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, tannier}@emse.fr

Abstract. Retrieval systems rank documents according to their retrieval status values (RSV) if these are monotonously increasing with the probability of relevance of documents. In this work, we investigate the links between RSVs and IR system evaluation.

1 IR Evaluation and Relevance

Kagolovsk *et al* [1] realised a detailed survey of main IR works on evaluation. Relevance was always the main concept for IR Evaluation. Many works studied the relevance issue. Saracevic [2] proposed a framework for classifying the various notions of relevance. Some other works proposed some definitions and formalizations of relevance. All these works and many others suggest that there is no single relevance: relevance is a complex social and cognitive phenomenon [3].

Because of the collections growth nowadays, relevance judgements can not be complete and techniques like the pooling technique are used to collect a set of documents to be judged by human assessors. Some works investigated this technique, its limits and possible improvements [4].

To evaluate and classify IR systems, several measures have been proposed; most of them based on the ranking of documents retrieved by these systems, and ranking is based on monotonously decreasing RSVs. Precision and recall are the two most frequently used measures. But some others measures have been proposed (the Probability of Relevance, the Expected Precision, the E-measure and the Expected search length, etc). Korfhage [5] suggested a comparison between an IRS and a so-called ideal IRS. (the normalized recall and the normalized precision). Several user-oriented measures have been proposed (coverage ratio, novelty ratio, satisfaction, frustration).

2 IR Evaluation Measures and RSV

2.1 Previous Use of RSVs

Document ranking is based on the RSV given to each document by the IRS. Each IRS has a particular way to compute document RSV according to the IR

M. Consens and G. Navarro (Eds.): SPIRE 2005, LNCS 3772, pp. 225–228, 2005.
© Springer-Verlag Berlin Heidelberg 2005

model on which it is based (0 or 1 for the Boolean model, $[0, 1]$ for the fuzzy retrieval, $[0, 1]$) or $\Re$ for the vector-space,etc). Little effort has been spent on analyzing the relationship between RSV and probability of relevance of documents. This relationship is described by *Nottelman et al.* [6] by a "normalization" function which maps the RSV onto the probability of relevance (linear and logistic mapping functions).

Lee [7] used a min-max normalization of RSVs and combined different runs using numerical mean of the set of RSVs of each run. *Kamps et al.* [8] and *Jijkoun et al.* [9] also used normalized RSVs to combine different kinds of runs.

2.2 Proposed Measures

We will use the following notation in the rest of this paper: d_i is the document retrieved at rank i by the system; $s_i(t)$ is, for a given topic t, the RSV that a system gives to the document d_i. Finally n is the number of documents that are considered while evaluating the system.

We assume that all the scores are positive. Retrieved documents are ranked by their RSV and documents are given a binary relevance judgement (0 or 1).

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 RSV scale depends on the IRS model and implementation. Different RSV scales should not act on the evaluation. Nevertheless, the relative distances between RSVs attributed by the same system are very significant; In order to free from the absolute differences between systems, we use a maximum normalization:

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

$s'_i(t)$ gives 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. For d_1, $s'_1 = 1$, we consider that $s_i = 0$ and $s'_i = 0$ for any non-retrieved document. We assume that a lower bound exists for the RSV and is equal to 0. If it is not the case we need to know (or to calculate) a lower bound and to perform a min-max normalization.

We propose a first pair of metrics, applicable to each topic; the figure r determines a success rate while e is a failure rate (p_i is the binary assessed relevance of document d_i):

$$r_1(n) = \frac{\sum_{i=1..n} s'_i \times p_i}{n} \text{ and } e_1(n) = \frac{\sum_{i=1..n} s'_i \times (1 - p_i)}{n}$$

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

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

$r_{2,1}(n)$ is a distance representing the estimation by the system of the "risk" of non relevance for the document. $e_{2,1}(n)$ is equivalent to $r_{2,1}(n)$ for relevant document. Documents with high NRSVs have a high influence on these metrics by increasing r_i (if they are relevant) and by penalizing the system through e_i (if they are not relevant).

A new problem arises at this step, if a document d_i is assessed as 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). One does not know if the confidence of the system was justified, whether this confidence was strong (high NRSV) or not (low NRSV). We can also 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 a non relevant document contributes by its distance to the top ranked document.

Measures r_1 and r_2 can be useful when comparing two IRSs, because they favor systems that give good RSVs to relevant documents. On the other hand, r_3 may allow a more objective evaluation of a single system performances.

3 Experiments

We experimented on TREC9 WebTrack results (105 IRSs). We used a correlation based on Kendall's τ in order to compare our measures with classical IR evaluation measures. 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 e_1 which is based on the normalized RSVs 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 NRSVs for relevant documents (resp. for non relevant documents) with a value expressing the distance between

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

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 irrelevant documents retrieved (a value that expresses the way the IRS valuate the risk of mistaking when ranking this irrelevant documents at a given position) is even less correlated with $P@N$ and with IPR at first recall levels.

4 Conclusion

RSV is used to rank the retrieved documents. Despite this central place, it is still considered as a system value with no particular semantics. We proposed IR measures directly based on normalized RSVs. Experiments on the TREC9 results 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] Kagolovsk, Y., Moehr, J.: Current status of the evaluation in information retrieval. Journal of medical systems **27** (2003) 409–424
[2] Saracevic, T.: Relevance: A review of and a framework for the thinking on the notion in information science. JASIS **26** (1975) 321–343
[3] Mizzaro, S.: How many relevances in information retrieval? Interacting with Computers **10** (1998) 303–320
[4] Zobel, J.: How reliable are the results of large scale information retrieval experiments. In: Proceedings of ACM SIGIR'98. (1998) 307–314
[5] Korfhage, R.: Information storage and retrieval. Wiley Computer publising (1997)
[6] Nottelman, H., Fuhr, N.: From retrieval status value to probabilities of relevance for advanced ir applications. Information retrieval **6** (2003) 363–388
[7] Lee, J.H.: Combining multiple evidence from different properties of weighting schemes. In: Proceedings of SIGIR '95. (1995) 180–188
[8] Kamps, J., Marx, M., de Rijke, M., Sigurbjrnsson, B.: The importance of morphological normalization for xml retrieval. In: Proceedings of INEX'03. (2003) 41–48
[9] Jijkoun, V., Mishne, G., Monz, C., de Rijke, M., Schlobach, S., Tsur, O.: The university of amsterdam at the trec 2003 question answering track (2003)