Research Journal of Information Technology

ISSN 1815-7432

Academic Journals Inc.

www.academicjournals.com

Research Journal of Information Technology 3 (1): 24-34, 2011
ISSN 1815-7432 / DOI: 10.3923/rjit.2011.24.34

Using Explicit Measures to Quantify the Potential for Personalizing Search

Fikadu Gemechu Erba, Zhang Yu and Liu Ting
Research Center for Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, People's Republic of China

Corresponding Author: Fikadu Gemechu Erba, Research Center for Information Retrieval, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, People's Republic of China

ABSTRACT

Currently existing web search engines return the same results for the same query issued to them. But, such systems do not satisfy the needs of different users having different information need underlying the same queries. In this study, we use explicit relevance judgment to show the variation in search results users find to be relevant. To get multiple judgments for the same query, we provide users with list of previously generated queries from our search engine and asked them to choose queries which are of interest to them and evaluate the search results quality for the query. Users are also asked to choose the queries they generated and evaluate the search results quality in the same fashion. The result we get shows that there is a great variation in users explicitly rating the same result for the same query and we use discounted cumulative gain to quantify this variation in relevance judgment. The result we get shows that with an increase in the number of people evaluating the same result for the same query, the gap between user satisfaction with an individual ranking and group ranking grows. Our experiments show that the best group ranking for a group of five people on average gives rise to a 26% improvement in discounted cumulative gain over the web ranking, while the best individual ranking leads to a 61% improvement over the web ranking.

Key words: Explicit relevance, search need, evaluate search result, user satisfaction, user rating, web search

INTRODUCTION

People use web search engines as a means for looking information they need on the web. They usually formulate the query, input to the search engine and the search engine returns a list of results relevant to the query issued. However, the results returned by the search engine may not be relevant to the users' information need and hence users need to modify and reformulate their queries (Jansen *et al*., 2007). Besides, users can have complex information needs while an individual query may only represent a piece of the user's information need. So, the biggest challenge for the search engine is to translate people's simple, short queries into lists of documents that satisfy their different information needs.

In previous work, it has been reported that relevance of search result to queries is correlated with the user search success (Huffman and Hochster, 2007). However, relevance of search result does not tell the complete picture, since queries can be ambiguous in many ways and users can

have different information needs underlying the same queries. For example, a web search for an ambiguous query DCG returns a wide range of results, including results about Dynamic Code Generator, Dynamic Compaign Generator, Diploma in Community Gynecology, DCG private limited company and few results about the Discounted Cumulative Gain. Although each of these results are relevant to someone who issues the query DCG, users who want to look for information on discounted cumulative gain would likely be not interested in most of what is returned. To this end some work has been done to help users articulate their information needs more clearly to the level of detail required (Kelly and Fu, 2007; Dumais *et al*., 2003; Anick, 2003).

In addition, there are greater differences in users rating the same results for the same query. This variability in rating the result for the same query is associated with the different search intents the users may have. For example, for a query swine flu two users rate the results differently, with the first rating four results to be highly relevant or relevant to the query whereas the second user rates three of the results to be highly relevant or relevant to the query. This difference in rating is attributed to the differences in users' intent (e.g., intent for the first user is statistics and treatment about the swine flu whereas intent for the second user is information about its transmission means). We also observe that there are differences in rating even when the users have written the same search intent behind the query.

Existing search engines have been designed with a notion of providing users with the same result for the same query submitted to them without considering the users' search context. Such systems do not fully satisfy different users' need even when they submit the same query to look for something on the web. The main difficulty lies in the way in which users articulate their information need and this affects performance of search engine in accurately returning appropriate result to each user (Carterette and Jones, 2007; Guo *et al*., 2010; Collins-Thompson and Bennett, 2009; He and Ounis, 2004; Hauff *et al*., 2008; Hassan *et al*., 2010).

