IJCSC 0973-7391 Volume 7 • Number 2 March 2016 - Sept 2016 pp. 85-89 available online at www.csjournalss.com

Survey of graded relevance metrics for information retrieval

Jaladhi Vyas
Department of computer science and engineering, Nirma University, Ahmedabad, India
jaladhi.vyas@nirmauni.ac.in

Abstract: A large number of metrics are available to evaluate the quality of rank web pages in information retrieval (IR). These metrics can be classified in different groups as follows: Binary Relevance, Graded Relevance, Rank Correlation Coefficient, and User Oriented Measures. Each group of metrics has difference characteristics. However, metrics that contains in the same group have the similar characteristics and uses. In this paper, I have discussed various types of metrics, used for the graded relevance measure.
Keywords: Discounted Cumulative Gain, Graded relevance, Information Retrieval, Rank Correlation Coefficient, User oriented measures

Introduction

One of the important issues in Information Retrieval is an evaluation of IR system because it measures the effectiveness of system by considering users' information needs. To deal with the problem of misinterpretation of the same result by different users, some metrics have been defined which correlates with the preferences of a group of users. Note that, there is no standard metric for evaluation of all tracks, so most of the tracks use basic metrics such as recall, precision and average precision as their base metric (i.e. use a combination of these) to form a new metric suitable for a particular track. CLEF, Cranfield, CLEF, TREC, INEX and NTCIR and many other evaluation initiatives have a strong tradition and they regularly perform experiments of user studies.

One of the open research problem is how to evaluate search engines effectively. The majority of the web search engines uses metrics which are based on cumulative gain e.g. Discounted Cumulative Gain (DCG) are heavily used in the majority of the web search engines. The significance of information retrieval (IR) evaluation based on graded relevance has started to receive attention, after the decades of binary relevance based TREC evaluation.

Metrics based on graded relevance

It may be sometimes difficult to specify the degree of relevance of the retrieved document as just either relevant or not, thus we can use Graded Relevance Based metrics. Metrics that I am going to discuss under this group are DCG, NDCG, ERR, NSDCG, RBP, Q-Measure and R-measure.

A. Discounted Cumulative Gain (DCG)

Binary relevance metrics are unable to differentiate between highly relevant documents and mildly relevant documents. This is one of the major drawbacks of it. The discounted cumulated gain (DCG) is a metric address the above problem effectively.

The method of calculating DCG metric is divided into 3 parts: First step is to compute the gain vector for a particular query by using a graded relevance score up to some particular rank position, e.g. If we want to calculate DCG for first 10 documents in ranking, then the length of the gain vector is 10. Secondly, compute the cumulative gain vector by adding all previously graded relevance score to current position. Finally, divide a cumulative gain vector by a discounting factor based on rank position to reduce the impact of gain as one further explore the rank list.

After doing this we can get DCG vector for a particular query. Given the gain vector Gj for a test query qj, the vector DCGj is given by

$$DCG_j[i] = \begin{cases} G_j[1] & \texttt{if } i = 1; \\ \frac{G_j[i]}{\log_2 i} + |DCG_j[i-1] & \texttt{otherwise} \end{cases}$$

To compare DCG with some ideal condition, we calculate IDCG vector. Here, NDCG is the ratio of DCG and IDCG.

One major issue with DCG is, it uses an assumption that, the usefulness of a document at rank i and the documents at rank less than i is independent. For example, one wants to calculate relevance score of document at rank position 2. If the documents in position 1 are relevant, it is possible that this document will be observed less and

due to this have few clicks. Conversely, if the first document is not too relevant, then document at rank position 2 is more possible to be examined and receive many clicks. Therefore, It is not possible to model the above two cases through a click model which depend only on the position. Thus, position models, fail to explain such a strong click through rate (CTR) difference. A real example, taken from the click logs of a commercial search engine, is shown below.

