A Classification of IR Effectiveness Metrics

Gianluca Demartini and Stefano Mizzaro

Dept. of Mathematics and Computer Science,
University of Udine, Udine, Italy
{demartin, mizzaro}@dimi.uniud.it

Abstract. Effectiveness is a primary concern in the information retrieval (IR) field. Various metrics for IR effectiveness have been proposed in the past; we take into account all the 44 metrics we are aware of, classifying them into a two-dimensional grid. The classification is based on the notions of *relevance*, i.e., if (or how much) a document is relevant, and *retrieval*, i.e., if (how much) a document is retrieved. To our knowledge, no similar classification has been proposed so far.

1 Introduction

Evaluation is an important issue in Information Retrieval (IR). Evaluation initiatives (Cranfield, TREC, CLEF, NTCIR, INEX) have a strong tradition, and user studies experiments are frequently performed. Whatever the approach (test collection or user study), the effectiveness metrics chosen are crucial. We are aware of 44 metrics proposed so far. We propose a novel classification of all of them, based on the notions of *relevance*, i.e., if (or how much) a document is relevant, and *retrieval*, i.e., if (how much) a document is retrieved. The simple and traditional approach is based on the *binary* relevance and retrieval assumptions: either a document is relevant or not, and either a document is retrieved or not. By relaxing these two assumptions, one can speak of: *ranking* relevance and/or retrieval (a document is more relevant/retrieved than another), and of *continuous* relevance and/or retrieval (the value of relevance/retrieval is a real number on a continuum, measuring the amount of relevance/retrieval). Combinations, like binary relevance and ranking retrieval are possible, and indeed frequent.

2 IR Metrics: A Survey and a Classification

Table 1 shows the (approximated) year in which each metric has been made public, the metric name, a bibliographic reference, the category(ies) to which it belongs (•), the category(ies) to which it can belong with straightforward extensions (○), and in which evaluation initiatives it is used (×). We take into account also the metrics used in INEX 05, made public a few weeks ago. For space limitations, being most of the metrics described in well known textbooks [19, 18, 13], we briefly recall only the following, less common, ones:[1]

[1] N is the set of documents in the database; R is the set of relevant documents; r is the set of retrieved documents; $\overline{x}$ is the complement of x; $|x|$ is the cardinality of x.

M. Lalmas et al. (Eds.): ECIR 2006, LNCS 3936, pp. 488–491, 2006.

Table 1. A classification of IR effectiveness metrics (sorted by year)

Year	Relevance: / Retrieval:	Binary B	Binary R	Binary C	Rank B	Rank R	Rank C	Cont. B	Cont. R	Cont. C	TREC	INEX	NTCIR
1960	Precision [19]	•											
	Recall [19]	•											
	Fallout [19]	•											
	Generality Factor [19]	•											
1965	E-Measure F-measure	•											
	R/P curve		•								×	×	×
	R/fallout curve		•										
	Normalized Recall [13]		•										
	Normalized Precision [13]		•										
	Expected Search Length [5]		•										
	Sliding Ratio [13]								•				
1970	Novelty Ratio [13]	•											
	Coverage Ratio [13]	•											
	Relative Recall [13]	•											
	Recall effort [13]	•											
	Utility [18]	•											
1975	MAP		•								×	×	×
	P@N		•								×	×	×
	R-Precision		•								×		×
	Interpolated MAP		•										
1990	Satisfaction [13]								•				
	Frustration [13]								•				
	Total [13]								•				
	Usefulness measure [8]					•							
	Average Search Length [14]		•										
1995	NDPM [20]					•							
	Ranked Half Life [2]									•			
	Relative Relevance [2]					•							
2000	Classification accuracy [1]	•											
	DCG [9]		•			•			○				
	AWP [10]		•			•							
	Weighted R-Precision [10]		•			•							
	ADM [7]	•	•	•	•	•	•	•	•	•			
	XCG [11]		•			•			○			×	
	bpref [4]		•										
	Q-measure [17]		•			•						×	
	R-measure [17]		•			•						×	
	Tolerance to Irrelevance [6]		•									×	
	Estimated Ratio of Relevant [16]		•									×	
	Kendall, Spearman [3]					•							
	Normalized xCG [12]		•			•						×	
	Mean average nxCG at rank n [12]		•			•						×	
	Effort-precision/gain-recall @ std. gain-recall p. [12]		•			•						×	
	Non-interpolated mean average effort-precision [12]		•			•						×	
	Interpolated mean average effort-precision [12]		•			•						×	

– *R/fallout curve*: a plot of the recall values corresponding to the fallout values.
– *Expected Search Length*: average number of documents which must be examined before the total number of relevant documents is reached.
– *Sliding Ratio*: sum of the relevance judgments of the documents retrieved so far divided by the sum of the relevance judgments of the documents the ideal system would have retrieved so far.
– *Novelty Ratio*: percentage of the relevant retrieved documents which were previously unknown to the user.
– *Coverage Ratio*: percentage of relevant and known documents which are retrieved.