Some work has been done to rate the relevance of results to the query using an explicit measures of relevance (Teevan *et al*., 2007; Ma *et al*., 2007). There are large variations in explicit judgment among individuals even when evaluating the relevance of result for the same query. This variation in explicit relevance judgment is the result of query ambiguity and difficulty inherently existing in the query generated by the users since users cannot clearly articulate their information need to the level of detail required to identify the different information needs underlying the same query. So it is important to make use of other information sources to disambiguate the query (Wang and Agichtein, 2010; Agichtein and Guo, 2010; Teevan *et al*., 2008). Besides, it is also observed that variation in users' relevance judgment of results of the same query was clearly reflected in their search intent. Search intent determination is an alternative means for personalized search and that is why search intent determination is a hot research direction nowadays (Agichtein and Guo, 2010; Guo and Agichtein, 2009).

To investigate how query ambiguity affects result quality, we conduct a study to examine the consistency of relevance judgments assigned by different individuals to the results of the same query. We use an explicit measure of relevance judgment to quantify the differences in result relevance between individuals. This helps us to better understand the potential benefit to be gained by personalizing the search result. In our study we use potential for personalizing curve to show the potential of the search result for personalizing and try to examine how this potential for personalizing varies with the number of people. This also enables us to understand how difficult it is for the search engine to satisfy interests of the large number of users who submit the same query with different information need.

The rest of this study is organized as follows: we begin present study with a description about the data we have collected followed by the methodology we have employed. We then give brief description of Discounted Cumulative Gain (DCG) and how it is used in quantifying the value of personalized search using the potential for personalizing curve. Finally, we present a conclusion of present study.

METHODS AND DATA SETS

There are various ways to evaluate whether a document is relevant to a query. They are generally classified into two broad categories namely explicit relevance judgment and implicit measures of relevance. In this work we only focus on the explicit relevance judgments which is the most commonly used measure for assessing the relevance of a document to a user's query in information retrieval literature.

The user search data were collected at research center of information retrieval, Harbin Institute of Technology, China using the search engine we developed. Since the users of our system were few, the data collection process was carried out for a period of 4 months from November 2010 to February 2010. Then after we asked some 5 participants of our work to explicitly evaluate the relevance of search result to query. The data we collected for an explicit relevance measure is summarized in Table 1. We collected explicit relevance judgments for 43 queries. Since our focus is on the difference in judgment across individuals for the same query, the table also provides information about how many of the queries are unique and how many have relevance judgments from multiple different users. We then give a detail description of how relevance judgments were collected, in particular how we were able to obtain many sets of judgments for the same query. In the subsequent sections we use this measure to understand how well web search engines currently perform and how well they could ideally perform to return the results to each individual.

Explicit relevance judgment: The simplest way to determine whether the results returned by the search engine are relevant to the query issued by an individual is to explicitly ask the individual. The best example of this is the Text REtrieval Conference (TREC) collection used in evaluation of information retrieval systems which is constructed using explicit judgments (Voorhess and Harman, 2005). In this approach, users are asked to rate the relevance of results to query based on a detailed description of an information need.

This approach is unrealistic for web search, in which people issue very short queries to describe their information needs (Spink and Jansen, 2004). It is less likely that the same short web query, when issued by different people, has the same unambiguous information goal behind it. Besides, TREC studies focus on whether documents are topically relevant to the query and not whether an individual user would be satisfied with the document. Hence, rather than people evaluate results for fully defined information goals, we asked our judges to indicate which results they would personally consider highly relevant, relevant and irrelevant to an information need specified with a more typical web query.

Table 1: Lists the number of people from which the explicit measure was collected, the total number of queries gathered for each, the number of unique queries and the number of queries with judgments from more than one individual. Also listed is how an explicit measure is quantified, which is labeled gain

Relevance measure	No. of users	No. of queries	No. of unique	≥2 users	Gain
Explicit judgments	5	43	23	20	2 if highly relevant 1 if relevant 0 if not relevant

We asked the judges in our study to evaluate how personally relevant the top 10 web search results were to them. Our participants have many commonalities (e.g., all were postgraduate students in engineering, lived almost in the same area and had similar computer literacy). For the purpose of collecting search results, we developed a search engine in our lab. The web results collected, containing a title, snippet and URL, from our search engine were presented to the participants in the same format as web results. The actual result page could be seen by using the URL, but was only viewed when the participant felt doing so was necessary to make a judgment. For each search result, each participant were asked to determine whether they personally found the result to be highly relevant, relevant, or not relevant to the query. To avoid participants' bias toward the rank, results were presented to them in random order.