Ranking 1		Ranking 2	
URL	CTR	URL	CTR
uk.myspace.com	0.97	www.myspace.com	0.97
www.myspace.com	0.11		

Here, one can observe that, same URL link has a higher click through rate. The larger difference indicates that, the user may not even explore to position 2 due to the excellent match of URL in position 1. One solution to solve the problem of position based model is to combine it with click based model which is called cascaded model. The dependency among URLs is taken into account by this model by assuming that the user views search results from top to bottom and the user has a fixed probability of being satisfied at each position.

B. Expected Reciprocal Rank (ERR)

This metric is used to model user persistence in finding relevant document. The formula to calculate ERR is as below.

$$ERR := \sum_{r=1}^{n} \frac{1}{r} P(\text{user stops at position } r),$$

Here, n indicates total number of documents. The formula to calculate ERR is given below.

$$ERR := \sum_{r=1}^{n} \frac{1}{r} \prod_{i=1}^{r-1} (1 - R_i) R_r.$$

For a given set of Ri , probability P that the user is satisfied and stops at position r is given by:

$$\prod_{i=1}^{r-1} (1 - R_i) R_r$$

The advantage of ERR is over DCG, MAP is that, it greatly reduces the contribution of document that appear after highly relevant one. The example shows the evaluation method of the above metric. Suppose we have a relevance scale of documents between 0 to 4, where 4 indicates highly relevant and 0 indicates non relevant document. The graded relevance score for first three documents are as follows 3, 2 and 4. By converting them to corresponding probability values we get (2^3-1/16), (2^2-1) /16, (2^4-1) /16. The table given below is used to calculate ERR.

Table I. Calculation of ERR for three different rank values

K	1/rank	Grad	P(satisfy at doc k)	P(stop at doc k)
1	1/1	3	7/16	7/16
2	1/2	2	3/16	3/16*(1-7/16)
3	1/3	4	15/16	15/16*(1-3/16)*(1-7/16)

Here, ERR=1*7/16+1/2*3/16*(17/16)+1/3*15/16*(1-3/16)*(1-7/16)= 0.63

C. Rank Biased Precision (RBP)

This metric is used to model user persistence in finding relevant document. The formula to calculate RBP is given below.

$$\mathrm{RBP}(R,p) = (1-p)\sum_{i=1}^{|R|} r_i p^{i-1}$$

Here, p is a value between 0 and 1 that shows the user's searching persistence, R indicates the input relevance vector which requires to be calculated, and the relevance of the document in a position i is shown by ri. A person is less patient if value of p is less and more patient if the value of p is more. Example: if p=0 then user looks only the first document and if p=1 then user read all documents one after another. This user model assumes that, the user reads the documents from top to bottom and the drawback of this metric is, the requirement of the designer for any experiment to choose the value of p.

D. Normalized Session Based DCG (NSDCG)

In all previously discussed metrics one assumption is used, that is, there is only one query per session. Evaluation metrics that assumes one query per session are not sufficient when searchers reformulation efforts matter. NSDCG metric is used in the following conditions:

- Needs of information may not be pre-defined
- Initial query formulation may not be optimal
- Highly relevant documents are desired
- He /She may learn from session progress

Session Based DCG (SDCG) is the most useful metric in 2010 Session Track. It includes sequences of a query as an additional dimension for evaluation and allow for further discount relevance document found only after additional search efforts. The session DCG (sDCG) for a given position q, query is defined as

$$sDCG@k(q) = \frac{1}{(1+log_{bq}(q))} * DCG@k(q)$$

Here, DCG@k (q) is originally DCG value at Kth position of q^{th} query, where bq is a discounting factor based on logarithm. The formula to calculate NSDCG is as below.

NSDCG (q) = (SDCG (q))/ (ISDCG (q))

E. Q-Measure and R-Measure

Q-measure and R-measure are similar to cumulative gain and average weighted precision (originally called Weighted Average Precision). However, they are more reliable than average precision. The Q-measure can be used to evaluate question-answers that involve ranked list of exact answers. Q-measure and R-measure are used as IR metrics with the NTCIR-4 CLIR test collections, while the Q- measure is used as a QA metric with the NTCIR-4 QAC2 test collection. Formulas to calculate q measure and r measures are as below:

$$cg(r) = \sum_{1\leq i\leq r} g(i) \; count(r) = \sum_{1\leq i\leq r} isrel(i)$$

$$R\text{-}measure = \frac{\beta cg(R) + count(R)}{\beta cig(R) + R}$$

If β =1 then,

$$BR(r) = \frac{cg(r) + count(r)}{cg_I(r) + r}$$

$$Q\text{-}measure = \frac{1}{R}\sum_{1\leq r\leq L} isrel(r)BR(r)\ .$$

Comparison of metrics based on top heaviness

The data use for comparison are based on TREC03 and TREC04 (robust track).

