Communications on Applied Electronics (CAE) – ISSN : 2394-4714
Foundation of Computer Science FCS, New York, USA
Volume 2 – No.7, August 2015 – www.caeaccess.org

A New Criterion for Evaluating News Search Systems

Mohammad Ubaidullah Bokhari
Chairman, Department of Computer Science, AMU.
Aligarh (India)

Mohd. Kashif Adhami
Research Scholar, Department of Computer Science, AMU.
Aligarh (India)

ABSTRACT
Measuring the effectiveness of web search engines had been widely studied for the past fifteen years and different methods have been proposed by the researchers. These studies helps in identifying the most effective search engine and are useful for both users at the personal level and search engine vendors at the business level. So in this paper, first we extensively review traditional web search evaluation methods under four major categories and then discuss the urge for news search evaluation. We discuss possible criteria and quality measures foe evaluating web-based news search systems. And finally we evaluate four news search systems under a new criterion-information richness, i. e., extracting the useful contents from search result record (SRR).

General Terms
Web Search Engines, Metasearch Engines, News Articles.

Keywords
News Search Engines, Search Result Records, Time-Sensitive Ranking.

1. INTRODUCTION
In the context on evaluation of information retrieval systems, lots of studies have been made for evaluating web search engines but less work has been done for news search evaluation. Many kinds of information are sensitive to time, especially in the case of news. The value of news article depends much on recency, i.e., the time of publication. Presently, growing number of people are reading news online which is mostly free and easily accessible with a web browser [2]. An important advantage in going for online news is that one can obtain recent news as well as old news that may not be readily available from newspapers. News articles can be accessed as soon as they are posted. Also, depending on the search capability of a news web site, users can readily locate news items that are of interest to them. However, there are many news organizations in the world and there are also many specialized news sites, such as those finance, sports, entertainment and local (catering to local communities). A news article of interest to a person may be posted in a newspaper web site unknown to her/him. To solve this problem, a number of news search engines have been created that allow users to search news articles from a number of news organizations around the world from a single search system. With the comprehensive list of news search engines it becomes necessary to measure the quality of these news search systems which can help news search users to decide-*Which news search engine to prefer?*.

There is a long list of news search/metasearch engines available currently on the web. Some of the popular ones are:

Google News- Launched originally in 2002. It aggregates recent content from top sites around the world. It uses the crawling mechanism for searching across thousands of news sources around the web. The service covers all the stories that were published recently and is available in tens of language so far. It features major topics and incorporate a story spotlight. Users can customize their homepage to elucidate latest headlines from interesting categories, in other words, provide the ability to browse categories of news where headlines are automatically assembled.

Yahoo News- it was founded in the mid 90's as the part of the Yahoo! Web portal. It is recognized as worldwide content aggregator by reputed services such as Reuters, Associated Press, Fox, BBC and more. It also uses the crawling mechanism for searching the web. It covers several topics, including sports, business, entertainment, tech, politics, science and health.

Bing News- it is a news aggregator-a part of the Microsoft's Bing project. It merges articles from multiple reliable sources, such as Reuters, Washington Pots, AP and NewYork Times. It is categorized into major sections like: Top Stories, World, Business, Politics, Sports, Entertainment, Sci/Tech and Health.

Newslookup- It can be called as news search engine, news headline, news feed and news services provider established in 2000. It also uses crawling mechanism and search several thousand news media sites providing latest run down of headlines by region, topic or person and supports configurable filtered search results.

Likewise there is a long list of news search systems and presently it becomes essential to have some news search evaluation studies for judging the quality of these systems.

2. NEWS SEARCH ENGINE TECHNOLOGIES
News search systems were constructed using conventional web crawler based technique or meta-search engine technology.

2.1 Web Crawler Based Technique
Crawler-based search engines use automated software agents called crawlers. These crawlers visit a web site, read the information on the actual site, read the site's meta tags and also follow the links that the site connects to performing indexing on all linked web sites as well. The crawler returns all the information back to the central depository, where the data is indexed. The crawler will periodically return to the sites to check for any information that has changed and its frequency is checked by the administrators of the search engine. Some human-powered search engines rely on humans

to submit information that is subsequently indexed and catalogued. Only information that is submitted is put into the index. In both the above cases, when a user query a search engine to locate information, actually searching is done through index created by the search engine. The user does not actually search the web. These indices are giant databases of information that is collected and stored and subsequently searched. If the index hasn't updated, sometimes search engine can return results consisting dead links since the search results are based on the index, if the index hasn't been updated since a web page became invalid the search engine treats the page as still an active link even though it no longer is. It will remain that way until the index is updated.