In this research, the queries were selected in two different manners at the participants' end. In the first approach participants were asked to select a query from a list of pre-selected queries. In the second approach users are asked to choose the query from the query they have generated by themselves. We call such queries self-generated. Participants were asked to write the search intent they had in mind when they issued self-generated queries, or were interested in for the pre-selected queries.

The reason why we needed both types of queries is that it allowed us to balance the value that could be obtained by studying naturally occurring self-generated queries with the need to collect multiple judgments for the same query. We requested people to select results from a pre-selected list which in turn helped us to explore the consistency with which different individuals evaluated the same results for the same query. Getting such type of data only from self-generated queries is difficult since it would demand from us a longer time until different participants coincidentally issued the same query. Hence the pre-selected queries provide a way to get overlap in queries across people. In our study, participants were encouraged to select only pre-selected queries that were of direct interest to them. This selection process somewhat mitigates the artificial nature of using the pre-selected queries. We collected self-generated queries so that we could directly compare overall patterns of judgments for self-generated tasks and pre-selected queries and explore any potential discrepancies. Russell and Grimes (2007) have shown that searchers behave somewhat differently for assigned and self generated search tasks (e.g., spending more time on self generated tasks and but generating fewer queries). But as shown in Fig. 1, there are no differences in the overall distribution of explicit relevance judgments for the two types of queries used in our studies.

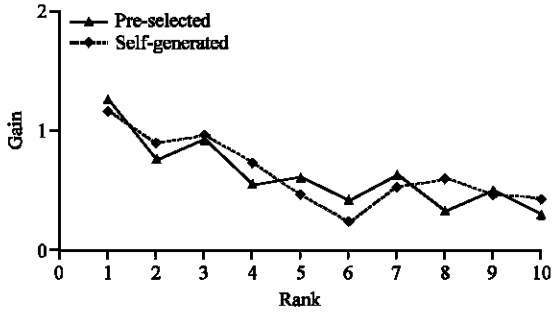

Fig. 1: The average gain based on explicit ratings for web search engine results as a function of rank. Separate curves are shown for the pre-selected (solid-line) and self-generated (dotted line) queries. Although there are some relationship between rank and relevance, many relevant results are not ranked top

Table 2: List of pre-selected queries for which results were explicitly evaluated by at least two users

Query	Users
active and passive voice	5
aids	5
bbc	5
blind channel equalization	2
expasy	2
hidden markov model	5
hk	3
ir	3
java pdb viewer	4
letter of recommendation	5
nokia n97 phone	5
office	5
personalization	5
phd scholarships in erasmus munds	5
software	3
solid works tutorial	3
swine flu	5
voa news	4
water	5
yahoo	5

We collected explicit relevance judgments for 43 queries. Of the queries, 37 were pre-selected and 6 were self-generated. The number of people who evaluated the results for the same pre-selected query ranged from 2 to 5. There were 20 unique queries that had judgments from at least 2 people. Table 2 shows the queries evaluated by more than one user.

POTENTIAL FOR PERSONALIZING SEARCH

Web search engines do not do a perfect job of ranking search results and the most likely reason is the variation in what different people consider relevant to the same query. In this section we try to put this variation in individual's relevance judgment in quantitative terms. This analysis enables us to better understand the potential benefit to be gained from personalizing search. First we give a description of how this potential for personalizing search is quantified and then describe findings using explicit judgments.

