Retrieval Effectiveness of News Search Engines: A Theoretical Framework

Mohammad Ubaidullah Bokhari
Department of Computer Science
Aligarh Muslim University
Aligarh

Mohd. Kashif Adhami
Department of Computer Science
Aligarh Muslim University
Aligarh

ABSTRACT
News search has now become an important internet activity as users are switching from hard copies to online news reading. Many modern news search engines like: Google News or Bing News are available for this purpose. We propose a theoretical framework for evaluating the retrieval effectiveness of news search systems. The framework exploits supervised machine learning approach for evaluating therefore we performed retrieval effectiveness tests on a small data set consisting relevancy features- Tfidf and Latent Semantic Indexing (LSI) as well as freshness feature-publication time, extracted from 1120 query-document pairs collected from search results of Google News, to evaluate the performance of various machine learned learning to rank algorithms on NDCG and ERR metric at different cut-offs. The motive behind this work is to conduct large-scale retrieval effectiveness studies for news search engines.

General Terms
Feature engineering, machine learning, retrieval effectiveness tests

Keywords
Learning to rank algorithms, ranking model, News search engine, News search quality.

1. INTRODUCTION
Using traditional web search engine for finding news is now days a bad practice since news search engines have evolved in recent years. You can clearly observe the difference in news search results when you type a same query at same time on two different search engines one being a traditional web search engine like *'google'* or *'bing'* and other, a fullfledged news search engine like *'google news'* or *'bing news'*. News search engines provide exceptionally good results for current event searching or breaking news because they crawl only news sites and revisit these sites several times per day. Thus results are usually focused and timely. Description of some of the modern news search engines are as follows:

Google News: is a free news aggregator furnished and operated by *Google.* It aggregates news from thousands of news websites. The beta version was launched in September 2002 and officially released in January 2006. Google News implemented searching, and the choice of sorting the results by date and time of publishing or grouping them (and also grouping without searching). Its result page design has been changed from June 2017 onwards.

Traditionally, news readers first pick a publication and then look for headlines that interest them. *Google News* do things a little differently, with the goal of offering the readers more personalized options and a wider variety of perspectives from which to choose. Google News offers links to several articles on every story, so one can first decide what subject interests the user and then select which publishers' accounts of each story the user like to read. Click on the headline that interests him and he'll go directly to the site which published that story. Articles from Google News are selected and ranked by computers that evaluate, among other things, how often and on what sites a story appears online. It also rank based on certain characteristics of news content such as freshness, location, relevance and diversity. As a result, stories are sorted without regard to political viewpoint or ideology and one can choose from a wide variety of perspectives on any given story. According to Google News will continue to improve by adding sources, fine-tuning the technology and providing it to readers in even more regions.

Bing News: is also a news aggregator (previously *Live Search News*) mechanized by artificial intelligence–is a part of Microsoft's Bing search engine which processes billions of global searches. Operating in the United States and other international markets Bing News displays the latest news stories on Bing.com/News on desktop and mobile, the Bing Search app, and through enterprise streams such as the Outlook News Connector, PowerBI and Bing for business. It was launched on June 2009. Bing News also aggregates the most recent news articles in response to user search queries algorithmically on Bing.com.

News headlines from various sources are aggregated and categorized into sections for users to browse, which include most read, trending, and breaking news stories as well as category-specific articles in areas such as business, politics, sports, science, tech and entertainment. The Bing News page also displays special events of national or global interest such as the U.S. presidential elections, Olympics, and award shows.

Depending on the user's location, localized news. Multimedia content are also incorporated on the news pages, including images and videos with smart-motion thumbnails similar to Bing Videos.

Bing News also allow users to type in a search term to browse through an archive of news articles relevant to the search query. In addition, users may refine their results by location and category, or search with an alternative related search term. RSS support was added on April 24, 2008, providing support for subscription to a specific news category or search results. In March 2011 Microsoft added Twitter"tweets" to its news results.