2.2 Metasearch Technology

A metasearch engine is a system that afford unified access to multiple existing search engines. Even though search engines and metasearch engines are built using very different techniques [3], from user's perspective there is essentially no difference between using any amongst the two. When a metasearch engine receives a query from a user, it passes the query to multiple existing search engines called *component search engines* and then it merges the results returned by these search engines and displays the combined results to the user. Technical issues for building metasearch engines have been discussed in [4,5…]. A metasearch engine makes it easy for a user to search several search engines concurrently while feeding just one query. A simple metasearch engine consists of a *user interface* for users to enter queries, a *search engine connection component* for submitting queries to its component search engines and obtaining result pages from them through programs, a *result extraction component* for extracting the search result records (SRRs) from the returned result pages and a *result merging component* for aggregating the results [5].

If a metasearch engine utilize a large number of search engines, then a *search engine selection component* is obliged. This component impel which search engines are likely to hold matching results for any given user query so that only these search engines are used for this query. Search engine selection is needed for efficiency. Sending a query to useless search engines will account for serious inefficiencies like huge network traffic caused by dispatching unwanted results and the waste of system resources for evaluating the query. Moreover, the metasearch engine may be overwhelmed by the irrelevant results returned by the useless search engines. In [4], for a news metasearch engine, a new component- the *publication time extraction component* was needed to determine the time and publication date of each retrieved news item. Publication time was required to perform time-sensitive ranking of retrieved results. Normally, among the relevant news articles, more fresh ones should be ranked ahead of stale ones.

3. WEB SEARCH EVALUATION

3.1 Overview of Web Search Evaluation Methods

We have categorized the web search evaluation methods into five sub-sections below.

3.1.1 Evaluation Based on Relevancy

Most of the web search evaluation studies are relevance based. Relevance based web search evaluation studies are similar to the traditional cranfield model based evaluation [6]. In mid nineties researchers developed different evaluation parameters [7, 8, 9] for selecting better search engines. All these early studies employed small test query sets therefore were criticized, also researchers themselves made the relevance judgements and reported results were only based on a limited account of trial runs. Research methodologies in late nineties included statistical techniques [10,11,12] and were more thorough than previous studies.

The most trivial effectiveness measures used for the evaluation purpose were precision and recall. Majority of the web search evaluation studies used precision for search engine evaluation as absolute recall cannot be measured for the web. Hawking et al [10] compared precision of top 20 results of five commercial search engines with precision at 20 of six TREC systems. Their experiments showed that the TREC systems outperformed the web search engines. Ljosland [13] conferred the evaluation of three search engines Altavista, AllTheWeb ang Google over 12 queries. They did not deduced any statistically significant difference amongst the precisions of search engines when reckoning only relevant documents but got statistically significant difference while concidering partially relevant documents in account. Chu and Rosenthal [7] evaluated three web search engines namely AltaVista, Excite and Lycos and used 10 queries to observe precision with a three-level relevance score, viz. relevant, somewhat relevant and irrelevant, for the upper 10 links. Meghabghad and Meghabghad [15] analyzed the effectiveness of five web search engines namely Yahoo, WebCrawler, InfoSeek, Excite and Lycos. Their results elucidated that Yahoo had the highest ratio amongst the five for both original and refined queries. Leighton and Srivastava [12] tested five search engines on the basis of precision on the first 20 results returned for 15 queries. Ding and Marchionini [8] evaluated precision and result overlap among three search engines namely InfoSeek, Lycos and OpenText using five queries.

The work by Clarke and Willett [14] is an example of recall based evaluation. They developed a method for comparing the recall of three sets of searches. These searches were conducted on the diverse indexed collections of AltaVista, Excite and Lycos. They also focused on critical evaluation of earlier research. Although various measures of recall and precision have been efficacious in information retrieval, there have been a number of attempts to devise measures of efficiency that aggregate recall and precision [16]. Eguchi et al. [17] gave an overview of the web retrieval task at the third NTCIR workshop. Although Trec is considered as a standard test-enviroment, the methodologies used and results obtained need to be regularly analyzed. The off-site link density of the real web is rated too low by the Web TREC collection and these collections are static therefore they do not reflect the high instability of web pages. This fact was pointed out by Gurrin and Smeaton [17] and thus web TREC methodologies can't serve as an important model for measuring the real effectiveness of search engines.

