precision value of 0.2 associated with situational relevance (test person 1) and 3.0 as Ranked Half-Life indicator value. This means that, for that person, the Target engine is capable of providing half the cumulated relevance frequency of the 15 assessed documents within the first three listed documents. For the Quorum engine, however, the situational RHL indicator value is 5.5 for the same need as assessed by both panel members; the precision is of higher value (0.4).

For a given document cut-off one might prefer to obtain a *RHL index* value which equals the computed RHL value normalised for the corresponding precision value (precision = 1.0). Table 1 demonstrates the Ranked Half-Life index values for the situational and intellectual topicality relevance types. The index serves to emphasise the characteristics of the engines' ranking capabilities for the same precision values across relevance types.

As shown in Table 1 the Target engine's *algorithmically ranked list* of documents contains the engine's original relevance scale which is different from that applied to the subjective, human relevance assessments. Without data transformation of the algorithmic relevance values one *cannot compare* the algorithmic and subjective RHLs directly.

Realistically speaking, in IIR experiments the document cut-offs might vary according to the engagement of each test person – a situation which then has to be normalised.

Thus, for a given document cut-off and a given value of precision the RHL indicator can be examined across all test persons and all test tasks for each of the involved types of relevance. Compared to ordinary precision measures the RHL indicator supplies additional valid information about the *degree* to which the engine is capable of ranking its output according to user-related relevance.

5 Discussion and conclusion

The application of the associative RR measure and the RHL indicator to various types of relevance bridges the interpretative distance between the involved subjective and objective types of relevance in IIR. Both measures are thus indicative parameters of systems performance, and users' degrees of satisfaction, *supplementing* recall and precision. The need for additional measures is reinforced by the fact that relevance can be assessed based on entire information objects, e.g. full-text documents. A result of this development is that IR evaluations will be carried out based more on the user-centred approach. The RR and RHL measures will in addition be of value in experimental scenarios in which users participate with their own information situations, since only the users in such cases are able to assess the relevance of the retrieved objects.

The Relative Relevance measure as well as the Ranked Half-Life indicator can obviously be applied to non-interactive IR experiments like TREC which include algorithmic rankings and assessors' relevance values assigned to these rankings. In TREC-like experiments the RR measure can be used directly to the two different types of assessments: the *algorithmic* and the *intellectual topicality* -- also across retrieval engines. The Ranked Half-Life indicator can also be applied across systems -- but either directly on the algorithmic level or limited to the subjective level alone. Transformation of scaling may solve this problem. The non-interactive scenario possesses one advantage over the IIR setting: the number of ranked and assessed documents per search session is vastly higher than commonly in IIR. This fact entails more elaborate performance analysis possibilities associated with both the proposed measures.

The computed measures allow in addition for an improved understanding and interpretation of the two basic classes of relevance in between. The measures seek to support an existing demand in IIR experimentation for:

- a measure which can bridge between the test-person's subjective relevance assessments and a given system's objective performance and;
- lead to a higher degree of attention drawn towards the nature of the computed performance data, by means of: 1) the position (rank) of the assessment and, 2) the degree of subjectivity (the user-indicated value) of the relevance assessments. Further, they point to a way of making the performance data comparable.

The Relative Relevance measure satisfies the need for interrelating the various types of relevance applied in evaluation of IR systems. The Ranked Half-Life indicator accomplishes to take into account both the position of the assessed information objects and the degree of subjectivity involved by providing information as to how well a system is capable of satisfying a user's need for information at a given level of relevance.

Acknowledgements

We would like to thank the six anonymous SIGIR reviewers for their helpful and constructive comments.

References

[1] Belkin, N.J. Cognitive models and information transfer. *Social Science Information Studies, (4)*, 1984, 111-129.

[2] Belkin, N.J., Oddy, R. & Brooks, H. ASK for information retrieval: Part I. *Journal of Documentation, (38)*, 1982, 61-71.

[3] Borlund, P. and Ingwersen, P. The development of a method for the evaluation of interactive information retrieval systems. *Journal of Documentation, (53)*3, 1997, 225-250.

