The Relationship between IR Effectiveness Measures and User Satisfaction

Azzah Al-Maskari
Dept. of Information Studies
University of Sheffield
Sheffield, S1 4DP, UK
Lip05aaa@shef.ac.uk

Mark Sanderson
Dept. of Information Studies
University of Sheffield
Sheffield, S1 4DP, UK
m.sanderson@shef.ac.uk

Paul Clough
Dept. of Information Studies
University of Sheffield
Sheffield, S1 4DP, UK
p.d.clough@shef.ac.uk

ABSTRACT
This paper presents an experimental study of users assessing the quality of Google web search results. In particular we look at how users' satisfaction correlates with the effectiveness of Google as quantified by IR measures such as precision and the suite of Cumulative Gain measures (CG, DCG, NDCG). Results indicate strong correlation between users' satisfaction, CG and precision, moderate correlation with DCG, with perhaps surprisingly negligible correlation with NDCG. The reasons for the low correlation with NDCG are examined.

Categories and Subject Descriptors
B.8 **[Performance and Reliability]**: *General*

General Terms
Measurement, Performance

Keywords
User satisfaction, IR Effectiveness measures

1. INTRODUCTION
Search engines are among the most popular and useful services on the web. Since the effectiveness of a retrieval system should be evaluated on the basis of how much it helps users achieve their task effectively and efficiently [2], the rating of search engine results by the user should be taken into account to evaluate search engines as a whole. To our knowledge there is no previous work that directly addresses the relationship between *precision*, Cumulative Gain *(CG)*, Discounted Cumulative Gain *(DCG)*, Normalized Discounted Cumulative Gain *(NDCG*) and users' satisfaction. Therefore, this paper presents the relationship between these measures.

2. METHODOLOGY
The study asked 26 users to search on four queries from a pool of 104 queries of which compiled by the author; these queries are of four categories (26 queries of famous artists, 26 queries of some diseases, 26 queries of famous politicians, and 26 queries of religions' principles). Each user was given one query from each category. The tasks were designed to emulate a simple information finding task (i.e., find pages that contain relevant information to Mozart). Users searched directly in Google[1], which was chosen for its popularity and high effectiveness as identified by [1]. Time allotted for each task was twelve minutes and users were asked to judge the first ten pages resulting from the best query they issued. They rated the effectiveness of each page at three levels of relevancy: *highly-relevant*[2] (2); *reasonably relevant*[3] (1); *not relevant*[4] (0). Users also rated their satisfaction with the "accuracy", "coverage" and "ranking" of the results.

3. FINDINGS
Table 1 shows the effectiveness of Google' results as quantified by NDCG and precision of the search results and users' satisfaction with the accuracy, coverage and ranking of the results. Users' satisfaction with accuracy-only relevant pages retrieved- was obtained by from the users in three level scales: very satisfied (2), reasonably satisfy (1), not satisfied (0). Users' satisfaction with coverage-all relevant information is retrieved- and ranking-ordering of relevant results- are calculated in the same way as their satisfaction with accuracy.

Table 1. Overall effectiveness of Google measured across users and queries (measures are computed at the first 10 pages)

NDCG	0.87
Precision	0.63
(U. S)[5] Accuracy	0.54
(U. S) Coverage	0.54
(U. S) ranking of results	0.50

Pearson's correlation was used to find the relationship between these measures, shown in Table 2. According to these figures, Cumulative Gain (CG) is the best measure to strongly model users' satisfaction with the results (accuracy, coverage, ranking); followed by precision and then by DCG. Perhaps surprisingly, NDCG did not correlate well with users' satisfaction of the results.

Table 2. Correlation between measures

Measures	Pearson's Correlation
CG vs. DCG	0.90
CG vs. NDCG	0.37
CG vs. Precision	0.88
CG vs. U. S. ranking of results	0.79

[1] http://www.google.co.uk/

[2] The page directly addresses the core issue of the topic
[3] The page only points to the topic, but it does not discus the themes of the topic thoroughly
[4] The page does not contain any information about the topic.
[5] U.S= Users' Satisfaction

SIGIR' 07, July 23-27, 2007, Amsterdam, The Netherlands.
ACM 978-1-59593-597-7/07/0007

CG vs. U.S. Accuracy	0.68
CG vs. U.S. Coverage	0.60
DCG vs. NDCG	0.46
DCG vs. Precision	0.75
DCG vs. U. S. ranking of results	0.69
DCG vs. U.S. Accuracy	0.60
DCG vs. U.S. Coverage	0.50
NDCG vs. Precision	0.14
NDCG vs. U. S. ranking of results	0.26
NDCG vs. U.S. Accuracy	0.10
NDCG vs. U.S. Coverage	0.10
Precision vs. U. S. ranking of results	0.70
Precision vs. U.S. Accuracy	0.59
Precision vs. U.S. Coverage	0.53
U. S. ranking of results vs. U.S. Accuracy	0.79
U. S. ranking of results vs. U.S. Coverage	0.67
U.S. Accuracy vs. U.S. Coverage	0.72

*The highlighted measures indicate strong and significant correlation ($p<0.05$)

4. DISCUSSION

NDCG was initially tested and proven to work well in test collection evaluations with the existence of a wide range of relevance judgments [3] [4]. However, the work discussed here has a very limited set of judgments, first 10 pages, which effected NDCG during the normalizing step, especially in cases where the ranking was identical to the ideal which led NDCG=1.0 (this situation was unlikely to occur in the past work [3] [4]). Hence NDCG does not correlate strongly with other measures (such as precision, users' satisfaction with coverage, accuracy and ranking of the results) when working with a limited number of relevance judgments (i.e., the case when only judging the first 10 pages).