Precision measures generally used by the studies on the web search performance do not dubbed links between web documents, but Samalpasis et al. [18] devised a method to measure relevance in hyperlinked networks of documents. They proposed a relevance metric, named Relative Distance Relevance (RDR), which was computed using the matrix of distances of document nodes in the network and on the binary relevance of these documents. Shang and Li [19] prompted a general approach for statistically evaluating precision of

search engines on the web. Search engine evaluation was done in two steps. First, relevance scores of hits were computed for each search engine and then search engines were ranked on the basis of relevance scores. They experimented with two query seta for six search engines- AltaVista, Fast, Google, Go, iWon and NorthernLight. The six search engines performed differently under different scoring methods and search modes. In another work [20], Li and Shang proposed a new statistical method for evaluating precision performance of search engines based on sample queries. Three search engines- AltaVista, Google and InfoSeek were compared using two query sets. The query sets were derived from the domain of distributed and parallel processing. They experimentally showed that their results of relevance judgment for three level scoring algorithms were consistent with the result of manual method.

3.1.2 Evaluation Based on Ranking

Ranking of documents present in the search results is an important criterion for web search evaluation. When a user feed query to the search engine, large number of results were returned. It is not possible for the users to see the whole set of results and they are eminently dependent on the ordering of the results returned by the search engine in obtaining desired information. For example, if a pure relevant is placed in the last hundred positions, out of the thousands of results returned, it is indeed unlikely that the user will visit the page as most users do not see beyond first ten or twenty results [21,22]. Therefore placement of the relevant pages in higher positions is required. The significance of placement of the relevant pages in higher position was taken into account by the measures such as precision at top 10 or 20, but the relative intermediary ranking of documents within top 10 or 20 is also important. The researchers can only make an analysis of the ranked results of the search engines for the purpose of search engine evaluation because the ranking algorithm are kept secret due to completion amongst the search engines and also to avert misuse by the mischievous users.

The general method of evaluating rankings is through human judgment. Su et al. [23] asked the judges to select and rank the five most relevant items out of the first twenty results returned. They analyzed four search engines- AltaVista, InfoSeek, Lycos and Open Text. They found Lycos's performance to be the best one. Courtois and Berry [24] judged the ranking of the search engines using three criteria namely-at least one occurance of all search terms, at least one occurance of all search terms appearing as a contiguous phrase and at least one occurance of all search terms present within title, headers or meta-tags. Overall, they found Excite had the best ranking for the top 20 documents. Gwizdka and Chignell [25] developed differential precision to measure the quality of ranking produced by search engines. Hawking et al. [26] evaluated the effectiveness of 20 search engines on a range of measures and one of the measure used was reciprocal rank of first relevant document- a measure closely related to ranking. Chowdhury and Soboroff [27] also used the same reciprocal rank measure for evaluating search effectiveness. Singhal and Kaszkiel [28] analyzed a well-performing TREC system with four search engines namely Excite, Google, Lycos and AltaVista. Contived the relevance judgment themselves and found that commercial web seach engines are better than a state-of-the-art keyword-based document-ranking algorithm. Vaughan [29] experimented with three commercial search engines-Google, AltaVista and Teoma with 24 participants using four queries. They got the first 10 links from each search engine ranked by these participants. The human rankings were compared with search engine rankings Google was found to be the best with highest average correlation.

3.1.3 Evaluation Based on User Satisfaction

Large number of studies had been made to evaluate web search engines using real user's point of view. It had been perceived that a more concrete evaluation of system performance can be made, if real users make the relevance judgments. Such evaluation process obviously might be affected by number of factors like the context of the query, time spent and emotional state of the user. Su [31] conferred the need for an evaluation methodology for interactive retrieval systems in realistic situations. She proposed the criteria for the evaluation of such systems from a user perspective. She inspected whether a single best measure of a system performance could be achieved by correlating twenty measures of retrieval performance with user's overall annotation and got value of search results as a whole to be the single best measure. This measure was dependent on the user's satisfaction with retrieved results as a whole. Before this, another study by Su [30] combined the objective measures of system performance with subjective measures of seeming satisfaction to produce some contradictory results. In [30] Su found that users with a low expectation of getting information communicated high satisfaction with a set of low precision results.

Spink [32] evaluated a meta-search engine, Inquirus, using user-centered approach. In her work, twenty-two annotators were asked to search their own respective information tasks and then, further rate the top 20 web documents on relevance on a four-point scale. Another work from Nasios et al. [33] analyzed the search engine capabilities with contemplating potential differences in the background of end users. In their study, web search engines were examined according to the search results' ability to satisfy an easily pleased user or hard to please user. They found that overall, AltaVista and HotBot performed best. Nahl [34] studied the nivice's search experience and concluded that usefulness of search results affects user's perception of ease of use and thus contributed to the importance of search engines. Chang and Wu [35] compiled a list of all search engine characteristics defined by previous researchers and argued whether they could be divided into two categories of factors, which would have different effect on the user's intention to use search engines. They described the attracting factors as the basic attributes that are "necessary to have" and retaining factors as additional and advanced attributes that are "nice to have" for users. Experimentally, they inferred that users do not perceive the several features correspondingly. The work from Johnson et al. [36] discussed the viability of the use of user-satisfaction as a multidimentional evaluative construct of search engines. They gave a conceptual framework for the evaluation search engines from a user perspective. Beg [37] measured the "satisfaction" a user gets when the search result is presented to him. For this principle, the response of the search user to the returned results presented before him is monitored and the feedback of the user is characterized by a seven component vector. After the feedback recovery, a weighted sum for each document selected by the user is computed on the basis of these seven components. Sorting the documents on the descending values of weighted sum will yield a sequence.