Computing the potential for personalizing search: In this work we used DCG to evaluate search results (Järvelin and Kekäläinen, 2002). DCG is a measure of effectiveness of a web search engine algorithms or related applications, often used in information retrieval. Using a graded relevance scale of documents in a search engine result set, DCG measures the usefulness, or gain, of a document based on its position in the result list. The gain is accumulated cumulatively from the top of the result list to the bottom with the gain of each result discounted at lower ranks. The basic idea behind the DCG approach is that higher ranked documents are worth higher value than lower ranked ones and also allows us to incorporate the notion of multiple level of relevance. More formally, DCG is defined as the sum of the gain of presenting a particular document times a discount of presenting it at a particular rank, up to some maximum rank j.

$$DCG\ (j) = \sum_{i=1}^{j} gain\ (i)\ discount\ (i) \tag{1}$$

For web search, gain is typically a relevance score determined from a human labeling and discount is the reciprocal of the log of the rank, so that putting a document with a high relevance score at a low rank results in a much lower discounted gain than putting the same discount at a high rank.

$$DCG\ (j) = G\ (1) + \sum_{i=2}^{j} \frac{G\ (i)}{\log_2 i} \tag{2}$$

The constants G (i) are the relevance scores. Human evaluators typically judge the documents on an ordinal scale, with the labels such as highly relevant, relevant and non-relevant. These are then mapped to a numerical scale for use in DCG computation. The three ordinal labels G(i) are assigned values of 2, 1 and 0 respectively as shown in Table 1.

Search result lists vary in length depending on the query. Comparing a search engine's performance from one query to the next cannot be consistently achieved using DCG alone, so the cumulative gain at each position for a chosen value of *j* should be normalized across queries. This is done by sorting documents of a result list by relevance, producing an Ideal Discounted Cumulative Gain (IDCG) at position j. For a query, the normalized discounted cumulative gain, or nDCG, is computed as:

$$nDCGG\ (j) = \frac{DCG(j)}{IDCG(j)} \tag{3}$$

The nDCG values for all queries can be averaged to obtain a measure of the average performance of a search engine's ranking algorithm. Note that in a perfect ranking algorithm, the DCG (j) will be the same as the IDCG (j) producing an nDCG of 1.0. All nDCG calculations are then relative values on the interval 0 to 1 and so is cross-query comparable.

We considered only results returned up to the rank level of ten for the sake of simplicity. So, for each query scores are cumulated for all ranks giving us a single summary measure for the quality of a set of results. This in turn helps us to make a relative comparison of the quality of results returned for queries since queries with more relevant documents have a higher DCG value than the one with less relevant results.

As an example, Table 3 shows the web search results for the query nokia n97 phone and the gain associated with each result for two different users. In this example, user I rated three results as relevant to his information need and user II rated one result as very relevant and three as relevant to his information need. Employing these scores we compute a DCG measure for each column giving a summary measure of the ranked list of results for a given user. The normalized DCG for user I and user II are 0.67 and 0.73 respectively. As shown in the column labeled I+II, on average the normalized DCG for the web ranking for users I and II is 0.70.

Taking the DCG value as a summary measure of the quality of a ranked list of web search results, the best possible ranking for a query is the ranking with the highest DCG value. From Eq. 2, DCG can be maximized by listing the results with the highest gain first. For example, for queries with explicit judgments where only one user evaluated the results, DCG can be maximized

Table 3: A ranked list of top ten web results for the query nokia n97 phone and the gain for two users based on their explicit judgments. The group gain for both users (I+II) represents the quality of the result for two of them

Result	Gain I	Gain II	I+II
gsmarena.com/nokia_n97-2615.php	0	1	1
mobilewhack.com/reviews/nokia_n97_phone.html	1	1	2
europe.nokia.com/find-products/devices/nokia-n97	0	0	0
europe.nokia.com/find-products/nseries	0	0	0
nokiausa.com/find.../phones/nokia-n97	1	0	1
nokiausa.com/find.../phones/nokia-n97-mini	0	0	0
telegraph.co.uk/.../Nokia-N97-Nokia-launches-iPhone-killer-N97-phone.html	0	1	1
engadget.com/2008/.../nokia-unveils-flagship-n97-phone	1	2	3
reviews.cnet.com/smartphones/nokia-n97.../4505-6452_7-33421200.html	0	0	0
phones4u.co.uk/reviews/nokia_n97	0	0	0
	User I	User II	Average
Normalized DCG	0.67	0.73	0.70

