Retrieval Status Values in Information Retrieval Evaluation

Amélie IMAFOUO and Xavier TANNIER
Ecole Nationale Supérieure des Mines de Saint-Etienne
158 Cours Fauriel - 42023 Saint-Etienne, Cedex 2, France
imafouo@emse.fr, tannier@emse.fr

Abstract

Evaluation is one major step of an IR process. It helps to classify and compare Information Retrieval (IR) systems according to their effectiveness or their efficiency. Many IR works carried different kinds of studies about evaluation methods and many evaluation measures have been proposed. Most of these measures are based on the ranking of documents retrieved by IR systems in response to queries. The retrieved documents are ranked according to their retrieval status values if these are monotonously increasing with the probability of relevance of documents. However, few IR works tried to know more about these RSV and their possible use for IR evaluation. In this work, we analyze different RSV computations and investigate the links between the RSVs and the IR systems evaluation.

1 Introduction

The aim of information retrieval systems (IRSs) is to retrieve documents that are relevant to the user queries. To reach this goal, they attribute a value to each candidate document; afterwards, they rank documents in the reverse order of this value. This value is called the Retrieval Status Value (RSV). These rankings are used to evaluate and compare IRS. Despite its important role in IR evaluation, the RSV has not been widely studied and is still considered as a meaningless system value.

In this work, we try to understand the real link between the RSV and IR evaluation. We propose new evaluation measures directly based on the document RSVs and we compute correlations between classical IR measures and our measures.

2 Related research

2.1 IR evaluation and relevance

Hernon and McClure stated that evaluation is the process of identifying and collecting data about specific services, establishing criteria by which their success can be assessed, and determining both the quality of the service and the degree to which the service accomplishes

stated goals. Some classifications of evaluation in IR have been proposed, like the distinction between evaluation of performance (what happens during the process), and evaluation of outcome (description of the results). Tague-Sutcliffe provides a list of important questions about evaluation of the IR system as a whole or by individual components, and about interactive systems evaluation. *Kagolovsk and Moehr* [6] realised a detailed survey of main IR works on evaluation.

One of the first methodologies of IR evaluation is the Cranfield method. It requires to have complete relevance assessments on a corpus queried by a set of information needs. Relevance was always the main concept for IR Evaluation as the aim of an IR system is to find relevant documents. Many works studied the relevance issue and in the litterature many terms/expressions are used as synonyms: user satisfaction, utility, usefulness, pertinence, topicality. *Rees and Schulz* [13] noted that relevance judgements are affected by about 40 variables. *Cooper* [2], *Wilson* [17] proposed some definitions and formalizations of relevance. *Saracevic* [14] proposed a framework for classifying the various notions of relevance. All these works and many others suggest that there is no single relevance. Relevance is a complex social and cognitive phenomenon [15]. It is not a univocal yes/no decision and can vary not only from person to person, but also for a same person, depending on circumstances and context. Cooper for example brings out the difference between utility and relevance. It is a multidimensional concept that implies many human cognitive processes. *Mizzaro* [10] showed that there are many kinds of relevance; each of these relevance can be seen as a point in a four-dimensional space:

- information resources : document, surrogate, information
- representation of user's problem: Real information need (IN), perceived IN, request, query
- time
- components: a combination of topic, task, context

For example *rel(Information, Real IN, t(f), topic, task, context)* stands for the relevance of the information received to the Real information need at time *t(f)* for the topic, the task and the context, it is the relevance the user is interested in.

Relevance judgements (assignments of a value of relevance by a judge at a certain point of time) have also been classified along five dimensions: kind of relevance judged, kind of judge, which information resources the judge can use, which representation of user's need the judge can use for expressing his relevance judgement and the time at which the judgement is expressed [10].

Nowadays, with the Text REtrieval Conference (TREC) for example, the evaluation principle has not changed, but the relevance judgements are not complete any more because of the collections growth. The pooling technique [16] is used to collect a set of documents to be judged by human assessors. It provides a common basis for IR reproducible and comparable

experiment but it encounters some limits and some propositions have been made to improve it:

- A variation of the standard pooling strategies can increase the number of relevant documents found for assessments [18].
- The interactive Searching and Judging methodology uses an interactive information retrieval system (IRS) to select the assessed documents and the Move-to-Front pooling selects a variable number of documents for each system according to their retrieval performances [3].
- A pseudo-relevance judgment methodology in which human judges are replaced by a randomly selected mapping of documents to topics [15].