[4] Brajnik, G., Mizzaro, S., Tasso, C. Evaluating User Interfaces to Information Retrieval Systems: A case Study on User Support. *In*: Frei, H.P., Harman, D., Schäuble, P. ,Wilkinson, R., eds. *Proceedings of the 19th ACM Sigir Conference on Research and Development of Information Retrieval.* Zurich, 1996. Konstanz: Hartung-Gorre, 1996, 128-136.

[5] Cleverdon, C.W & Keen, E.M. *Factors determining the performance of indexing systems.* Vol. 1: Design. Vol. 2: Results. Cranfield, UK: Aslib Cranfield Research Project, 1966.

[6] Cooper, W.S. A Definition of Relevance for Information Retrieval. *Information Storage and Retrieval, (7)*1, 1971, 19-37.

[7] Cooper, W.S. On Selecting a Measure of Retrieval effectiveness, Part 1. *Journal of the American Society for Information Science, (24)*2, 1973, 87-100.

[8] Harman, D. Overview of the first TREC conference. *In*: Korfhage, R., Rasmussen, E., Willett, P., eds. *Proceedings of the 16th ACM Sigir Conference on Research and Development of Information Retrieval.* Pittsburgh, 1993. New York, N.Y.: ACM Press, 1993, 36-47.

[9] Harman, D. Overview of the second text retrieval conference (TREC-2). *Information Processing & Management, (31)*3, 1995, 271-289.

[10] Harman, D. (Ed.). *Overview of the Third Text REtrieval Conference (TREC-3).* National Institute of Standards and Technology Special Publication 500-225, Gaithersburg, Md. 20899.

[11] Harman, D. (Ed.). *Overview of the Fourth Text REtrieval Conference (TREC-4).* National Institute of Standards and Technology, Gaithersburg, Md. 20899.

[12] Harman, D.K. The TREC Conferences. *In*: Sparck Jones, K. and Willitt, P., eds. *Readings in Information Retrieval.* San Francisco: Morgan Kaufmann, 1997, 247-256.

[13] Harter, S.P. Psychological Relevance and Information Science. *Journal of American Society for Information Science, (43)*, 1992, 602-615.

[14] Hillman, D.J. The Notion of Relevance (1). *American Documentation, (15)*1, 1964, 26-34.

[15] Ingwersen, P. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation, (52)*1, 1996 , 3-50.

[16] Keen, E.M. Presenting results of experimental retrieval comparisons. *Information Processing & Management, (28)*4, 1992, 491-502.

[17] Ingwersen, P. Search procedures in the library analysed from the cognitive point of view. *Journal of Documentation, (38)*, 1982, 165-191.

[18] Robertson, S.E. and Hancock-Beaulieu, M.M. On the evaluation of IR systems. *Information Processing & Management, (28)*4, 1992, 457-466.

[19] Saracevic, T. Evaluation of Evaluation in Information Retrieval. *In*: Fox, E.A., Ingwersen, P., Fidel, R., eds. *Proceedings of the 18th ACM Sigir Conference on Research and Development of Information Retrieval.* Seattle, 1995. New York, N.Y.: ACM Press, 1995, 138-146.

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

[21] Saracevic, T. Relevance reconsidered '96. *In*: Ingwersen, P. and Pors, N.O., eds. *Information Science: Integration in Perspective.* Copenhagen: Royal School of Librarianship, 1996, 201-218.

[22] Saracevic, T., Mokros, H. & Su, L.T. Nature of interaction between users and intermediaries in on-line searching: qualitative analysis. *ASIS Proceeding*, 1990, 47-54.

[23] Saracevic, T. et al. A Study of Information Seeking and Retrieving: 1. Background and Methodology. *In*: Sparck Jones, K. and Willitt, P., eds. *Readings in Information Retrieval.* San Francisco: Morgan Kaufmann, 1997, 175-190.

[24] Pao, M.L. Term and citation retrieval: a field study. *Information Processing & Management, (29)*1, 1993, 95-112.

[25] Tenopir, C. & Cahn, P. Target & Freestyle: Dialog and Mead join the relevance ranks. *On-line*, (May), 1994, 31-47.

[26] van Rijsbergen, C.J. *Information Retrieval.* Second edition. London: Butterworths. 1979.

[27] Wallis, P., Thom, J.A. Relevance judgements for assessing recall. *Information Processing & Management, (32)*3, 1996, 273-286.