Evaluating Scalability in Information Retrieval with Multigraded Relevance

Amélie Imafouo, Michel Beigbeder

Evaluating Scalability in Information Retrieval with Multigraded Relevance

Amélie Imafouo and Michel Beigbeder

Ecole Nationale Supérieure des Mines of Saint-Etienne
158 Cours Fauriel - 42023 Saint-Etienne, Cedex 2, France
{imafouo, beigbeder}@emse.fr

Abstract. For the user's point of view, in large environments, it can be desirable to have Information Retrieval Systems (IRS) that retrieve documents according to their relevance levels. Relevance levels have been studied in some previous Information Retrieval (IR) works while some others (few) IR research works tackled the questions of IRS effectiveness and collections size. These latter works used standard IR measures on collections of increasing size to analyze IRS effectiveness scalability. In this work, we bring together these two issues in IR (multigraded relevance and scalability) by designing some new metrics for evaluating the ability of IRS to rank documents according to their relevance levels when collection size increases.

1 Introduction

Nowadays, many factors support a growing production of information. A regular increase of 30% was noted between 1999 and 2002 in the information production [1]. The problem of accessing this mass of information comes under the field of domains like digital libraries and information retrieval but currently few works of these domains have taken into account the size effect in their approaches. The size of large collections (or web) coupled with and the ambiguity of user query make it difficult for search engines to return the most recent and relevant information in real-time. The need to learn more about they way collections size acts on retrieval effectiveness becomes increasingly pressing. In this work, we present works dealing with multigraded relevance and in the last part we present the metrics designed to evaluate the ability of IR systems to rank documents according to their relevance levels.

2 Multigraded Relevance Levels in IR

2.1 The Relevance as a Complex Cognitive and Multidimensional Concept

Relevance is the central concept for IR evaluation, usually considered as a binary notion. However, some research works showed that the relevance is a complex

cognitive concept, that has many dimensions ([2], [3] , [4], [5]. Many different aspects of relevance have also been discussed by proposed definitions and classifications ([4], [6], [7], [8]). It is not an easy job to judge documents and give them a relevance level regarding a topic as many variables affect the relevance (*Rees et al.* [9]: about 40 variables, *Cuadra et al.* [10]: 38 variables). All these works and many others suggest that there is no single relevance (there are many relevances) and that relevance is a complex cognitive problem.

2.2 Multigraded Relevance in IR

In the user's point of view, it is desirable to have IRS that retrieve documents according to their relevance level [11]. IR evaluation methods should then credit (or at least recognize) IRS for their ability to retrieve highly relevant documents at the top of their results list, by taking into account various relevance levels of a document for a given query; they have been studied in some previous IR works (*Tang et al.* [12]: a seven-point scale, *Spink et al.* [13] used a three-point scale). Some test collections provide multigraded relevance assessments (TREC Web Track collection: three point scale [14], *INEX* collections: a multilevel scale, *NTCIR* evaluation campaign [15]). *Kekäläinen et al.* [11] used a four-points scale for relevance level : *highly relevant, fairly relevant, marginally relevant, not relevant*. Each of these relevance level has to be expressed by a numerical value for computing measures. One of the remaining question is the choice of these values and the semantic they should have. Their work also proposed *generalized non binary recall and precision*, that are extensions of standard binary recall and precision taking into account multiple relevance levels [11]. The *Discounted Cumulated Gain* and the *Cumulated Gain* are also proposed by the same authors in [16]. We present them using our formalism in section 3.2. *Sakai* [17] also proposed a measure based on of the *Cumulated Gain*.

Our conceptions meet those of *Kekäläinen et al.* [16] concerning the fact that multiple relevance levels should be taken into account when evaluating IRS. While information grows continuously, for the users lambda, one of the main issue for IRS will become to retrieve documents with highly relevance level at the top of the results list. We design metrics to allow the evaluation of this ability in IRS as collections size increase.

3 Protocols for Scalability Evaluation with Multigraded Relevance

Let C_1 and C_2 be two collections of different sizes such as $C_1 \subset C_2$ and S an IRS. The aim is to analyze how S behaves on each collection to determine if its effectiveness improves, remains the same or decreases when the collection size increases. Our measures are based on the comparisons of the relevance levels of the first documents in the results lists for the two collections.

3.1 Relevance Level *Importance*

For a given topic, we assume that a relevance level is given to every document regarding this topic. Let $\{rel_i\}$, $i = 1, \ldots, n$ be the set of possible relevance levels; two documents are *equivalent* if they have the same relevance level regarding this topic.

