Search Engines Comparison on the Basis of Session Duration and Click Hits

Rajesh Kumar Goutam[1] , Sanjay K. Dwivedi[2]
[1] Department of Computer Science, Babasaheb Bhimrao Ambedkar University
Lucknow, 226025, India

[2] Department of Computer Science, Babasaheb Bhimrao Ambedkar University
Lucknow, 226025, India

Abstract

The evaluation of search engines has greatly diversified in recent years. Evaluation campaigns are required to continuously re-consider their tasks and updating evaluation function in order to satisfy the users. We presented two user action dependent approaches to rank the results, namely Session duration time and Click Hits. Furthermore, we have conducted an experiment with 25 TREC queries to do comparison of five popular search engines.

Keywords: *Information Retrieval, Search Engine performance, Search Engine Comparison, Session Duration, Click Hits.*

1. Introduction

Search engines are constructed to quickly find useful information from the web. Thousand of search engines are existing and each with different evaluating function. It is essential to know which search engine provides most relevant results. Several evaluation techniques have been developed and most of them are heavily dependent upon precision and recall. The potential advantages of manual evaluation techniques are their accuracy with respect to users requirements. The main problem with the statistical techniques is that it is subjective, time-consuming and also requires human evaluators to evaluating the results.
Automatic evaluation is considered much better for rapid changing web behavior, does not require any extra overhead cost and expert judgments. Session duration is the time period from the beginning query until the searcher departed the search engine for the last time [6]. In this paper, we followed the approach; if two ranking methods mark the same results in top ranking then these are considered relevant [5]. In the following section, we utilized the session duration to correlate with click hits to know the closeness between rankings of retrieved results when these are arranged separately.

2. Related work

Web search engine evaluation is complex and expensive process, which requires human evaluators to provide relevance judgments that indicate the degree of relevance of each document obtained for each query. Clicks data from information seekers, used in aggregate, can provide valuable indication about the relevancy of retrieved document. Joachims [1] utilized clicks to decide user preferences between retrieved results and presented an approach to mining logfiles of search engines with the aim of enhancement in their retrieval performance automatically. With the help of experimental support, author showed that click hits can resolve the problem of search engines evaluation efficiently without the expert's editorial grade. Terveen et al. [2] addressed the question whether two rankings can be correlated using human judgments to measure the quality of search engines and whether link-based metrics are good indicators of webpage quality. They found that correlation may be the best way to compare the search engines efficiently.
Massimo Melucci[7] utilized kendall's correlation method to compare two rankings ranked on the basis of dwell time and snippets. Donald Metzler et al. [3] examined Expected Reciprocal Rank (ERR) that correlates better with clicks metrics than other editorial metrics. Authors also correlated an editorial with a click metric and found that number of clicks in a session negatively correlated with editorial metrics. Clark et al. [4] worked with Reciprocal Rank Fusion (RRF) and combined the rankings from multiple IR systems and found better results than any individual retrieval system. Yilmaz et al. [5] believed in Excepted Browsing Utility (EBU) and extended the NDCG and RBP to incorporate probability of clicks in their discount curves.

3. Evaluation of Information Retrieval Systems

The information gathering process is inherently vague. In most systems, documents and queries include natural language. The data of the documents needs to be analyzed, which is a complex task for IR systems. Therefore, documents holding text are represented by natural language words frequently, without syntactic or semantic context. This is often known as the bag-of-words approach. These language keywords or query terms can imperfectly represent an item because their context and relations to other terms with in the documents and queries are lost in the indexing process. Information retrieval systems can be implemented through various ways by choosing a model and specific language processing tools. They perform in a complex system and their performance for a specific data collection cannot be guessed. As a consequence, the practical evaluation of performance is a main concern in information retrieval systems. Researchers are facing challenge to find measures which can be utilized to determine whether a system is better than another one. The most traditional basic measures are recall and precision. Recall indicates the capability of a system to select relevant documents, whereas precision measures indicate towards how good a retrieval system is in finding only relevant documents without any intruder efforts.

4. Clickthrough Data In Search Engines

Search engine evaluation is not easy as it was before. Now the scenario has been changed, web is expanding continuously in size so it is not possible to examine documents manually for their relevancy.

Automatically identifying the effectiveness of retrieval functions based on observable users behavior holds guarantee for making retrieval evaluation quicker, cheaper, and more user centered. Click hits have proven themselves effective for judging the quality of search engines and retrieval functions [1].

The potential advantages are clear unlike human evaluators judgments, usage data can be captured essentially at zero cost, it is existing in real time, and reflects the decisions of the users, not those of judges holds the experience in their area. Clearly, information seeker do not click on links at random, but on the basis of furnished information such as title and summery. Clicks through data is informative and known for most “perfect” relevance judgments, but key issue is: how can this information be gathered? We shall discuss the solution of this problem in following section.