	TREC03	TREC04
#Topics	50	49
#Documents	Approx. 529,000	
Pool depth	125	100
Average N	925.5	654.6
Range N	[292, 2050]	[132, 1371]
Average R	33.2	41.2
Range R	[4, 115]	[3, 161]
S-relevant	8.1	12.5
A-relevant	–	–
B-relevant	25.0	28.8
#Teams	16	14
#All runs	78	110
#Runs used for rank correlation	30	30

Figure 1. List of TREC03 and TREC04 data

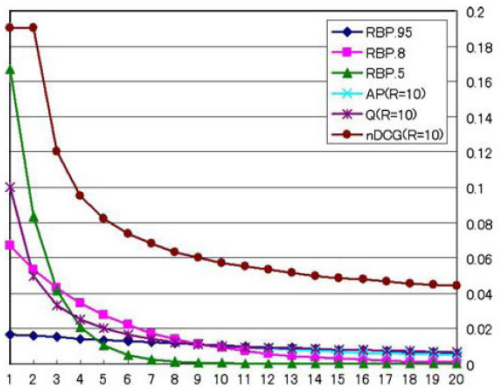

Figure 2. Comparison of different metrics based on top heaviness for TREC03 and TREC04 data

The above figure shows the comparisons of the ‘‘top-heaviness’’ among Average Precision (AP), RBP, nDCG and Q by considering a ranked output which contains just one relevant document from Rank 1 to 20. The situation when R = 10 is shown in the top of the graph, and the situation when R = 100 is shown in the bottom, under a binary relevance environment. It is clear from the figure that the value of R does not affect the RBP curves. It is clear from the figure that, RBP.5 ignores a relevant document retrieved below rank 10, therefore one can say it is perhaps too top-heavy. Due to this, evaluation becomes unbalanced.

Comparison of metrics based on discriminative power

The comparison based on discriminating power for metrics such as Q(0), AP(0), nDCG(0), bpref_R and RBP with the original 100% relevance data is shown in the following figure.

	Disc. power (%)	Diff. required			
(a) TREC03			(b) TREC04		
Q	80/120 = 66.7	0.07	Q	63/91 = 69.2	0.08
Q′	77/120 = 64.2	0.07	Q′	62/91 = 68.1	0.08
AP	77/120 = 64.2	0.07	AP	61/91 = 67.0	0.07
AP′	77/120 = 64.2	0.09	AP′	61/91 = 67.0	0.07
nDCG	71/120 = 59.2	0.08	nDCG	58/91 = 63.7	0.08
nDCG′	71/120 = 59.2	0.08	nDCG′	58/91 = 63.7	0.09
bpref_R	69/120 = 57.5	0.08	bpref_R	57/91 = 62.6	0.09
RBP.8	57/120 = 47.5	0.08	RBP.95	45/91 = 49.5	0.05
RBP.95	55/120 = 45.8	0.04	RBP.8	36/91 = 39.6	0.09
RBP.5	45/120 = 37.5	0.12	RBP.5	30/91 = 33.0	0.12

Figure 3. List of metrics with their discriminative power for TREC03 and TREC04 data.