Figure 3 shows some example systems: showing in the first column the rank position, (r), followed by the gain: g(r), (g(2) = highly relevant, g(1) = reasonably relevant, g(0) =not relevant). The next columns are the cumulative gain, cg(r), discounted cumulative gain, dcg(r), normalized discounted cumulative gain, ndcg(r), and precision, (p). Systems A, B, C, D, E, and F all show the advantages of NDCG by incorporating all levels of relevance judgment, while giving precedence to more relevant items. However, NDCG has some drawbacks in disregarding the degree of relevancy if a system returns only documents with one level of relevancy (e.g. G & H), in these systems NDCG=1, because they are considered to have a prefect order (ideal ranking) and cannot be normalized (i.e., rearranging the documents in descending order of relevancy). This problem also occurs if a system has only one relevant item at the top of the rank (I & J); NDCG doesn't distinguish between the degrees of relevancy in these cases. Moreover, in the last two systems (K, L), NDCG is equal to 0.67; though these systems differ in relevancy but both have one relevant item located at the same rank. Therefore, for the last six systems, precision is a more appropriate measure than NDCG. Precision also has its limitation as shown in the first six systems (A, B, C, D, E, F) in disregarding the location of relevant items as well as in not considering multiple relevance judgments. Therefore, examining a combination of measures appears to be better because some measures work well in some conditions while others work better in others.

System A

r	g(r)	cg(r)	dcg(r)	ndcg(r)	p
1	2	2	2	1	
2	1	3	4	1	
3	0	3	5	1	
				1.00	0.67

System B

r	g(r)	cg(r)	dcg(r)	ndcg(r)	p
1	2	2	2	1	
2	0	2	3	0.75	
3	1	3	4.63	0.96	
				0.90	0.67

System C

r	g(r)	cg(r)	dcg(r)	ndcg(r)	p
1	1	1	1	0.5	
2	2	3	4	1	
3	0	3	5	1	
				0.83	0.67

System D

r	g(r)	cg(r)	dcg(r)	ndcg(r)	p
1	1	1	1	0.5	
2	0	1	2	0.5	
3	2	3	4.26	0.85	
				0.62	0.67

System E

r	g(r)	cg(r)	dcg(r)	ndcg(r)	p
1	0	0	0	0	
2	2	2	3	0.75	
3	1	3	4.63	0.92	
				0.56	0.67

System F

r	g(r)	cg(r)	dcg(r)	ndcg(r)	p
1	0	0	0	0	
2	1	1	2	0.5	
3	2	3	4.26	0.85	
				0.45	0.67

System G

r	g(r)	cg(r)	dcg(r)	ndcg(r)	p
1	1	1	1	1	
2	1	2	3	1	
3	1	3	4.6	1	
				1.00	0.00

System H

r	g(r)	cg(r)	dcg(r)	ndcg(r)	p
1	2	2	2	1	
2	2	4	5	1	
3	2	6	7.26	1	
				1.00	1.00

System I

r	g(r)	cg(r)	dcg(r)	ndcg(r)	p
1	2	2	2	1	
2	0	2	3	1	
3	0	2	4	1	
				1.00	0.33

System J

r	g(r)	cg(r)	dcg(r)	ndcg(r)	p
1	1	1	1	1	
2	0	1	2	1	
3	0	1	3	1	
				1.00	0.33

System K

r	g(r)	cg(r)	dcg(r)	ndcg(r)	p
1	0	0	0	0	
2	2	2	3	1	
3	0	2	4	1	
				0.67	0.33

System L

r	g(r)	cg(r)	dcg(r)	ndcg(r)	p
1	0	0	0	0	
2	1	1	2	1	
3	0	1	3	1	
				0.67	0.33

Figure 3. Imaginary systems that reflect NDCG vs. Precision

5. CONCLUSION

We have quantified the quality of Google based on a range of common IR measures with over 104 user queries. We have concluded that CG and precision correlate better than NDCG with users' satisfaction of the results, though NDCG has been proven to work well in search engine evaluation. We have also shown that both NDCG and precision have their own shortcomings, therefore, a combination of measures is better to complement each other when evaluating the effectiveness of IR systems.

6. ACKNOWLEDGMENTS

We would like to thank Ministry of Manpower, Oman, and the Tripod project (IST-FP6-045335) for funding this study.

7. REFERENCES

[1] Hawking, D., Craswell, N., Bailey, P. & Griffihs, K. Measuring Search Engine Quality. *Information Retrieval,* 4, 33-59.2001

[2] Ingwersen, P. & Järvelin, K. *The turn: integration of information seeking and retrieval in context*, Springer. 2005

[3] Järvelin, K. & Kekäläinen, J. IR evaluation methods for retrieving highly relevant documents. *ACM SIGIR Proceedings.* Athens, Greece. 2000

[4] Järvelin, K. & Kekäläinen, J. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20**,** 422-446. 2002