Yahoo News: originated as an internet-based news aggregator by Yahoo!. Articles originally came from news services such as the Associated Press, Reuters, Fox News, Al Jazeera, ABC News, *USA Today*, CNN, BBC News, etc.

In 2001, Yahoo! News launched the first "most-emailed" page on the web. It was well-received as an innovative idea, expanding people's understanding of the impact that online news sources have on news consumption. Yahoo allowed comments for news articles until December 19, 2006, when commentary was disabled. Comments were re-enabled on March 2, 2010. Comments were temporarily disabled between December 10, 2011, and December 15, 2011, due to glitches.

By 2011, Yahoo had expanded its focus to include original content, as part of its plans to become a major media organization. Veteran journalists (including Walter Shapiro and Virginia Heffernan) were hired, while the website had a correspondent in the White House press corps for the first time in February 2012. An Amazon-owned marketing data collection company (Alexa) claimed Yahoo! News one of the world's top news sites, at this point.

Newslookup: Newslookup.com is a news search engine, news headline, news feed and news services provider established in 2000, by Michael Kynast. The search engine crawls several thousand news media sites providing time based live run down of headlines by region, topic or person and provides configurable filtered search results.

Configurable options at search time:Search by media type: Newspapers, Television, Radio, Internet, Search by source region of news company, Limit results by html document parts, such as html meta keywords, meta description, document title and document body, Search by news site, Boolean query language support, Phrase support, Cached copies of crawled documented when allowed by robots.txt and meta robots, Results sorting by relevance and date, Group results by site or non-grouped.

So there is a need to evaluate the retrieval effectiveness of theses news search systems because lots of web search evaluation studies[1-8] have been done but news search evaluation studies are almost negligible [9-11].

2. OVERVIEW OF THE FRAMEWORK

Our proposed theoretical framework for conducting retrieval effectiveness tests on news search systems is shown in Fig 1. For the purpose of discussion, we broadly divide our framework into four modules., namely- *Result Collection*, *Feature Engineering*, *Supervised Learning* and *Quality Measurement.*

Result Collection: To collect news results, breaking/latest news queries need to be entered into the news search engine to be investigated. To compare various news search engines, same query should be entered at the same time to the multiple contestant news search engines. This can be done either manually which is candid but time consuming or can be done automatically via socket programming. Lewandowski and Sunkler [12] developed a tool for collecting search engine results. Collecting results means you have to save SERPs of all the news search engines going to participate in retrieval effectiveness test.

Feature Engineering: In the data file, which serves as input to the machine learning algorithm, each row corresponds to a query-url pair. The first column is relevance label of the pair, the second column is query id and the following columns are features. The larger value the relevance label has, the more relevant the query-url pair is. An instance of MSLR-WEB10K dataset from Microsoft learning to rank datasets website is shown in Fig. 1. For building ranking models for news search results, relevancy features are needed together with freshness features. Other important features can be derived from the urls returned such as pagerank of the news document or news website rank etc.

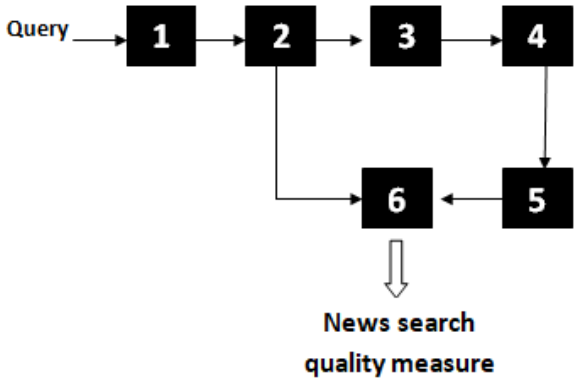
```
0 qid:1 1:3 2:0 3:2 4:2 ... 135:0 136:0

2 qid:1 1:3 2:3 3:0 4:0 ... 135:0 136:0
```