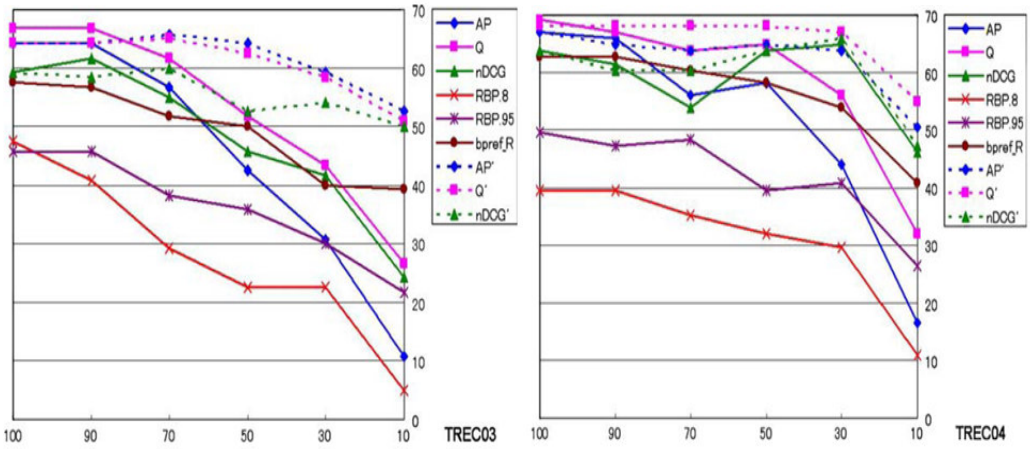

Figure 4. Metrics with their reduction rate (x-axis) versus discriminative power for TREC03 and TREC04 data.

It is clear from the figure that, metrics such as Q0, AP0 and nDCG0 are more robust than other matrices for TREC03 and TREC04, to incomplete relevance assessments. It is clear from the figure that, for TREC4 the original nDCG perform well, however, not for TREC03. Thus, the winners in terms of robustness, to incomplete relevance assessment are Q0, AP0 and nDCG0.

Conclusion

Evaluation of Information Retrieval System is crucial and also must be taken seriously, as the same set of retrieved documents might have different perceptions of relevance. The relevance of a document is totally subjective matter. Measuring performance of IR system using one metric depends on TREC we are using and every metric is designed to use in some specific TREC. I.e. one metric gives better output in one TREC for some particular IR system, it changes in other TREC for the same system. So instead of comparing the evaluation result of IR system by using various metrics of several TREC comparisons should be done between different metrics of same TREC.

References

1. Kalervo Järvelin , Jaana Kekäläinen, Cumulated gain-based evaluation of IR techniques, ACM Transactions on Information Systems (TOIS), v.20 n.4, p.422-446, October 2002 [doi>10.1145/582415.582418].
2. M. R. Grossman, Wachtell, Lipton, Rosen & Katz ,B. Hedin, H5 D. W. Oard and V. Cormack, Overview of the TREC 2010 Legal Track, University of Waterloo, University of Maryland, College Park K. Elissa, "Title of paper if known," unpublished.
3. Book: Retrieval Evaluation, Baeza-Yates & Ribeiro-Neto, Modern Information Retrieval, 2nd Edition.
4. Bateman, J. (1998). Changes in relevance criteria: A longitudinal study. In Proceedings of the 61st American Society for Information Science annual meeting 35 (pp. 23-32).
5. Allan, J., Callan, J., Croft, B., Ballesteros, L., Broglio, J., Xu, J., & Shu, H. (1997). Inquery at TREC 5. In E.M. Voorhees & D.K. Harman (Eds.), Information technology: The Fifth Text Retrieval Conference (TREC-5) (pp. 119-132). Gaithersburg, MD: National Institute of Standards and Technology.
6. Cuadra, C.A., & Katter, R.V. (1967). Experimental studies of relevance judgments: Final report. Vol. I. Project summary. Santa Monica, CA: System Development Corporation.
7. Green, R. (1995). The expression of conceptual syntagmatic relationships: A comparative survey. Journal of Documentation, 51(4), 315-338.
8. Rees, A.M., & Schultz, D.G. (1967). A field experimental approach to the study of relevance assessments in relation to document searching. Cleveland: Case Western Reserve University.
9. Saracevic, T. (1996). Relevance reconsidered '96. In P. Ingwersen & N.O. Pors (Eds.), Proceedings of the second international conference on conceptions of library and information science: Integration in perspective (pp. 201-218). Copenhagen: The Royal School of Librarianship.