This sequence is compared with the full list sequence in which the documents were initially short-listed and then spearman rank order correlation coefficient is computed. The performance of the search engines can be evaluated by repeating this procedure for a representative set of queries and averaging the correlation coefficient value.

3.1.4 Automatic Evaluation

The work from Soboroff et al. [38] was the first in which the relevance judgment was performed automatically. They replaced human relevance judgments with a number of randomly selected "pseudo-relevant" documents from pool generated in the TREC environment. Their work showed that the system rankings correlated positively and significantly to the real TREC rankings. Wu and Crestani [39] proposed the reference count method for automatic ranking of retrieval systems. For each query, in their method, they first acknowledged the list of documents returned by a retrieval system, and then noted references for each document of the list. They also analyzed the effectiveness of various versions of their method with the method used in [38] and showed that their method outperformed the random selection method in automatic ranking of retrieval systems.

In [40], Aslam and Savell noted that both the random selection method [38] and the reference count method [39] was not sufficient at predicting the performance of top performing systems and they presented a hypothesis that both methods were suffering from a tyranny of the masses effect [] which states that better systems were doing something different from the more generic systems in the competition and in the absence of actual relevance judgments were being suffered for this truth. In another study [41], Aslam et al. provided a sampling approach in order to replace human relevance judgment. In [42], Hersh and Kim described the variations on experiments by Vorhees [1], Soboroff et al. [38] and Aslam et al. [41]. The analysis was done by them using the test collections and submitted runs from 2006 medical tasks [42] and ImageCLEF 2005. They concluded that use of other approaches were needed if human judgments were to be replaced. Amitay et al. [43] used a list of (onTopic terms) supposed to be relevant and separated the list of terms believed to be irrelevant (offTopic terms) to a particular query. Their results were found to be consistent with the human-based results. Shang and Li used a large automatic test design to evaluate six popular search engines using 3000 queries from two different domains [19]. They computed relevance scores using three different algorithms and elucidated statistical comparisons of the ranking. Can et al. [44] also proposed an automatic performance evaluation method named AWSEEM. In this work, thet proposed to replace human-based relevance annotations with a set of automatically generated relevance judgments. Researchers showed that their results were consistent with human-based evaluations. In another study [45], Nuray and Can described the automatic ranking of retrieval systems in imperfect environment. Beitzel et al. [46] too opted a method similar to the AWSEEM method and used the Open Directory Project (ODP) [47] categories to determine relevant documents for the evaluation of web search engines. Chowdhury and Soboroff proposed method for automatically comparing search engine performance based on how they rank the known item search result [27]. They constructed a large number of query document pairs and found the rank of the paired documents in response to a query and scored each ranked list using the reciprocal rank of the target document. The overall score was given by the mean reciprocal rank over all paired documents. Mowshowitz and Kawaguchi [48] used the overlap of URLs of the matching pages for evaluating the performance of 12 search engines with 12 queries and the top 20 links. They used bias to evaluate the performance of the search engines. Nuray and Can [49] also discussed some new methods for automatic ranking of retrieval systems. They merged the retrieval results of multiple systems using various data fusion methods. They considered the top-ranked documents as the pseudo-relevant documents and used them to evaluate and rank systems. They experimented with TREC data and found that their evaluations were strongly correlated and statistically significant with human-based evaluations of the same systems. Sharma and Jansen [50] used implicit feedback in the real-time for the development of the evaluation system. Ali and Beg [51] presented an approach for automatic evaluation without human involvemet for a large number of queries. Their system initially learns ranking rules on the basis of implicit user feedback using rough set theory for a small number of queries. These ranking rules were then employed for fusion of different evaluation techniques for the evaluation of search systems with large number of queries for achieving more realistic evaluation.

3.2 Lack of News Search Evaluation

Although a vast number of evaluation studies are present for the web search engine evaluation as discussed in the previous section still researchers lack news search evaluation studies. Lots of news search engines and news meta-search engines are presently available on the internet and web users having news search intent are frequently turning to these online news sources. So there is a need for studies regarding the effectiveness of web-based news search systems. Few attempts were made before, for evaluating news search systems.