Several measures have been proposed and most of them apply to batch evaluations on a test collection (a corpus + a set of information needs + a set of assessments) like CACM, CISI or TREC. To answer the question of the choice of an appropriate evaluation measure, there is a need to define precisely what is being evaluated. These measures deal generally with system outputs.

2.2 Existing measures

Measures used to evaluate IRS are often based on the ranking of documents retrieved by these systems, and ranking is based on monotonously decreasing RSVs. Recall is a measure which gives the fraction of relevant documents which has been retrieved by the system. Precision is the fraction of all retrieved documents which is actually relevant. Precision and recall are the two most frequently used measures for example through the recall/precision curve. There are some other measures based on precision and recall like Average-precision at seen relevant documents or R-precision.

The use of recall and precision is quite problematic when one deals with large collections or when there is only a weak ordering between documents. To handle this latter problem, *Raghavan et al* [12] proposed the Probability of Relevance (PRR) and Expected Precision (EP) which measures the precision we can expect to obtain for a given recall level. Finally, some measures combine precision and recall to obtain a single value like the harmonic mean of recall and precision, the E-measure, which allows to weight the precision and the recall, according to the interest of the user [1] and the Expected search length [2], which gives the number of documents to read before finding relevant documents.

Korfhage [8] suggested a comparison between an IRS and a so-called ideal IRS. Thus the normalized recall is the distance between the recalls of a given system and the recalls of the ideal system which is defined as one giving all relevant documents before the first nonrelevant one. A normalized precision is defined in an analogous way. These measures keep the basic problems detailed below of ordinary recall. The sliding ratio, satisfaction, frustration are measures based on weighted relevance judgments that implies to have relevance weights

instead of binary judgment (relevant /non relevant) [8]. Several user-oriented measures have been proposed. The three typical ones are:

- coverage ratio: the proportion of relevant documents known to the user that are actually retrieved;
- novelty ratio: the proportion of the relevant retrieved documents that are previously unknown to the user;
- relative recall: the ratio of the relevant retrieved documents examined by the user to the number of documents the user would have liked to examine.

All these measures are based on the ranking of documents retrieved and on the statement that one knows which documents are relevant and which ones are not. Within a framework like TREC, this statement encounters some limits.

2.3 Limits of actual measures

The continuous growth of test collection makes it impossible to build a complete set of relevance judgements for each topic. Owing to the non-completeness of assessments, the final recall is biased. *Zobel* [18] showed that at best 50 to 70% of relevant documents are detected with the pooling method. As a result of the previous problem, non judged documents are considered as non relevant when evaluating IRS. This favors systems using strategies that are close to the participant systems, and penalizes new (or better) systems with different strategies, able to find new relevant documents .

If a topic corresponds to a lot of relevant documents, evaluating the system by comparing the number of retrieved relevant documents with the whole set of relevant documents is not judicious, because any way a basic user will not read all the documents . Moreover, it is unfair to penalize a system that missed relevant documents , because one can not say if a relevant but non-retrieved document is more relevant than any retrieved document ; usually one can only attribute a binary value. Precision and recall then have strong variations when the number of judged documents varies. Only a $P@N$ - precision after N first documents retrieved - curve corrects partly this problem. The Mean average precision is one of only to take into account the rank of relevant documents.

The RSV is used by IRS to rank documents. This means that evaluation is in fact based on RSV.

2.4 Previous use of RSVs

Joon Ho Lee [9] normalized each RSV by the maximum (and the minimum) RSV. To combine differents runs, he suggested to weight individual run depending on their overall performance and to chose the summation function (numerical mean of the set of RSVs). This function is one of the several functions for combining the RSVs tested by *Fox and Show* [4]. Through experiments, Lee [9] noticed significant improvements by combining two retrieval runs in

which one performs cosine normalization and the other does not. Following Lee [9], *Kamps et al.* [7] experimented with two approaches to morphological normalization at INEX 2002. To experiment with combinations of different kinds of run, they normalize the RSVs, since different runs may have radically different RSVs and mapped the values to $[0, 1]$ using $RSV_i' = (RSV_i - min_i)/(max_i - min_i)$ and they assigned a weight to the documents using a linear interpolation factor that represented the relative weight of a run. This combined run was better than the best underlying baserun in 3 of 4 cases. For their participation at TREC 2003 campaign in the Question/Answering Track, *Jijkoun et al.* [5] computed a score that is the sum of two weighted components: the normalized original RSV (global similarity $RSV_{normalized} = RSV/Max(RSV)$) and the spanning factor (local similarity).