We define a total order relation on the set of the relevance levels noted $\succ$: $rel_i \succ rel_j$ when $i > j$. This total order relation gives the preference wished on retrieved documents but it gives no indication about the *importance* of a particular relevance level regarding the other relevance levels. However, it is the *importance* of a relevance level that characterizes the quality/quantity of information expected from a document of this relevance level. One may need to highly credit (resp penalize) retrieval systems that return the documents with the highest relevance level at the top (resp not at the top) of their results list: in this case, the highest relevance level must have a high *importance* (compared to the *importance* of the other relevance levels) when evaluating retrieval results. On the other side, some applications need to retain many documents of good relevance levels, the difference between a document of good relevance level and a document of high relevance level is not important. Thus, a function I that models the *importance* of relevance levels depends of the types of applications the IRS is designed for and is characterized by the following properties (a positive and increasing function):

- $I(rel_i) > 0$
- $I(rel_i) > I(rel_j)$ if $rel_i \succ rel_j$ i.e. $i > j$

The choice of the number of relevance levels and the attribution of numerical values to relevance levels is still an open problem in IR. *Kekäläinen* [18] used different empirical weighting schemes (see figure 1). Giving a numerical values of *importance* to relevance levels means nothing in the absolute; but in the relation with others relevance levels, these values can be associated to a semantics as we show it through the *gain function* (section 3.3).

	Highly relevant (HR)	Fairly Relevant (FR)	Marginally Relevant (MR)	Not Relevant (NR)
scheme 1	1	1	1	0
scheme 2	3	2	1	0
scheme 3	10	5	1	0
scheme 4	100	10	1	0

Fig. 1. Four schemes for assigning numerical values to relevance levels [18]. These values give the *importance* of the relevance level for us.

3.2 Cumulated Gain at a Given Rank

The *Cumulated Gain, CG* as proposed by *Kekäläinen et al.* [16], is computed at rank r by the sum of relevance levels of documents retrieved at any rank $k \leq r$:

$$\begin{cases} CG(1) = I(RelLevel(d_1) \\ CG(i) = CG(i-1) + I(RelLevel(d_i) \end{cases}$$

The *Discounted Cumulated Gain(DCG)* also computes relevance gains with a discount factor which is a decreasing function of the rank: the greater the rank, the smaller share of the document relevance level is added to the cumulated gain. This factor is needed to reduce progressively the impact of the gain of relevant information according to the rank (steep reduction with a function like the inverse of the rank $disf(k) = 1/k$ if the first documents are those we want to focalize on during the evaluation or less steeply with a function like the inverse of the log of the rank $disf(k) = 1/log_b(k)$ as in [16]). By averaging over a set of queries, the average performance of a particular IR method can be analyzed. Averaged vectors have the same length as the individual ones and each component i gives the average of the ith component in the individual vectors. The averaged vectors can directly be visualized as gain-by-rank graphs. The actual CG and DCG vectors are also compared to the best theoretically possible. We described the building of the best theoretically results list in section 3.4, as we re-use it for our metrics.

3.3 Information Gain Between Two Relevance Levels

For a given topic, in front of two documents with two different relevance levels, the same amount of relevant information is not expected from the two documents. It is interesting to quantify the relevant information gained (or lost) when moving from a relevance level to another, that is a function of the relevance levels: $Gain(rel_i, rel_j) = g(rel_i, rel_j)$, with these characteristics:

- $g(rel_i, rel_j) > g(rel_i, rel_k)$ if $rel_j \succ rel_k$ i.e. if $j > k$
- $g(rel_i, rel_j) < g(rel_k, rel_j)$ if $rel_i \succ rel_k$ i.e. if $i > k$
- $g(rel_i, rel_i) = 0$. There is neither a gain nor a loss of information when one stays on the same relevance level (even if one change the document because the documents of the same relevance level for a given topic are in the same equivalence class).

By deduction, we have: $g(rel_i, rel_j) < 0$ if $rel_i \succ rel_j$ i.e. if $i > j$.

Indeed if we have $rel_i \succ rel_j$, then this means that the quantity of relevant information contained in the document of relevance level rel_i is higher than the quantity of relevant information contained in a document of relevance level rel_j. Thus, when moving from a document of relevance level rel_i to a document of relevance level rel_j, one loses relevant information and so $g(rel_i, rel_j) < 0$.