Fig. 1: An instance of MSLR-WEB10K dataset

Supervised Learning: The data file described above serves as input to the machine learning algorithm. In supervised learning the target values or relevance label are already known before and the learning function maps the values onto it during optimization. In other words, Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output. $Y = f(X)$. The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data. It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance. Supervised learning problems can be further grouped into regression and classification problems. A classification problem is when the output variable is a category, such as "red" or "blue" or "disease" and "no disease" and a regression problem is when the output variable is a real value, such as "dollars" or "weight". Some common types of problems built on top of classification and regression include recommendation and time series prediction respectively. Some popular examples of supervised machine learning algorithms are: Linear regression for regression problems, Random forest for classification and regression problems, Support vector machines for classification problems.

Quality Measurement: With the help of this framework news search quality can be measured in terms of ranking. In this section the final machine-learned ranking can be compared with the search engine ranking. To measure the difference between two rankings spearman's rank correlation coefficient can be used. Higher the value of correlation coefficient, better is the ranking and hence better is the search engine. [8,13] are the evaluation studies which utilized this measure for quality measurement.



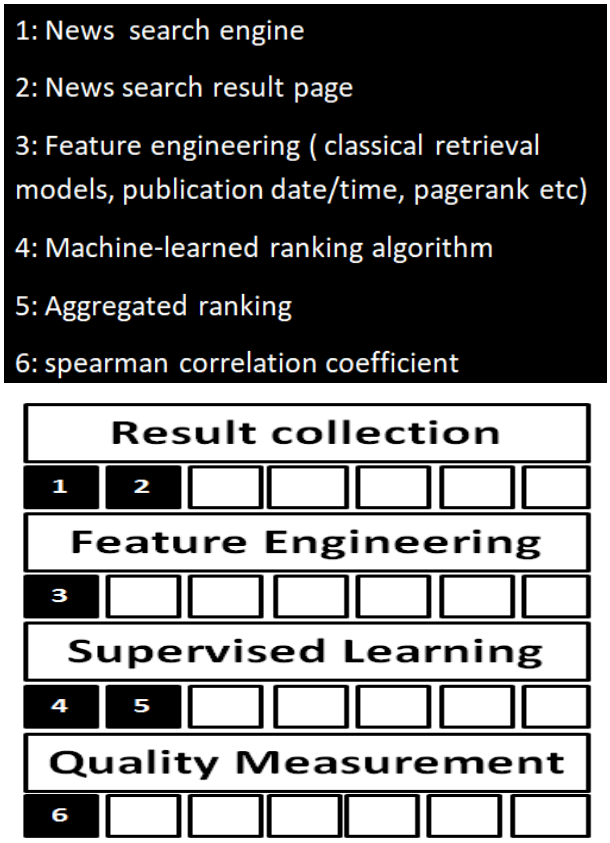


Fig. 2: Framework for evaluating retrieval effectiveness of news search engines.

3. RANKING ALGORITHM SELECTION

In this paper we will test the performance of six ranking algorithms included in the Lemur's RankLib library on our specified dataset collected from news search results. Description of these algorithms is as follows:

RankLib is a library with several learning-to-rank algorithm implemented. RankLib is the part of open source project called The Lemur Project. RankLib is written in java and the current version includes following algorithm:

RankNet: is a pairwise approach as described in [15]. RankNet uses a neural network together with gradient descent steps to control the learning rate in each iteration step. The neural network has two hidden layers and uses back-propagation to minimize a cost function to perform the pairwise ranking.

RankBoost: is a pairwise technique based on boosting [14].

RankBoost operates in rounds and choose the feature in each round that minimizes a loss function.

AdaRank: As described in [16], it is another boosting technique which combines weak rankers to create the ranking function. The algorithm is inspired by AdaBoost or "Adaptive Boosting" [19], a well recognized machine learning algorithm. AdaRank adopts the listwise approach to the learning to rank problem and tries to minimize the performance measures directly instead of indirect minimization of a loss functionas the similar algorithms above.