Rasolofo et al. [52] constructed a realistic current news test collection using the results obtained from 15 news web sites. These web sites included ABC News, BBC, AllAfrica and others. They used 107 topical queries for evaluation. They showed that high retrieval effectiveness can be achieved in realistic news metasearch application using low cost merging methods, even if the primary servers give very little information. They compared various result merging strategies using a news test collection but did not actually evaluated the various news sources. Only precision-based evaluation was done for individual news servers, which is of our interest as far as news search evaluation is concerned. They computed the average precision over queries for which the news server returned at least one document and again computed the same for servers returning at least one relevant document. For the relevance assessment five research assistants were recruited. Judging was carried out using a web browser. They implemented a Perl CGI script which presented a merged list of all the results pages for a particular query. When analyzing the mean performance provided by these news servers, the average retrieval performance was computed and the standard deviation around this mean was relatively high indicating that the performance difference between best server and the worst varies largely and this shows the retrieval performance varies a greater extent in this kind of online news services.

We found the work from Liu et al. [53] was the only one explicitly related to evaluation of news search systems. They developed AllInOneNews- a news search system based on

metasearch engine technology. This news metasearch engine directly connects to the search engines of individual newspapers and news sites. Also they provided several evaluation criteria to measure the quality of news search systems. In particular they proposed a new and novel scheme to measure and compare the retrieval effectiveness of news search/metasearch engines for time-sensitive information and then compared three news search/metasearch engines namely Google News, Mamma News and their own developed system-AllInOneNews using their developed criteria. They identified strengths and weaknesses of thes systems in processing different types of queries.

Table 1: Analysis of various web search evaluation methods

Categories	Basis of evaluation	Evaluation measures	Significance	Drawbacks
Relevance	Search results should be topically related	Precision, recall, relevance scores	Most widely used	Relevance judgment needed for relevance- a highly debatable term
Ranking	Document present in the search results should be ranked	Rank order correlation coefficient	Search engine that places relevant position in higher positions get more score.	Correlation of search results to human ranking needed.
User satisfaction	User's satisfaction with the search results.	User's rating, overall user satisfaction.	More realistic evaluation.	Search user feedback needed, subjective.
Automatic	Relevance judgments are automatically generated.	Precision, precision@k relevance scores, mean reciprocal rank.	Extensible, can be performed regularly.	Lacks human intelligence, in general.

3.3 Possible Quality Measures/Criteria for Evaluating News Search Systems.

Due to dynamic nature of the news, it is necessary for news search systems to retrieve not only the relevant document but also the *fresh* relevant document. It means the retrieval system must take recency of the web page also into account so that news search users can get fresh news reports. So time-sensitive effectiveness evaluation is important in the case of news search systems. Liu et al. [53] proposed five criteria for measuring the quality of news search systems namely Traditional Effectiveness, Time-sensitive Effectiveness, Redundancy, Diversity and Information Richness.

4. SEARCH RESULT RECORD (SRR) BASED EVALUATION OF NEWS SEARCH SYSTEMS

The information returned from the search result record (SRR) can be considered as, an important criteria for evaluating the effectiveness of news search systems. The information that may be contained in an SRR includes news source's name, news article size, the URL to the full document, the Publication time/date, a short excerpt/summary of the full document etc. These pieces of information as a whole can be termed as information richness of the SRRs. So in this section we will discuss the information richness-based evaluation of four news search/metasearch systems. Table 2 gives the description of parameters used in the evaluation for describing the SRRs. Although Liu et al. [53] evaluated before, three news search systems based on SRRs but here we included some more parameters for describing SRRs and our evaluation is more extensive as far as information richness based evaluation is concerned, Liu et al. have also used four other criteria.

The analysis of information richness based evaluation is given in table 3. First we discuss briefly these results and then graphically elucidate the overall outcome. On 30//7/2015 we fed the query 'ashes 2015' in four news search/metasearch engines namely-Google News, Yahoo News, Bing News and Newslookup. We have taken few assumptions necessary for describing the results. Response time (RT) is the total time taken by the search engine in returning the retrieved results in other words we can say the total time taken for processing the query. Google News and Newslookup were the two systems which included RT in their SRRs. For Result sorting criteria (RSC) the former three news search engines returned single options i. e. to sort, by relevance or date (recency) but the last one- Newslookup provided lots of options for relevance as well as date both. We can get the sorted results according to the options given. All the four systems displayed the total no. of results retrieved (N) and except Yahoo News, which displayed it at the bottom, other engines have it at the top position. In the case of date/time, only Bing News doesn't have the date and included only time, others showed both values. We also found that source URL was not mentioned in the case of first three search systems and only Newslookup included it in their SRRs. Also the SRRs included some site description which can help search users in deciding which link to click. More the site description included in the result page, more easy for users to click the intended link accordingly. So we found that Newslookup included more lines, around 5-6, to describe each link as compared to others.
In our work we made some hypothesis to evaluate these four news search systems and our motive behind this is to elucidate two facts:

1) New benchmark criterion, in our case-information richness, can be explored for evaluating the effectiveness of web-based news search systems.
2) New evaluation measures, even-though traditional measures are present for web search evaluation, can be devised for the developed criterion.

For answering the question- *Which news search engine is best ?*. We analyzed our results and drawn some important conclusions. In the case of response time (RT), for the test query we used-'ashes 2015', we found:

RT for Google News : 0.17 sec for retrieving 71,00,000 documents.

RT for Newslookup : 0.451 sec for retrieving 28,401 documents.

RT for Yahoo News : not given , total no. ret. doc.-139.

RT for Bing News : not given, total no. of ret. doc.-26,000

Thus we can say Google News performed best retrieving highest number of documents with less response time than Newslookup.

For overall performance, considering all the seven parameters, we found Newslookup to be best having a positive value for each parameter unlike others.

Table 2: Parameters used for SRR based evaluation

Parameters	**Description**
Response time (RT)	Total time taken for processing the query.
Result sorting criteria (RSC)	If the results are sorted by relevance, recency or any other criteria.
No. of SRRs (N)	Total no. of SRRs displayed on a page.
Date/Time (D/T)	If the SRR indicate the date and time.
Highlighted query terms (HQT)	If the query terms are highlighted in the SRRs.
Source URL (SU)	Whether source URL mentioned in the results.
Site description (SD)	Usefulness or few lines of site's description.

Table 3: information richness based evaluation of four news search systems.

Parameters	**Google News**	**Yahoo News**	**Bing News**	**Newslookup**
RT	Yes	No	No	Yes
RSC	Yes, changeable, single option.	Yes, changeable, single option.	Yes, changeable, single option.	Yes, lots of options like: relevancy<60,30,14,7 days, relevancy<24,48,72 hours, date<24,48,72 days etc.
N	Yes, at top position.	Yes, at the last.	Yes, at the top.	Yes, at the top.
D/T	Yes, both	Yes, both	Only time given.	Yes, both
HQT	Yes	Yes	Yes	Yes
SU	No	No	No	Yes
SD	Low around 2 lines.	Medium, around three lines.	Medium around three lines.	High around 5-6 lines.

5. CONCLUSION AND FUTURE WORK

In this paper we extensively reviewed the traditional web search evaluation methods under four major categories and then discussed the need for exploring new methods for news search evaluation which had been a less focused area. We discussed some possible criteria and evaluation measures for evaluating news search systems and finally we evaluated four news search systems under a new criterion-information richness which means we have extracted information from search result record (SRR) pages and used it for effective evaluation. Our analysis revealed Google News to be the best having low response time and highest number of documents retrieved whereas whewn we cincider the whole seven SRR parameters then Newslookup seemsed to have more positive response.

In future new benchmark criterion can be explored for evaluating the effectiveness of web-based news search systems and more parameters or measures, such as time-sensitive patrameters to measure freshness of retrieved documents, can be devised for the developed criterion.

6. REFERENCES

[1] E. M. Voorhees. Variations in relevance judgments and measurement of retrieval effectiveness. In *Proc. of the 21 st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp. 315-323, 1998.

[2] C. M. Kelly and G. D. Moulin. The web cannibalizes media, Technical report, The forrester group, May, 2002.

[3] C. Yu and W. Meng. Web Search Technology. In *The Internet Encyclopedia* edited by Hossein Bidgoli, Wliey Publishers, pp. 738-753, 2003.

[4] K. L. Liu, W. Meng, J. Qiu, C. Yu, V. Raghavan, Z. Wu, Y. Lu, H. He and H. Zhao. AllInOneNews: Development and Evaluation of a Large-Scale News Metasearch Engine. In *Proc. of SIGMOD'07,* Beijing, China, 2007.

[5] W. Meng, C. Yu and K. L. Liu. Building Efficient and Effective Metasearch Engines. ACM Computing Surveys, 34(1), pp. 48-84, 2002.

[6] C. W. Cleverdon, J. Mills and E. M. Keen. Factors Affectinng the Performance of Indexing Systems, ASLIB, Cranfield Research Project, Vol. 2, pp. 37-59, Bedford, UK, 1966.