In the same way, $g(rel_i, rel_j) > 0$ if $rel_j \succ rel_i$ i.e. if $i < j$

It is obvious that the gain function between two relevance levels depends on the *importance* associated to each of the relevance levels.

Thus $g(rel_i, rel_j) = h(I(rel_i), I(rel_j))$. An example of an h function is modelled by the mathematical distance (we can build a *distance* between different relevant levels, using their associated numerical value of *importance*).

For example (simple case) $d(rel_i, rel_j) = d(I(rel_i), I(rel_j))$. So we can choose:

$$\begin{cases} g(rel_i, rel_j) = -d(I(rel_i), I(rel_j)) \, if \, rel_i \succ rel_j \\ g(rel_i, rel_j) = d(I(rel_i), I(rel_j)) \, else \end{cases}$$

We respect all the properties of the function g.

3.4 Information Gain at a Rank When Collection Size Increases

We assume in this study that the measures proposed will be used to evaluate the effectiveness of a system on a collection that grows (from a first collection C_1 to a second collection C_2 with $C_1 \subseteq C_2$). When a collection C_1 grows by the addition of new documents and becomes a collection C_2, our assumption is that the effectiveness of a good retrieval system should at the worst case stay the same (from C_1 to C_2), whatever be (the relevance level of) the documents added. This effectiveness should not decrease, whatever be the documents added, as all the documents in C_2 were already in the collection C_1. And when new relevant documents are added, a good retrieval system effectiveness should stay the same or increase from C_1 to C_2.

For a given topic t, $d_k^t(C)$ is the document retrieved at rank k when the collection C is queried using the topic t. the information gain at rank k between the results lists of both collections is computed using the gain function as follows: $Move_k^t(C_1, C_2) = g(RelLevel(d_k^t(C_1)), RelLevel(d_k^t(C_2)))$

This Move expresses the relevant information gain (resp loss) at rank k when moving from C_1 results list for the topic t to C_2 results list for the topic t. We obtain a vector of weighted Moves by applying a discount factor $< \; disf(1) \times Move_1^t(C_1, C_2), \ldots, disf(N) \times Move_N^t(C_1, C_2) \; >$

Measure Type 1. There are two possibilities for using these vectors:

- For a given cut-off level N, either we sum the vectors' elements topic by topic to have a unique value for each topic. Thus we define the first metric as follows[1]:

$$Measure1_N^t(C_1, C_2) = \sum_{k=1}^{N} disf(k) \times Move_k^t(C_1, C_2)$$

- either we sum the weighted Moves rank by rank for all the topics and we obtain a single vector of N elements:

$$< disf(1) \times \textstyle\sum_t (Move_1^t(C_1, C_2)), \ldots, disf(N) \times \sum_t (Move_N^t(C_1, C_2)) >$$

  This sum-vector has the same size as vectors of weighted Moves for each topic; we can then visualize the vector as a gain/loss versus rank graph.

[1] For two collections C_i and C_j, this measure can only be computed on the set of topics t for which the IRS S provides a results list of N or more documents for both collections.

Thus, by querying an IRS on a set of collection $\{C_i\}$ such as $C_i \subset C_{i+1}$, we obtain information gains realized when collection size increases, and we can analyze the impact of collection size on the information gain. According to our assumptions, the measure $Measure1_N^t(C_1, C_2)$ should not be negative for a good retrieval system, as $C_1 \subset C_2$.

Measure Type 2. For a given collection C, the IRS S provide a result list $Retrieved^t(C)$ for a given topic t. Then we build a results list $Retrieved\_ideal_N^t(C)$ so-called *ideal* for this topic in the same way as [16].

Example: Consider $HR \succ FR \succ MR \succ NR$ the relevance levels of [16] (see table 1 and a topic t with 7 documents HR, 10 documents FR, 20 documents MR. We choose $N = 30$. The *ideal* results list of size 30 for topic t is as follows:

$$\underbrace{HR, \ldots, HR}_{7times}, \underbrace{FR, \ldots, FR}_{10times}, \underbrace{SR, \ldots, SR}_{13times}$$

We can now build the weighted vectors of Moves between the results list for the collection C and the *ideal* results list:
$< disf(1) \times Move_1^t(C, Ideal\_C), \ldots, disf(N) \times Move_N^t(C, Ideal\_C)$.