Regarding IR evaluation measures, most of them are based on the ranking of documents retrieved. This ranking is based on the RSV given to each document by the IRS. Each IRS has a particular way to compute document documents RSV according to the IR model on which it is based. In the Boolean model, RSVs are either zero or one. Fuzzy retrieval allows for RSVs in the intervall $[0, 1]$. The vector-space model can be used with the cosine metric (RSVs in $[0, 1]$) or the scalar product (RSVs in $\Re$). So different IR methods compute RSVs in different ways and different scales. The actual relationship between the RSV of a document and its probability of relevance can be approximated by a function [11]

$$f : \Re \longmapsto [0, 1], f(RSV(d, q)) \approx Pr(rel|q, d)$$

Little effort has been spent on analyzing the relationship between RSVs and probability of relevance of documents. For ad-hoc retrieval, it is sufficient to rank documents according to their RSVs. However, advanced applications like filtering or distributed retrieval require estimates of the actual probability of relevance and RSVs are not sufficient (for example in resource selection, there's need for approximating the number of relevant documents in the result set or for merging the documents retrieved from the selected collections into a single ranked list). *Nottelman and Fuhr* [11] describe the relationship between the RSV of a document and its probability of relevance by a "normalization" function which maps the RSV onto the probability of relevance. They proposed linear and logistic mapping functions for different retrieval methods. In the rest of this section, we propose some measures to link RSV to IRS evaluation.

3 Proposed Measures

We will use the following notation in the rest of this paper:

- d_i : for a given topic, document retrieved at rank i by the system.
- $s_i(t)$: for a given topic t, it is the RSV that a system gives to the document d_i: $s_i(t) = RSV(d_i, t)$
- n : number of documents that are considered while evaluating the system.

- $P(t)$: set of relevant documents for the topic t among the n first documents.
- $\bar{P}(t)$: set of non relevant documents for the topic t among the n first documents.

We assume that:

1. all the scores $s_i(t)$ are positive ($\forall i\, s_i(t) >= 0$);
2. the retrieved documents are ranked by their RSVs;
3. documents are given a binary relevance judgement (0 or 1).

Some limits recalled above for existing measures may induce to propose the measure giving the proportion of relevant documents among the first n. This metric is already known as: $P@n = \frac{|P|}{n}$; and seems to fit well with the information need of a lambda user who wants as many documents as possible to be relevant among the ones that he or she will read (the ones at the top of results).

The RSVs are generally considered as meaningless system values; yet we guess that they have stronger and more interesting semantics than the simple rank of the document. Indeed, two documents that have close RSVs are supposed to have close probabilities of relevance. In the same way, two distant scores suggest a strong difference in the probability of relevance, even if the documents have consecutive or close ranks. But the semantics (inherent to the IRS) and the RSVs scale depend on the IRS's model and implementation. We already noticed that RSVs scale vary from an IRS to another. These different scales should not act on the evaluation. An absolute use of these RSVs would therefore not be equitable. On the other hand, the relative distances between RSVs attributed by the same system are very significant; if a system attributed RSVs between 0 and 100000, a distance of 1000 between two documents corresponds to close probabilities of (supposed) relevance. In order to free from the absolute differences between systems, we used a maximum normalization of the RSVs so that the biggest one always equals 1 (for a given system and a given topic: $s_1(t) = max_j\, s'_j(t)$). So:

- For a topic t,
$$\forall i\, s'_i(t) = \frac{s_i(t)}{s_1(t)}$$
- Thus, $\forall i\, s'_i(t) = NRSV(d_i, t)$, $s'_i(t) \in [0, 1]$ and $s'_i(t) < s'_{i+1}(t)$

For the topic t, $s'_i(t)$ represents an estimation by the system of the relative closeness of the document d_i to the document considered as the most relevant by the system (d_1) for topic t. This closeness is obviously the highest for the top ranked document d_1 and it should be the lowest for all non-retrieved documents (according to our hypothesis, $\forall i, s'_i > 0$ et we consider that $s_i = 0$ and $s'_i = 0$ for any non-retrieved document). We assume that a lower bound exists for the RSVs and is equal to 0. If it is not the case we need to know (or to calculate) a lower bound and to perform the min-max normalization for a topic t by the formula $\forall i, s'_i(t) = \frac{s_i(t) - min_j\, s'_{j(t)}}{s_1(t) - min_j\, s'_j(t)}$

We propose a first pair of symmetrical metrics, applicable to each topic; the figure r determines a success rate while e is a failure rate:

$$\begin{cases} r_1(n) = \frac{\sum\limits_{i=1..n} s'_i \times p_i}{n} \\ \\ e_1(n) = \frac{\sum\limits_{i=1..n} s'_i \times (1-p_i)}{n} \end{cases}$$

where p_i is the binary assessed relevance of document d_i:

$$\begin{cases} p_i = 1 \text{ si } d_i \in P \\ p_i = 0 \text{ si } d_i \in \bar{P} \end{cases}$$