Table 4: The best possible ranking of search results for a query nokia n97 phone for user I and user II. The rightmost section shows the best possible ranking if the same list must be returned to user I and user II. The normalized DCG for the best ranking when only one person is taken into account is 1. When one or more than one person must be accounted for, the normalized DCG drops

				Best ranking for group (I+II)			
Best ranking for user I		Best ranking for user II			Gain		
Web result	Gain I	Web result	Gain II	Web result	I	II	I+II
mobilewhack.com/rev...	1	engadget.com/2008/...	2	engadget.com/2008/...	1	2	3
nokiausa.com/.../nokia-n97	1	gsmarena.com/nokia_n97...	1	mobilewhack.com/rev...	1	1	2
engadget.com/2008/...	1	mobilewhack.com/rev...	1	gsmarena.com/nokia_n97...	0	1	1
gsmarena.com/nokia_n97...	0	telegraph.co.uk/.../Nokia-N97...	1	nokiausa.com/.../nokia-n97	1	0	1
europe.nokia.com/.../devices...	0	europe.nokia.com/.../devices...	0	telegraph.co.uk/.../Nokia-N97...	0	1	1
europe.nokia.com/.../nseries	0	europe.nokia.com/.../nseries	0	europe.nokia.com/.../devices...	0	0	0
nokiausa.com/.../nokia-n97-mini	0	nokiausa.com/.../nokia-n97	0	europe.nokia.com/.../nseries	0	0	0
telegraph.co.uk/.../Nokia-N97...	0	nokiausa.com/.../ nokia-n97-mini	0	nokiausa.com/.../ nokia-n97-mini	0	0	0
reviews.cnet.com/smartphones...	0	reviews.cnet.com/smartphones...	0	reviews.cnet.com/smartphones...	0	0	0
phones4u.co.uk/reviews...	0	phones4u.co.uk/reviews...	0	phones4u.co.uk/reviews...	0	0	0
	I		II		I	II	Avg.
Normalized DCG	1.00	Normalized DCG	1.00	Normalized DCG	0.95	0.98	0.97

by ranking highly relevant documents first, relevant documents next and irrelevant documents last. The best ranking of the results for nokia n97 phone for Users I and II individually can be seen in the two left columns of Table 4. Because these lists are the best possible for the individual, the normalized DCG for these rankings is 1 in each case. Note that this is an ideal case in which we have explicit judgments of how relevant each result is for each user. The best a search engine could do for this user is to match these judgments by returning the results in this order.

When there are more than one set of ratings for a result list, the ranking that maximized DCG ranks the results that have the highest collective gain across raters first. For queries with explicit judgments, this means that results that all raters thought were highly relevant are ranked first,

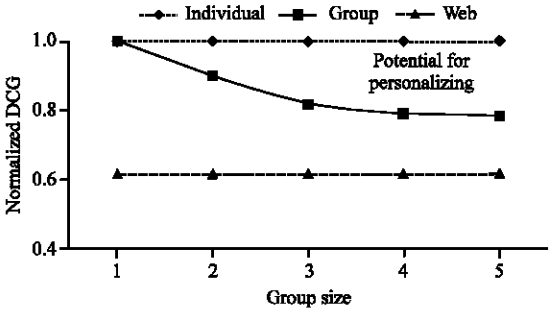

Fig. 2: As more people are taken into account, the average DCG for each individual drops for the ideal group ranking, but is constant for the ideal personalized ranking

followed by those that most people thought were highly relevant but a few people thought were just relevant, followed by results most people thought were relevant, etc. The collective gain for users I and II for the query nokia n97 phone is shown in the rightmost column of Table 3 and the results ranked according to that gain can be seen in the rightmost column of Table 4. Because the list must satisfy more than one person, it is no longer the best list for either individual. Instead, as shown in the Table 4, the normalized DCG for the best group ranking is 0.95 for User I and 0.98 for User II, for an average of 0.97. This group normalized DCG, which is 0.97 is lower than the normalized DCG for the best individual ranking (1.00 for both I and II).