5. Incorporating Implicit Feedback

We consider two types of approaches to ranking the results with the implicit feedback: (1) treating click hits as independent evidence for ranking results, and (2) integrating session duration into the correlation calculation. We utilized both these users signals to calculate the correlation that help to re-arrange the ranked results.

The general approach to re-rank the results automatically, utilizes users actions for the presented queries in previous search sessions. Each result is assigned a numeric score in terms of expected relevance/user satisfaction based on previous interactions. merging multiple rankings is very popular method to re-rank the obtained results holding promise to arrange the results in efficient way, we selected a simple and robust approach to arrange the retrieved results according to click hits and session duration separately to calculate the correlation to know the similarity in the preference ordering arranged on the basis of click hits and session duration.

6. Methodology

One of the objectives of this study is to present a comparison of the performance of five popular search engines in terms of Session Duration and Click Hits.

Table 1: Selected TREC queries

Selected TREC queries
Specialist in childhood education.
US whether service
Family Education rights
Crystals in urine results
Surface area evaporation
Earthquake in California
Kidney and Lung Cancer
Job Safety Analysis
Global warming
Wright Brothers
School Bus Safety
Online coloring Books
Heroin found in Cigar
Unique rare coins
Hand washing gel
Old age skin care
platinum
Nuclear commission
Water wastage in India
Skilled nursing facility in California

planet mercury discovery
iodine in blood
weight of mail
listeria infections
equation for sources of energy

To perform the performance comparison between the different search engines, it is necessary to have several standard queries from various field, we utilized TREC queries collection. We completely avoided popular queries such as popular entertainer; popular location, popular songs etc. because some time web search engines cache previous results from these highly queried topics. Another requirement was the select a clicks capturing device that holds capability to count Clicks automatically, Mousotrom 5.0 made it possible.

Mouse Movement Recorder which holds the quality to track of the mouse movement path. The main purpose of using mouse movement recorder is to recognize the monitor area where user performs the maximum actions. Suppose cursor moves in the middle of the screen in horizontal direction repeatedly this indicates towards the fact that user is reading the results. If recorded cursor make hits on the top left of the screen, means user is performing action at toolbar.

7. Performance Comparison of Search Engines

Ranking the retrieved results is a basic problem in the process of information retrieval. Most common approaches are heavily dependent upon the similarity of query and a page text. But instead of following the traditional approach we emphasized upon Session Duration and Click Hits. We make some predictions from the searchers interactions. Table 2 shows the some predictions from the users actions.

The ultimate goal of including implicit feedback into ranking is to improve the relevancy of the retrieved results. To perform this task, we have selected five popular search engines named Google, AltaVista, Yahoo, Excite and MSN. We executed each of selected queries to every search engines involved. We focus our whole concentration to measure the time devoted to find the first desired and satisfactory result for each query.

We totally avoided META tags consideration during the relevance calculation. We focused our whole concentration towards the session time for first most satisfactory result and click hits over results. The findings of the experiment are summarized in table 3.

Table 2: Prediction for experiment

Mouse Action	***Meaning***
Distance traveled	Users Efforts
Key Strokes	Query formulation
Time	Depth of desired results
Left Button Click	Action on sorted list of results responded by search engines
MouseWheel frequency	Concentration on particular document
Mouse traversing speed	Attention on results
Hits on top left of screen	Working with toolbar.
Hits on lower right	Desired for more results