Coordinate Ascent: The Coordinate Ascent method is described as an optimization method in [18]. The method optimize through minimization of measure-specific loss, more specifically the mean average precision (MAP).The Coordinate Ascent suffer from getting stuck in local minimas, when searching for the global minima of the MAP, but by doing many restarts (typically 10) this can be avoided.

LambdaMART: is an ensemble method consisting of boosted regression trees (MART) in combination with LambdaRank [19]. LambdaRank is a neural network algorithm for learning to rank with the same basic idea as RankNet (backpropagation to minimizea loss function). LambdaRank's loss function is meant as an upgradedversion of the loss function in RankNet with faster running time and better performance on measures. The authors of LambdaRank points out that their algorithm could be combined with boosted trees. The developers of the LambdaMART algorithm implemented an algorithm that does what Burges et al. (2006) [20] did advice.

ListNet: is an algorithm described in [21]. ListNet isusing a neural network approach with gradient descent to minimize a loss function, similar to RankNet. ListNet differs from RankNet as the method uses the listwise approach instead of the pairwise approach taken in RankNet. In this way ListNet tries to utilize the benefits of the group structure of the training data.

4. RETRIEVAL EFFECTIVENESS TESTS

4.1 Dataset

Our dataset consists of 112 news queries which were entered into the *'Google News'* within the time period comprising from 14 May 2017 to 18 May 2017, i.e. four days. For each query top ten results are collected. We used Tf-idf and Latent Semantic Indexing (LSI) for extracting relevancy features and publication date/time as freshness feature. Codes for tf-tdf and LSI were written in R programming platform. We extracted publication date/time of the retrieved news document either from the news article itself as mostly the news articles includes published date/time in regular date/time format such as *DD/MM/YY* or from the search result page which definitely includes published date/time but in some other format such as *9 hrs ago, 59m, 1h, 14 May 23:41 PM or 20161216 07:44 UTC***.** Thereafter difference between query issue time and document publication time was calculated. The dataset also includes relevance label from human annotated judgement. First the human relevance judgement was done on five-point scale, a single human judge was used for the judgement, namely- *0-irrelevant, 1-fair, 2-good, 3-excellent and 4-perfect,* then inspired by [23] recency was coupled with following recency demotion guidelines:

Table 1: Recency demotion guidelines

Less than 6hrs	Very fresh	Upgrade 2-level up
Less than 12hrs	fresh	Upgrade 1-level up
More than 24hrs	old	Demote 1-level
More than 72hrs	Stale	Demote 2-level

4.1 NDCG values

All the six ranking algorithms were implemented using RankLib-2.7.jar file in the terminal. We have done 5-fold cross validation test which means the command will sequentially split the training data into 5 chunks of roughly equal size. The i-th chunk is used as the test data for the i-th fold. The training data for each fold consists of the test data from all other folds. The values reported in the table below is the average values from the test data for five folds.

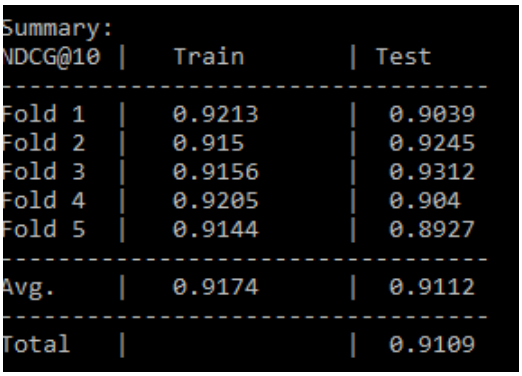


Fig. 3: 5-fold cross validation test results on the terminal

Since the Normalized Discounted cumulative gain (NDCG) and Expected Reciprocal Rank (ERR) are evaluation metrics used for graded relevance judgment we performed our retrieval effectiveness tests using these two metrics.

Table 2: NDCG values for RankNet, RankBoost and AdaRank.