As for the previous case, we have two possibilities of using these vectors for evaluation:

- At the cut-off level N and for the topic t we compute :

  $$Measure2_N^t(C) = Measure1_N^t(Retrieved^t(C), Retrieved\_ideal_N^t(C))$$

  This measure expresses the information gain when moving from the collection C results list to an *ideal* results list. While the collection C size increases, we can then analyze the variation of its results list compared to an ideal results list.
- we sum the weighted vectors elements rank by rank for all the topics and we obtain a single vector of N elements; we can then visualize the vectors as a gain-versus-rank graph.

  $< disf(1) \times \sum_t (Move_1^t(C, Ideal\_C)), \ldots, disf(N) \times \sum_t (Move_N^t(C, Ideal\_C)) >$

4 Discussions and Conclusions

In this work, we propose some metrics based on the notion of multigraded relevance levels for evaluating the way IRS scale. Their goal is to provide some information on the coherence between the ranking of documents retrieved by an IRS and the relevance levels of these documents as collection size increases. Some recent metrics in IR used a notion of relevance with multiple levels, e.g. the *Discounted Cumulated Gain* or the *Cumulated Gain*. For a given collection and an IRS, these metrics compute the relevant information gain obtained as one

goes through the results list returned by an IRS on a given collection. Our metrics compute the relevant information gain obtained when a single IRS is used on a collection which grows. Thus we evaluate the ability of the IRS to rank the documents according to their relevance levels when collection size grows. All the metrics that use multigraded relevance need to associate a numerical value to each relevance level and this is still not well studied in IR: in this study, we formalize the (obvious) constraints linked to the attribution of numerical values to relevance levels through the *importance function* and the *gain function*.

We are now working on the relation between our metrics and the existing metrics that used multigraded relevance levels through some experiments.

References

1. Lyman, P., Varian, H.R., Swearingen, K., Charles, P., Good, N., Jordan, L.L., Pal, J.: How much informations 2003. http://www.sims.berkeley.edu/research/projects/how-much-info-2003/ (2003)
2. Mizzaro, S.: How many relevances in information retrieval? Interacting with Computers **10** (1998) 303–320
3. Barry, C.L.: User-de.ned relevance criteria: an exploratory study. Journal of the American Society for Information Science **45** (1994) 149Ũ159
4. Saracevic, T.: 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** (1975) 321–343
5. Schamber L., E.M.B., Nilan, M.S.: A re-examination of relevance: toward a dynamic, situational definition. Information Processing and Management **26** (1990) 755Ũ776
6. Wilson, P.: Situational relevance. Information Storage and Retrieval **9** (1973) 457–471
7. Cooper, W.S.: A definition of relevance for information retrieval. Information Storage and Retrieval (1971)
8. Cosijn, E., Ingwersen, P.: Dimensions of relevance. Information Processing and Management **36** (2000) 533Ũ550
9. Rees, A.M., Schulz, D.G.: 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 (1967)
10. Cuadra, C.A., Katter, R.V.: The relevance of relevance assessment. In: Proceedings of the American Documentation Institute. Volume 4., American Documentation Institute, Washington, DC (1967) 95–99
11. Kekäläinen, J., Järvelin, K.: Using graded relevance assessments in ir evaluation. Journal of the American Society for Information Science and Technology **53** (2002) 1120–1129
12. Tang, R., William M. Shaw, J., Vevea, J.L.: Towards the identification of the optimal number of relevance categories. Journal of the American Society for Information Science **50** (1999) 254–264
13. Spink, A., Greisdorf, H., Bateman, J.: From highly relevant to not relevant: examining different regions of relevance. Information Processing and Management: an International Journal **34** (1998) 599–621

14. Voorhees, E.M.: Evaluation by highly relevant documents. In: Proceedings of the 24th annual international ACM SIGIR Conference. (2001) 74–82
15. : Ntcir workshop 1: Proceedings of the first ntcir workshop on retrieval in japanese text retrieval and term recognition, tokyo, japan. In Kando, N., Nozue, T., eds.: NTCIR. (1999)
16. Järvelin, K., Kekäläinen, J.: Ir evaluation methods for retrieving highly relevant documents. In: Proceedings of the 23th annual international ACM SIGIR Conference. (2000) 41–48
17. Sakai, T.: Average gain ratio: A simple retrieval performance measure for evaluation with multiple relevance levels. In: SIGIR'03. (2003)
18. Kekäläinen, J.: Binary and graded relevance in ir evaluations -comparison of the effects on rankings of ir systems. Information Processing an Management **41** (2005)