Table 3: Query-by-query comparison of Session Duration and Clicks Hits

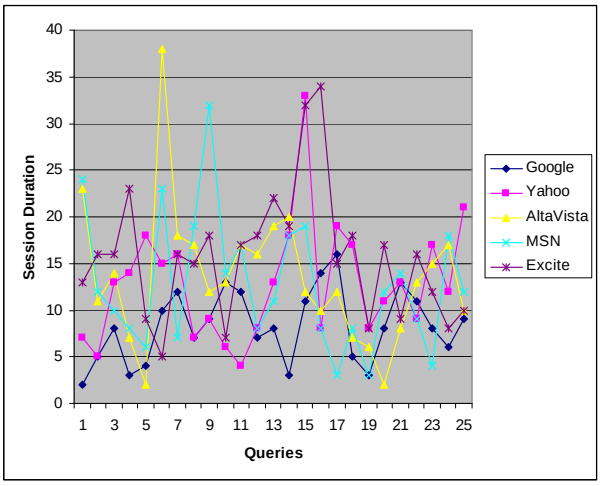
Queries	***Google***		***Yahoo***		***AltaVista***		***MSN***		***Excite***	
	Session Duration (Minute)	***Click Hits***	***Session Duration (Minute)***	***Click Hits***	***Session Duration (Minute)***	***Click Hits***	***Session Duration (Minute)***	***Click Hits***	***Session Duration (Minute)***	***Click Hits***
1	*2*	*108*	*7*	*78*	*23*	*72*	*24*	*71*	*13*	*45*
2	*5*	*90*	*5*	*83*	*11*	*44*	*12*	*40*	*16*	*42*
3	*8*	*74*	*13*	*151*	*14*	*55*	*10*	*34*	*16*	*47*
4	*3*	*34*	*14*	*168*	*7*	*30*	*8*	*29*	*23*	*75*
5	*4*	*40*	*18*	*180*	*2*	*12*	*6*	*22*	*9*	*35*
6	*10*	*78*	*15*	*156*	*38*	*92*	*23*	*70*	*5*	*19*
7	*12*	*121*	*16*	*162*	*18*	*43*	*7*	*28*	*16*	*49*
8	*7*	*72*	*7*	*72*	*17*	*65*	*19*	*57*	*15*	*51*
9	*9*	*82*	*9*	*109*	*12*	*40*	*32*	*76*	*18*	*58*
10	*13*	*123*	*6*	*86*	*13*	*42*	*14*	*54*	*7*	*31*
11	*12*	*128*	*4*	*60*	*17*	*60*	*17*	*68*	*17*	*68*
12	*7*	*130*	*8*	*78*	*16*	*64*	*8*	*32*	*18*	*54*
13	*8*	*78*	*13*	*102*	*19*	*43*	*11*	*55*	*22*	*37*
14	*3*	*44*	*18*	*192*	*20*	*56*	*18*	*32*	*19*	*47*
15	*11*	*142*	*33*	*220*	*12*	*40*	*19*	*51*	*32*	*156*
16	*14*	*172*	*8*	*78*	*10*	*32*	*8*	*34*	*34*	*168*

17	*16*	*184*	*19*	*220*	*12*	*63*	*3*	*12*	*15*	*75*
18	*5*	*56*	*17*	*197*	*7*	*40*	*8*	*39*	*18*	*83*
19	*3*	*46*	*8*	*102*	*6*	*28*	*3*	*16*	*8*	*33*
20	*8*	*98*	*11*	*120*	*2*	*10*	*12*	*39*	*17*	*68*
21	*13*	*113*	*13*	*156*	*8*	*90*	*14*	*152*	*9*	*104*
22	*11*	*114*	*9*	*123*	*13*	*122*	*9*	*103*	*16*	*156*
23	*8*	*99*	*17*	*182*	*15*	*165*	*4*	*56*	*12*	*102*
24	*6*	*76*	*12*	*124*	*17*	*182*	*18*	*205*	*8*	*76*
25	*9*	*102*	*21*	*222*	*10*	*164*	*12*	*122*	*10*	*83*
Average	***8.28***	***96.16***	***12.84***	***136.84***	***13.56***	***66.16***	***12.76***	***59.88***	***15.72***	***70.48***

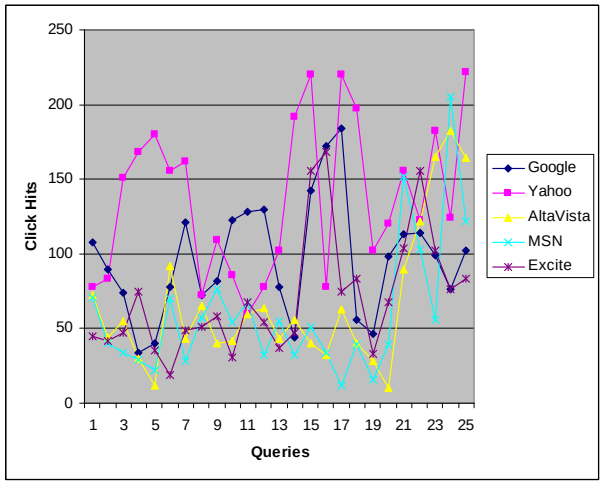
Table 4: Search engines comparison

Ranking	***Session Duration***	***Clicks Hits***	***Correlation***
1st Rank	Google	Yahoo	Google (0.67)
2nd Rank	MSN	Google	Yahoo (0.62)
3rd Rank	Yahoo	Excite	AltaVista(0.34)
4th Rank	AltaVista	AltaVista	MSN (0.15)
5th Rank	Excite	MSN	Excite (-0.19)

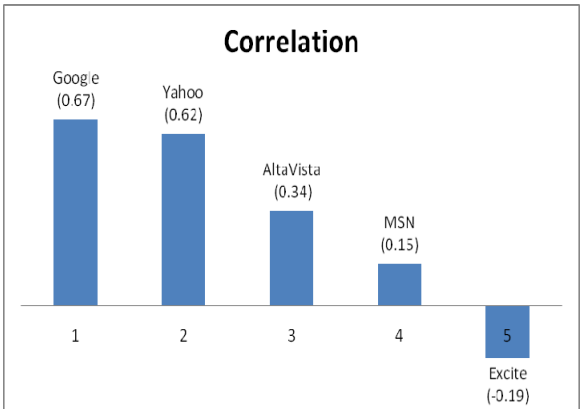
Graph 1 depicts the comparison of five search engines for their Session Duration. The table 3 indicates that Google presents quality results with in limited time at the top of the retrieved list of results whereas the Graph 2.0 denotes the clicks for the purpose of searching relevant results shows 'Yahoo' considers click hits to rank the results in its ranking algorithm.