NDCG values	RankNet	RankBoost	AdaRank
@1	0.7715	**0.8589**	0.8029
@2	0.7552	**0.8456**	0.8164
@3	0.7924	**0.8430**	0.8008
@4	0.7341	**0.8443**	0.8240
@5	0.7441	**0.8531**	0.8325
@6	0.8118	**0.8731**	0.8536
@7	0.8268	0.8876	0.8713
@8	0.7786	**0.9037**	0.8580
@9	0.8212	**0.9150**	0.8882
@10	0.8755	**0.9251**	0.9196

Table 3: NDCG values for Coordinate Ascent, LambdaMART and ListNet.

NDCG values	Coordinate Ascent	Lambda MART	ListNet
@1	0.8571	0.7404	0.8260
@2	0.8292	0.7337	0.8008
@3	0.8107	0.7711	0.8052
@4	0.8274	0.7674	0.8226
@5	0.8472	0.7771	0.8403
@6	0.8711	0.8088	0.8559
@7	**0.8884**	0.8271	0.8722
@8	0.8995	0.8479	0.8939
@9	0.8956	0.8724	0.9016
@10	0.9226	0.8843	0.9090

We observed NDCG values at different cut-off values (1-10), shown in Fig. 6. Again, **RankBoost** outperforms its contestants, except for NDCG@7 value where Coordinate Ascent had the maximal value.

Fig. 7 denotes the average NDCG values across all the cut-offs and obviously RankBoost had the maximal value, having a difference of **0.01006** with the second highest competitor, i.e. Coordinate Ascent.

Table 4: Average NDCG values.

Ranking algorithm	Avg. NDCG
RankNet	0.79112
RankBoost	**0.87494**
AdaRank	0.84673
Cordinate Ascent	0.86488
LambdaMART	0.80302
ListNet	0.85275

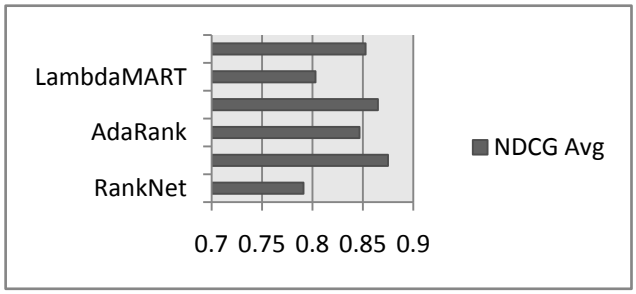


Fig. 4: Average NDCG values.

Similarily RankBoost outperforms at each cut-offs for ERR values.

Table 5: ERR values for RankNet, RankBoost and AdaRank.

ERR	RankNet	RankBoost	AdaRank
@1	0.6352	**0.7636**	0.7047
@2	0.7948	**0.8309**	0.7879
@3	0.7193	**0.8471**	0.8111
@4	0.7602	**0.8509**	0.814
@5	0.8130	**0.8518**	0.8307
@6	0.7917	**0.8527**	0.8194
@7	0.8189	**0.8529**	0.8194
@8	0.7966	**0.8531**	0.8374
@9	0.8128	**0.8532**	0.8371
@10	0.8172	**0.8534**	0.8370

Table 6: ERR values for Coordinate Ascent, LanmbdaMART and ListNet.

ERR	Coordinate Ascent	Lambda MART	ListNet
@1	0.7460	0.6596	0.7138
@2	0.8163	0.6583	0.7938
@3	0.8357	0.7699	0.8187
@4	0.8377	0.7891	0.8279
@5	0.8506	0.7783	0.8294
@6	0.8482	0.7903	0.8329
@7	0.8430	0.7927	0.8264
@8	0.8470	0.7978	0.8244
@9	0.8490	0.8090	0.8289
@10	0.8434	0.8066	0.8309

5. IMPLICATIONS FROM THE RETRIEVAL TESTS

Various implications derived from these tests are as follows:

1-On small dataset on around 100 queries (we took 113 queries for news search task), **RankBoost** [1] performed best both on NDCG and ERR values.

2-RankBoost performed consistently well around all NDCG cut-offs, see Fig. 4, against ListNet.