[7] H. Chu and M. Rosenthal. Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology. In *Proc. of 59 th American Society for Information Science,* Baltimore: MD, USA, pp. 127-135, 1996.

[8] W. Ding and G. Marchionini. A Comparative Study of Web Search Service Performance. In *Proc. of 59 th Annual Meeting of the American Society for Information Science.* Vol: 33, Baltimore:MD, USA, pp. 136-142, 1996.

[9] H. V. Leighton. Performance of Four World Wide Web Index Services: InfoSeek, Lycos, Webcrawler and WWW Worm, 1996.

[10] D. Hawking, N. Craswell and P. Harman. Results and Challenges in Web Search Evaluation. Comput. Netw. Vol:31, pp. 1321-1330, 1999.

[11] S. Lawrence and C. L. Giles. Searching the World Wide Web. Science, 280: 98-100, 1998.

[12] H. V. Leighton and J. Srivastava. First 20 Precision Among World Wide Web Search Services. J Am Soc Information Science, Vol: 50, pp. 870-881, 1999.

[13] M. Ljosland. Evaluation of Web Search Engines and Search for Better Ranking Algorithms. In *Proc. of SIGIR'99 Workshop on Evaluation of Web Retrieval,* 1999.

[14] J. Clarke and P. Willett. Estimating the Recall Performance of Web Search Engines. In *Proc. of ASLIB'97,* 49(7), pp. 184-189, 1997.

[15] D. B. Meghabghab and G. V. Meghabghab. Information Retrieval in Cyberspace. In *Proc. of the American Society for Information Science(ASIS) Mid-Year Meeting*, San Diego: CA, USA, pp. 224-237, 1996.

[16] C. T. Meadow. Text Information Retrieval Systems. Toronto, Canada: Academic Press: 1992.

[17] C. Gurrin and A. Smeaton. Improving the Evaluation of Web Search Systems. In *Proc. of the 25 th European Conference on IR Research (ECIR '03),* Pisa, Italy, In: Sebastiani F. editor, Lecture Notes in Computer Science, Vol: 2633, pp. 25-40, Springer: New York, USA, 2003.

[18] M. Samalpasis, J. Tait and C. Bloor. Evaluation of Information Seeking Performance in Hypermedia Digital Libraries, Interact. Comput., Vol: 10(3), pp. 269-284, 1998.

[19] Y. Shang and L. Li. Precision Evaluation of Search Engines. World Wide Web, Vol:5(2), pp. 159-173, 2002.

[20] L. Li and Y. Shang. A new statistical method for performance evaluation of search engines. In *Proc. of the 12 th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2000),* pp. 208-215, 2000.

[21] C. Silversien, M. Henzinger, M. Marais and H. Moricz. Analysis of a very large web search engine query log, ACM SIGIR Forum, Vol: 33(1), pp. 6-12, ACM Press: New York, NY, USA, 1999.

[22] A. Spink, S. Ozmutlu, H. C. Ozmutlu and B. J. Jansen. U.S. versus Europran web searching trends, SIGIR Forum, Vol: 36(2), pp. 32-38, 2002.

[23] L. T. Su, H. Chen and X. Dong. Evaluation of web-based search engines from the end-user's perspective: A pilot study. In *Proc. of the 61 st American Society for Information Science, Vol: 35, pp. 348-361,* Pittsburgh: PA, USA, 1998.

[24] M. P. Courtois, M. W. Berry. Results ranking in web search engines. Online: 23(3), pp. 39-46, 1999.

[25] J. Gwizdka and M. Chignell. Towards information retrieval measures for evaluation of web search engines, 1999.

[26] D. Hawking, N. Craswell, P. Bailey and K. Griffiths. Measuring search engine quality. Information Retrieval, Vol:4, pp. 33-59, 2001.

[27] A. Chowdhury and I. Soboroff. Automatic evaluation of world wide web search services. In *Proc. of the 25 th annual international ACM SIGIR conference on research and development in information retrieval,* Tampere, Finland, ACM Press, pp. 421-422, 2002.

[28] A. Singhal and M. Kaszkiel. A case study in web search using TREC algorithms. In *Proc. of the 10 th International World Wide Web Conference,* Hong Kong, pp. 708-716, 2001.

[29] L. Vaughan. New measurements for search engine evaluation proposed and tested. Information Processing and Management, Vol: 34, pp. 557-579, 1998.

[30] L. T. Su. Value of search results as a whole search engines by undergraduate students. In *Proc. of the 62 nd*

American Society for Information Science, Vol: 36, pp. 98-114, Washington DC, USA, 1999.