- *Relative Recall* (aka *sought recall*): percentage of the documents the user would have liked to examine which are relevant, retrieved, and examined.
- *Recall effort*: ratio of desired to examined by the user documents.
- *Satisfaction* (and *Frustration*): sliding ratio on documents in R ($\overline{R}$) only.
- *Total*: weighted mean of satisfaction and frustration.
- *Usefulness measure*: which of two IR systems delivers more useful information to the user.
- *Average Search Length*: average number of documents examined moving down in a ranked list before the average position of a relevant document is reached.
- *NDPM*: normalized distance between user and system ranking of documents.
- *Ranked Half Life*: degree to which relevant documents are located on the top of a ranked retrieval result.
- *Relative Relevance*: degree of agreement between the types of relevance applied in a non-binary assessment context.
- *Classification Accuracy*: if the classification is correct $((|r \cap R| + |\overline{r} \cap \overline{R}|)/|N|)$.
- *Average Weighted Precision (AWP)*: based on Cumulative Gain (CG), but more statistically reliable since it performs comparison with an ideal ranked output before averaging across topics.
- *Weighted R-Precision*: an extension of R-Precision.
- *Average Distance Measure (ADM)*: average difference between the relevance amount of documents and their estimates by the IR system.
- *eXtended Cumulative Gain (XCG)*: extends DCG-based metrics via the definition of a set of relevance value functions modeling different user behaviors.
- *bpref*: the average number of nonrelevant documents before a relevant document in the ranking, using the documents in the pool only.
- *Q-measure*: is based on CG, but it is better than AWP because it imposes a penalty for going down the ranked list.
- *R-measure*: is based on CG and it is the counterpart of Q-measure for R-Weighted Precision.
- *Tolerance to Irrelevance (t2i)*: maximum time that the user would keep reading nonrelevant documents before she proceeds to the next result.
- *Estimated Ratio of Relevant*: expectation of the number of relevant documents a user sees in the list of the first k returned documents, divided by the number of documents a user would see in the collection.
- *Kendall, Spearman*: statistical correlation between the ranked retrieval result and the user ranking of the documents.
- *Normalized xCG*: reflects the relative gain the user accumulated up to that rank, compared to the gain she could have attained if the system would have produced the optimum best ranking.
- *Mean average nxCG at rank n*: the average of nxCG[i] values for i=1 to n.
- *Effort-precision/gain-recall at standard gain-recall points*: the amount of relative effort (where effort is measured in terms of number of visited ranks) that the user is required to spend when scanning a systems result ranking compared to the effort an ideal ranking would take in order to reach a given level of gain (relative to the total gain that can be obtained).
- *Non-interpolated (Interpolated) mean average effort-precision*: the average of effort-precision values at each natural (standard) gain-recall point.

3 Conclusions and Future Work

The evolution over time shows that: (i) INEX initiative has caused a steep increase in the number of metrics; and (ii) the earlier metrics are usually classified under binary relevance and retrieval, and more recent metrics are often rank- or continuous-based, thus reflecting the changes in the underlying notion of relevance [15]. We hope that this classification will be useful for IR researchers, and will enable them to choose more consciously the most appropriate metrics for their purpose.

References

1. R. Belew. *Finding Out About*. Cambridge Univ. Press, 2000.
2. P. Borlund and P. Ingwersen. Measures of relative relevance and ranked half-life: Performance indicators for interactive IR. In *21st SIGIR*, pages 324–331, 1998.
3. R. Brache. Personal communication, 2005.
4. C. Buckley and E. Voorhees. Retrieval evaluation with incomplete information. In *27th SIGIR*, pages 25–32, 2004.
5. W. S. Cooper. Expected search length: A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. *JASIST*, 19:30–41, 1968.
6. A. de Vries, G. Kazai, and M. Lalmas. Tolerance to irrelevance: A user-effort oriented evaluation of retrieval systems without predefined retrieval unit. In *RIAO 2004 Conference Proceedings*, pages 463–473, 2004.
7. V. Della Mea and S. Mizzaro. Measuring retrieval effectiveness: A new proposal and a first experimental validation. *JASIST*, 55(6):530–543, 2004.
8. H. Frei and P. Schauble. Determining the effectiveness of retrieval algorithms. *IPM*, 27(2):153–164, 1991.
9. K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *TOIS*, 20:422–446, 2002.
10. N. Kando, K. Kuriyama, and M. Yoshioka. Information retrieval system evaluation using multi-grade relevance judgments. In *IPSJ SIGNotes*, 2001.
11. G. Kazai. Report of the INEX 2003 metrics working group. In *Proceedings of the 2nd INEX Workshop*, pages 184–190, 2004.
12. G. Kazai and M. Lalmas. INEX 2005 evaluation metrics. http://inex.is.informatik.uni-duisburg.de/2005/inex-2005-metricsv4.pdf.
13. R. R. Korfhage. *Information Storage and Retrieval*. John Wiley & Sons, 1997.
14. R. M. Losee. Upper bounds for retrieval performance and their use measuring performance and generating optimal boolean queries: Can it get any better than this? *IPM*, 30(2):193–204, 1994.
15. S. Mizzaro. Relevance: The whole history. *JASIS*, 48(9):810–832, 1997.
16. B. Piwowarski and P. Gallinari. Expected ratio of relevant units: A measure for structured information retrieval. In *INEX'03 proceedings*, pages 158–166, 2004.
17. T. Sakai. New performance metrics based on multigrade relevance: Their application to question answering. In *NTCIR 4 Meeting Working Notes*, 2004.
18. G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1984.
19. C. J. van Rijsbergen. *Information Retrieval*. Butterworths, 2nd edition, 1979.
20. Y. Y. Yao. Measuring retrieval effectiveness based on user preference of documents. *JASIS*, 46(2):133–145, 1995.