Graph 1.Session Duration comparison



Graph 2.Click Hits comparison



Graph 3.Correlation comparison

8. Rank Correlation

Ranking is a natural process performed by the information retrieval system which associates numerical scores to the retrieved results for measuring the occurrence of relevancy in each result. Rank correlation denotes to the suite of statistical methods to examine the degree to which two rankings are correlated, correlation refers the closeness of the values of one ranking to be in the identical order of the values of the other ranking.

We examine the ranked list of 10 results ranked by decreasing session duration also known as reference ranking. X = (1,2,3,4,5,6,7,8,9,10) is supposed to be the reference ranking placing d^{th} result at rank d. In this way webpage d follows d-1 results and precedes 25-d results. A different ranking X'=(8,2,3,4,1,6,7,5,9,10) is the ranking presented to the end user ranked according to

decreasing number of hits occurred previously on the results. We utilized the Kendall's tau (τ) coefficient to calculate the correlation when the results are arranged by using the data furnished in the table 3.

It is obvious from table 4 that Google holds the probability to have maximum identical pairs [6] in X and X' that indicates towards the fact that Google considers session duration and click hits during the ranking process of results. if results ranked according session time are still existing in top ranking when the results are arranged according to number of click hits then it should be considered that the top ranked results are relevant because two user action based ranking process suggest approximately same results.

9. Discussion and Conclusion:

We conclude that most widely used measures in information retrieval evaluation system such as precision and recall do not seem to be suitable to the W3 environment. We found users feedback provided an automatic evaluative approach capable to rank the results efficiently, utilizing the real time user interaction data to directly predict the search engine quality without creating retrieval evaluation metrics. This makes it possible to evaluate retrieval performance more economically and user-centered.

The significance of this study is that capturing user actions as implicit relevance judgments can attain significant advantages in evaluating search engines performance by reducing the cost and time of evaluating obtained results.

In this paper, Google was found to be out performer than other search engines on the basis of session duration time. This indicates that users seek information with Google in lesser amount of time. Correlation shows that Google considers previous users actions in its ranking algorithm and holds approximately common top ranked results when the results are arranged by session duration and clicks hits separately.

The results achieved in the experiment showed that Yahoo is oscillating at first ranking when the results are sorted with click hits indicates towards the fact that users devote time with keen interest on the retrieved results. This may be because of Yahoo's incorporation of Click Hit in its ranking functionality.

There was no significant effect of user interaction found over Excite and MSN that shows that few search engines still believe in traditional human evaluation.

References

[1] Joachims T. "Optimizing search engines using clickthrough data", Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD). 2002, pp.133-142.

[2] Brian Amento, Loren Terveen., "Does Authority Mean Quality? Predicting Expert Quality Ratings of Web Documents", SIGIR 2000, pp.296-303.

[3] Donald Metzler, Chappelle, Ya Zhang, Grinspan., "Expected Reciprocal Rank for Graded Relevance", ACM. 2009, pp.621-630.

[4] G.V. Cormack, Clarke, Buttcher, "Reciprocal Rank Fusion Outperforms Condorcet and individual Rank Learning Methods", SIGIR,. 2009, pp.758-759.

[5] Emine Yilmaz, Milad Shokouhi, Nick Craswell, Stephen Robertson "Incorporating user behavior information in IR evaluation", SIGIR, 2009, pp.57-62.

[6] Bernard J. Jansen, Amanda Spink, Chris Blakely, Sherry Koshman "Defining a Session on Web Search Engines", Journal of the American Society for Information Science and Technology, 2008, pp.862-871.

[7] Massimo Melucci "Recent On Rank Correlation in Information Retrieval Evaluation", SIGIR Forum, 2007, pp.18-33.

Sanjay Kumar Dwivedi Head & Reader, Department of computer science at Baba Saheb Bhimrao Ambedkar University, Lucknow 226025 (U.P.) India (A Central University). His research interest is in Artificial Intelligence, Data Mining, Search Engines and e-Governance. He has published number of research papers in journals and conferences both.

Rajesh Kumar Goutam Research Scholar in Department of computer Science at Baba Saheb Bhimrao Ambedkar University, Lucknow - 226025 (U.P.) India (A Central University). His research interest is in Artificial Intelligence, Data Mining, Search Engines, web technology and e-Governance.