[31] L. T. Su. A comprehensive and systematic model of user evaluation of web search engines: II An evaluation by undergraduates. Journal of American Society Information Science Technology, Vol: 54(13), pp. 1193-1223, 2003.

[32] A. Spink. A user centered approach to evaluating human interaction with web search engines: an exploratory study, Information Processing and Management, Vol: 38(3), pp. 410-426, 2002.

[33] Y. Nasios, G. Korinthios and Y. Despotopoulus. Evaluation of search engines. Report undertaken by the National Technical University of Athens on behalf of the European Commission and Project PIPER, 1998.

[34] D. Nahl. Ethnography of novices' first use of web search engines: affective control in cognitive processing. Internet Reference Services Quarterly, Vol: 3(2), pp. 51-72, 1998.

[35] L. Li and Y. Shang. A new statistical method for performance evaluation of search engines. In *Proc. of the 12 th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2000),* pp. 208-215, 2000.

[36] F. C. Johnson, J. R. Griffiths and R. J. Hartley. Devise: A framework for the evaluation of internet search engines. Library and Information Commission Research Report 100, 2001.

[37] M. M. S. Beg. A subjective measure of web search quality. International Journal of Information Science, Elsevier, Vol: 169(3-4), pp. 365-381, 2005.

[38] I. Soboroff, C. Nicholas and P. Cahan. Ranking retrieval systems without relevance judgments. In *Proc. of the 24 th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New Orleans, LA, USA, pp. 66-73, 2001.

[39] S. Wu and F. Crestani. Methods for ranking information retrieval systems without relevance judgments. In *Proc. of the ACM Symposium on Applied Computing*, Melbourne, Florida, USA, pp. 811-816, 2003.

[40] J. A. Aslam and R. Savell. On the effectiveness of evaluating retrieval systems in the absence of relevance judgments. In *Proc. of the 26 th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Toronto, Canada, pp. 361-362, 2003.

[41] J. A. Aslam, V. Pavlu and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *Proc. of the 29 th Annual International ACM SIGIR Conference on Research and Developmant in Information Retrieval,* Seattle, WA, USA, pp. 541-548, 2006.

[42] W. Hersh and E. Kim. The impact of relevance judgments and data fusion on results of image retrieval test collections. In *Proc. of the 2 nd MUSCLE/ImageCLEF Workshop on Image and Video Retrieval Evaluation*, Alicante, Spain, pp. 29-38, 2006.

[43] E. Amitay, D. Carmel, R. Lempel and A. Soffer. A scaling IR-system evaluation using term relevance sets. In *Proc. of the 27 th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Sheffield, UK, pp. 10-17, 2004.

[44] F. Can, R. Nuray and A. B. Sevdik. Automatic performance evaluation of web search engines. Information Processing and Management, Vol: 40(3), pp. 495-514, 2004.

[45] R. Nuray and F. Can. Automatic ranking of retrieval systems in imperfect environments. In *Proc. of the 26 th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Toronto, Canada, pp. 379-380, 2003.

[46] S. M. Beitzel, E. C. Jensen, A. Chowdhury and D. Grossman. Using titles and category names from editor-driven taxonomies for automatic evaluation. In *Proc. of the 12 th International Conference on Information and Knowledge Management,* New Orleans, LA, USA, pp. 17-23, 2003.

[47] Open Directory Project. http://dmoz.org/.

[48] A. Mowshwitz and A. Kawaguchi. Assessing bias in search engines. Information Processing and Management, Vol: 35(2), pp. 141-156, 2002.

[49] R. Nuray and F. Can. Automatic ranking of information retrieval systems using data fusion. Information Processing and Management, Vol: 42(3), pp. 595-614, 2006.

[50] H. Sharma and B. J. Jansen. Automatic evaluation of search engine performance via implicit user feedback. In *Proc. of the 28 th annual international ACM SIGIR conference on research and development in information retrieval,* Salvador, Brazil, pp. 649-650, 2005.

[51] R. Ali and M. M. S. Beg. Automatic performance evaluation of web search systems using rough set based rank aggregation. In *Proc. of the First International Conference on Intelligent Human Computer Interaction 2009 (IHCI 2009)*, Springer (India) Publisher: Allahabad, India, pp. 344-358, 2009.

[52] Y. Rasolofo, D. Hawking and J. Savoy. Result merging strategies for a current news metasearcher. Information Processing and Management, Vol: 39, pp. 581-609, 2002.

[53] K. L. Liu, W. Meng, J. Qiu, C. Yu, V. Raghavan, Z. Wu, Y. Lu, H. He and H. Zhao. AllInOne News: Development and evaluation of a large-scale news metasearch engine. In *Proc. of SIGMOD'07,* Beijing, China, 2006.