$r_1(n)$ is the average normalized RSV (NRSV) considering only the relevant documents; e_1 is the average NRSV considering only non relevant documents . The second proposed pair of metrics is derived from r_1 and e_1:

$$\begin{cases} r_2(n) = \underbrace{\frac{\sum\limits_{i=1..n} (1-s'_i) \times (1-p_i)}{n}}_{r_{2,1}} + \frac{\sum\limits_{i=1..n} s'_i \times p_i}{n} \\ \\ e_2(n) = \underbrace{\frac{\sum\limits_{i=1..n} (1-s'_i) \times p_i}{n}}_{e_{2,1}} + \frac{\sum\limits_{i=1..n} s'_i \times (1-p_i)}{n} \end{cases}$$

$r_{2,1}(n)$ concerns non relevant documents, the average distance from the top ranked document d_1. This distance represents in a way the estimation by the system of the "risk" of non relevance for the document. That is why one can think fair to add it to the success rate.

$e_{2,1}(n)$ is symmetrical to $r_{2,1}(n)$. These metrics give a higher weight to documents that have high NRSVs: these documents take part in success rate (r_i if they are really relevant; on the other hand the system is more severely penalized (through the e_i measures) if they are not relevant. Conversely, a low score has little influence on the measures.

N.B.: the "standard" precision measure does not take RSVs in to account. It is equivalent to success rates r_1 and r_2 where s'_i is equal to 1 for all retrieved documents.

A new problem arises at this step, in relation with human assessment. If the assessor considers that a document is not relevant, it sounds fair to penalize the system according to s_i: the more the NRSV is high, the more the system should be penalized, since it is a sign of erroneous confidence.

But if the human judge assesses that the document is relevant, it seems difficult to evaluate the system according to s_i. Indeed the assessor cannot say *how much* the document is relevant (in the case of binary judgment[1]). One does not know if the confidence of the system was

[1] In the case of n-ary judgment for example graded relevance, the metrics should then be adapted.

Table 1: Kendall tau between IRS ranking

-	IPR at 0	IPR at 0.1	IPR at 0.2	IPR at 1	MAP	P@5	P@10	P@100	P@1000
r_1	0.92	0.83	0.80	0.87	0.81	0.90	0.83	0.64	0.53
e_1	-0.50	-0.06	0.18	0.59	0.52	-0.61	-0.43	-0.14	-0.11
r_2	0.31	0.29	0.31	0.46	0.55	0.71	0.50	0.15	0.20
e_2	-0.51	-0.064	0.20	0.59	0.59	-0.68	-0.47	-0.09	-0.09
r_3	0.31	0.31	0.33	0.46	0.54	0.64	0.46	0.18	0.21

justified, whether this confidence was strong (high NRSV) or not (low NRSV). Because of this, we can notice that if a system retrieves n relevant documents (out of n), the success rates r_1 and r_2 will be less than 1, which is unfair. Thus we propose a new measure:

$$r_3(n) = \frac{\sum_{i=1..n} p_i + \sum_{i=1..n} (1 - s_i)(1 - p_i)}{n}$$

Any relevant document retrieved contributes to this measure for 1, and an non relevant document contributes by its distance to the top ranked document, that is to say $(1 - s'_i)$ in order to take into account the fact that the system "assessed" through the RSV that the document relevance was uncertain.

Furthermore, we can observe that $r_3 = 1 - e_1$.

Measures r_1 and r_2 can be useful when comparing two IRS, because they favor the systems that give good RSVs to relevant documents (a same document differently weighted by two systems contributes to their evaluation in a different manner). On the other hand, r_3 may allow a more objective evaluation of a single system's performances.

4 Experiments

We used the results lists of IRS that took part in the TREC9 WebTrack campaign. 105 runs had been submitted to this track. We computed our measures for each of these result lists. We used a correlation based on Kendall's τ in order to compare our measures with classical IR evaluation measures. Kendall's tau computes the distance between two rankings and produces a result between -1.0 (perfect inverse) and 1 (equivalent rankings). We used various cut-off levels for precision and recall; here we show results for Interpolated recall at 0.0 and at 0.1 and for P@5 and P@10 and give comments for the others levels: IPR stands for Interpolated Precision at Recall level.