For those queries where we have measures of relevance from multiple people, it is possible to find the best possible ranking for each individual, as well as the best possible ranking for different sized groups of individuals. As additional people are added to the group, the gap between user satisfaction with the individual rankings and the group ranking grows. This gap between best possible web search result ranking for an individual and the best possible web search result ranking for the group is what we call potential for personalizing search.

Quantifying potential for personalizing using explicit measures: We tried to quantify the potential for personalizing using explicit measures of relevance we collected. Figure 2 shows the potential for personalizing curve drawn for different groups of people. It also shows the average normalized DCG for the best ranking individual (dotted line), group (solid line) and web rankings (dashed line) as a function of the number of individuals in the group. These data were derived from the 20 pre-selected queries for which we collected explicit relevance evaluations of the results from more than two users as shown in Table 1.

We could only explore small groups employing explicit data. Search engines that do not adapt the search experience to individual users must try to find the best possible result ranking for the much larger group of people consisting of all possible searchers. It is impossible for us to explore the preferences of everyone, since it is not practical to collect relevance judgments from everyone even implicitly. We use potential for personalizing curve as an alternative means to make a good estimate about what the potential would be among large groups by looking at how it increases as group size increases.

As shown in the Fig. 2, with perfect personalization the average normalized DCG for an individual is 1. As more people's interests are taken into account to generate a ranking, the average normalized DCG for each individual drops for the ideal group ranking. The gap represents the

potential value to be gained by personalizing the search results. There is also a gap between the current normalized DCG for the Web results and the best group ranking, which represents the potential improvement to be gained merely by improving results without consideration of individuals.

As we can see in Fig. 2, the best group ranking for a group of five people on average gave rise to a 26% improvement in DCG over the current web ranking (0.78 vs.0.62), while the best individual ranking led to a 61% improvement (1.00 vs. 0.62). From the shape of the curve, we can infer that with an increase in size of the number of people in the group the DCG value decreases and becomes closer to the web ranking which aims to satisfy large number of searchers interests for the same query. The experimental result we obtained is not unique to this work but it is in complement with the result obtained by Teevan *et al.* (2007) in their effort to characterize the value of personalizing search. Teevan *et al.* (2008) shown that the potential of ambiguous query for personalization varies greatly with the size of the people and the general pattern of the result obtained is also consistent with the result we obtained although there is a large difference in the size of the data used in both cases.

CONCLUSIONS

We tried to explore the differences in what people consider relevant to the same query by using explicit measures. For our study, we collected explicit relevance judgments from our participants to know how the judgment varies among individuals even when they have the same intentions behind the query. We employed this measure of relevance to measure the variability in judgments and behaviors for the same query. The result we found shows that there are significance differences among individuals. Hence, there is a large gap between how well the search engines could perform if they were to adapt results to individuals and how well they currently perform by returning a single ranked list of results designed to satisfy everyone. This observed gap is what we called a potential for personalizing search and quantified empirically using normalized discounted cumulative gain.

Although, explicit relevance judgments allow us to examine the consistency in relevance assessments across different individuals, it is impractical for a search engine to collect relevance judgments. It is cumbersome for people to provide explicit judgments and challenging to gather sufficient data. Therefore, in the future we plan to explore the variation in relevance judgments among individuals for the same query using implicit measures of relevance.

ACKNOWLEDGMENTS

We would like to thank Habtamu Beri, Olivia Juba, Yigezu Balcha and Yilma Taye for their unreserved help in collecting explicit relevance feedback data.

REFERENCES

Agichtein, A. and Q. Guo, 2010. Towards inferring web searcher intent from behavior data. Proceedings of the 28th International ACM Conference on Human Factors in Computing Systems, April 10-15, Atlanta, GA, USA., pp: 1-4.

Anick, P., 2003. Using terminological feedback for web search refinement: A log based study. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28-Aug. 1, Toronto, Canada, pp: 88-95.

Carterette, B. and R. Jones, 2007. Evaluating search engines by modeling the relationship between relevance and clicks. Proceedings of the 21st Advances in Neural Information Processing Systems, (ANIPS'07), University of Massachusetts Amherst, Burbank, CA, pp: 217-224.

Collins-Thompson, K. and P.N. Bennett, 2009. Estimating query performance using class predictions. Proceedings of the 32th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 19-23, Boston, USA., pp: 672-673.

Dumais, S.T., E. Cutrell, J.J. Cadiz, G. Jancke, R. Sarin and D. Robbins, 2003. Stuff I've Seen: A system for personal information retrieval and re-use. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 28-Aug. 1, Toronto, Canada, pp: 72-79.

Guo, Q. and A. Agichtein, 2009. Beyond session segmentation: Predicting changes in search intent with client-side user interactions. Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information retrieval, July 19-23, Boston, Massachusetts, USA., pp: 636-637.

Guo, Q., R.W. White, S.T. Dumais, J. Wang and B. Anderson, 2010. Predicting query performance using query, result and user interaction features. Proceedings of the 9th International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information, April 28-30, Paris, France, pp: 1-4.

Hassan, A., R. Jones and K.L. Klinkner, 2010. Beyond DCG: User behavior as a predictor of successful search. Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, Feb. 3-6, New York, USA., pp: 221-230.

Hauff, C., V. Murdock and R. Baeza-Yates, 2008. Improved query difficulty prediction for the web. Proceedings of ACM 17th Conference on Information and Knowledge Management, Oct. 26-30, California, USA., pp: 439-448.

He, B. and I. Ounis, 2004. Inferring query performance using pre-retrieval predictors. Proceeding of Symposium on String Processing and Information Retrieval, Oct. 5-8, Padova, Italy, pp: 43-54.

Huffman, S.B. and M. Hochster, 2007. How well does result relevance predict session satisfaction. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 23-27, Amsterdam, Netherlands, pp: 567-574.

Jansen, B.J., M. Zhang and A. Spink, 2007. Patterns and transitions of query reformulations during web searching. IJWIS, 3: 380-390.

Järvelin, K. and J. Kekäläinen, 2002. Cumulative gain-based evaluation of IR techniques. ACM Trans. Inform. Syst., 20: 422-446.

Kelly, D. and X. Fu, 2007. Eliciting better information need descriptions from users of information systems. Inform. Process. Manage., 43: 30-46.

Ma, Z., G. Pant and O.R.L. Sheng, 2007. Interest-based personalized search. ACM Trans. Inform. Syst., Vol. 25.

Russell, D.M. and C. Grimes, 2007. Assigned and self- chosen tasks are not the same in web search. Proceedings of the 40th Annual Hawaii International Conference on System Sciences, Jan. 3-6, Big Island, Hawaii, USA., pp: 1396-1403.

Spink, A. and B. Jansen, 2004. Web Search: Public Searching of the Web. Kluwer Academic Publishers, Dordrecht: The Netherlands.

Teevan, J., S. Dumais and E. Horvitz, 2007. Characterzing the value of personalizing search. Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Informational Retrieval, July 23-27, Amsterdam, Netherlands, pp: 757-758.

Teevan, J., S.T. Dumais and D.J. Liebling, 2008. To personalize or not to personalize: Modeling queries with variation in user intent. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Informational Retrieval, July 20-24, Singapore, pp: 163-170.

Voorhess, E.M. and D.K. Harman, 2005. TREC: Experiment and Evaluation in Information Retrieval. MIT Press, USA.

Wang, Y. and E. Agichtein, 2010. Query ambiguity revisited: Click through measures for distinguishing informational and ambiguous queries. Proceedings of 11th Annual Conference of the North American Chapter of the ACL HLT, June 1-6, Los Angeles, California, USA., pp: 361-364.