The ranking obtained with the measure r_1 which is based on the normalized RSV for relevant documents is highly correlated with precision on the first documents retrieved (P@N). This correlations decreases as N increases.

Conversely, the ranking obtained with the measure e_1 which is based on the normalized RSV for non relevant documents is inversely correlated with P@N and with IPR at first recall levels (this was excepted, since e_1 represents a failure rate).

The measures r_2 (resp. e_2) that combines normalized RSV for relevant documents (resp. for non relevant documents) with a value that expresses the distance between non relevant documents (resp. relevant documents) and the first document are less (resp. less inversely) correlated with $P@N$ and with IPR at first recall levels.

The measure r_3 that combines contribution from relevant documents retrieved (1) and contribution from non relevant documents retrieved (a value that expresses the way the IRS valuate the risk of mistaking when ranking this non relevant documents at a given position) is even less correlated with $P@N$ and with IPR at first recall levels.

Despite the fact that IR evaluation methods (the trec_eval for example) use the RSVs as a base to evaluate IR systems, the document RSVs are always ignored and considered as system values with no real signification. These correlations, and especially those between r_1 (resp. e_1) and $P@N$ show that there is a strong link between relevant (resp. non relevant) document rank and their RSVs.

5 Conclusion

The measures currently used in IR evaluation are mainly based on IRS outputs. For well established IR evaluation campaign like TREC, the document RSV is the main field of these outputs. Indeed it is used to rank the retrieved documents, to compute IR evaluation measures and the to compare IRS. Despite this central place, the RSV is still considered as a system value with no particular semantics. In this work, we propose IR measures directly based on normalized RSVs. Experiments on the TREC9 participant result lists show a high correlation between these measures and some classical IR evaluation measures. These correlations indicate possible semantics besides documents RSVs. The proposed measures are probably less intuitive than precision and recall but they put forth the question of the real place of RSV in IR evaluation.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval*. ACM Press, second edition, 1999.

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

[3] G.V. Cormack, C.R. Palmer, and C.L.A.Clarke. Efficient construction of large test collections. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 282–289, 1998.

[4] E.A. Fox and J.A. Shaw. Combination of multiple searches. In *Proceedings of the 2nd Text REtrieval Conference (TREC-2), National Institute of Standards and Technology Special Publication 500-215*, pages 243–252, 1994.

[5] V. Jijkoun, G. Mishne, C. Monz, M. de Rijke, S. Schlobach, and O. Tsur. The university of amsterdam at the trec 2003 question answering track.

[6] Y. Kagolovsk and J.R. Moehr. Current status of the evaluation in information retrieval. *Journal of medical systems*, 27(5):409–424, October 2003.

[7] J. Kamps, M. Marx, M. de Rijke, and B. Sigurbjörnsson. The importance of morphological normalization for xml retrieval. In *Proceedings of INEX'03*, pages 41–48, 2003.

[8] R. Korfhage. *Information storage and retrieval*. Wiley Computer publising, 1997.

[9] Joon Ho Lee. Combining multiple evidence from different properties of weighting schemes. In *Proceedings of SIGIR '95*, pages 180–188, 1995.

[10] S. Mizzaro. How many relevances in information retrieval? *Interacting with Computers*, 10(3):303–320, 1998.

[11] H. Nottelman and N. Fuhr. From retrieval status value to probabilities of relevance for advanced ir applications. *Information retrieval*, 6(3-4):363–388, 2003.

[12] V.V. Raghavan, G.S. Jung, and P. Bollman. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems*, 7(3):205–229, July 1989.

[13] A. M. Rees and D. G. Schulz. A field experimental approach to the study of relevance assessments in relation to document searching. 2 vols. Technical Report NSF Contract No. C-423, Center for Documentation and Communication Research, School of Library Science, Case Western Reserve University, Cleveland, Ohio, 1967.

[14] T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *JASIS*, 26:321–343, 1975.

[15] I. Soboroff, C. Nicholas, and P. Cahan. Ranking retrieval systems without relevance judgments. In *Proceedings of the 24th annual international ACM Conference on research and Development in Information Retrieval (ACM-SIGIR)*, pages 9–13, September 2001.

[16] E. Voorhees and D. Harman. Overview of the sixth text retrieval conference (trec-6). In *NIST Special Publication 500-420, The sixth text retrieval conference*, November 1997.

[17] P. Wilson. Situational relevance. *Information storage and retrieval*, 9(8):457–471, 1973.

[18] J. Zobel. How reliable are the results of large scale information retrieval experiments. In *Proceedings of ACM SIGIR'98*, pages 307–314, August 1998.