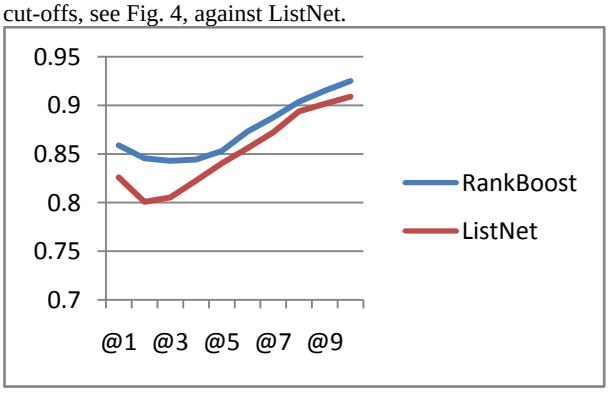


Fig. 5: RankBoost vs ListNet (across different NDCGG values).

3- The average NDCG value for Coordinate Ascent is 0.86488**,** which is **0.01006** less than average value of RankBoost. Although NDCG@7 value for Cordinate Ascent is higher but the NDCG curve is not consistent in terms of values and forms downward sharp cliffs at NDCG@3 and NDCG@9.

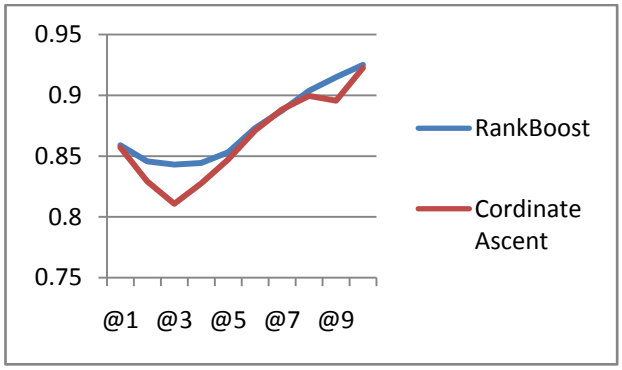


Fig. 6: RankBoost vs Coordinate Ascent (across different NDCGG values).

4- When we compare ListNet with Coordinate Ascent (both are list-wise algorithms), the later had 0.01213, average NDCG value higher than ListNet.

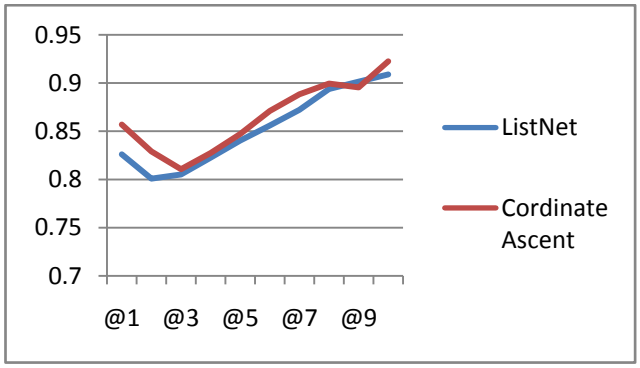


Fig. 7: ListNet vs Coordinate Ascent (across different NDCGG values).

5- RankBoost gave a clean sweep for ERR values at each cut-off.

6- A learning to rank algorithm with pair-wise approach can be fruitful for small datasets (not necessarily) in machine learning.

6. CONCLUSION

We proposed a theoretical framework for evaluating the retrieval effectiveness of news search systems. Since this framework utilizes supervised machine learning approach for news search evaluation, we conducted retrieval effectiveness tests on small dataset of about 112 queries for news search results to compare the performance of different learning to rank algorithms included in RankLib library. Experimental results shows that RankBoost [14], which follows pairwise approach, performed well both on NDCG and ERR metric, with ListNet and Coordinate Ascent, both following list-wise approach, were close rivals. These tests will encourage researchers to conduct retrieval effectiveness tests on news search results with large datasets and evaluate state-of-the-art news search engines.

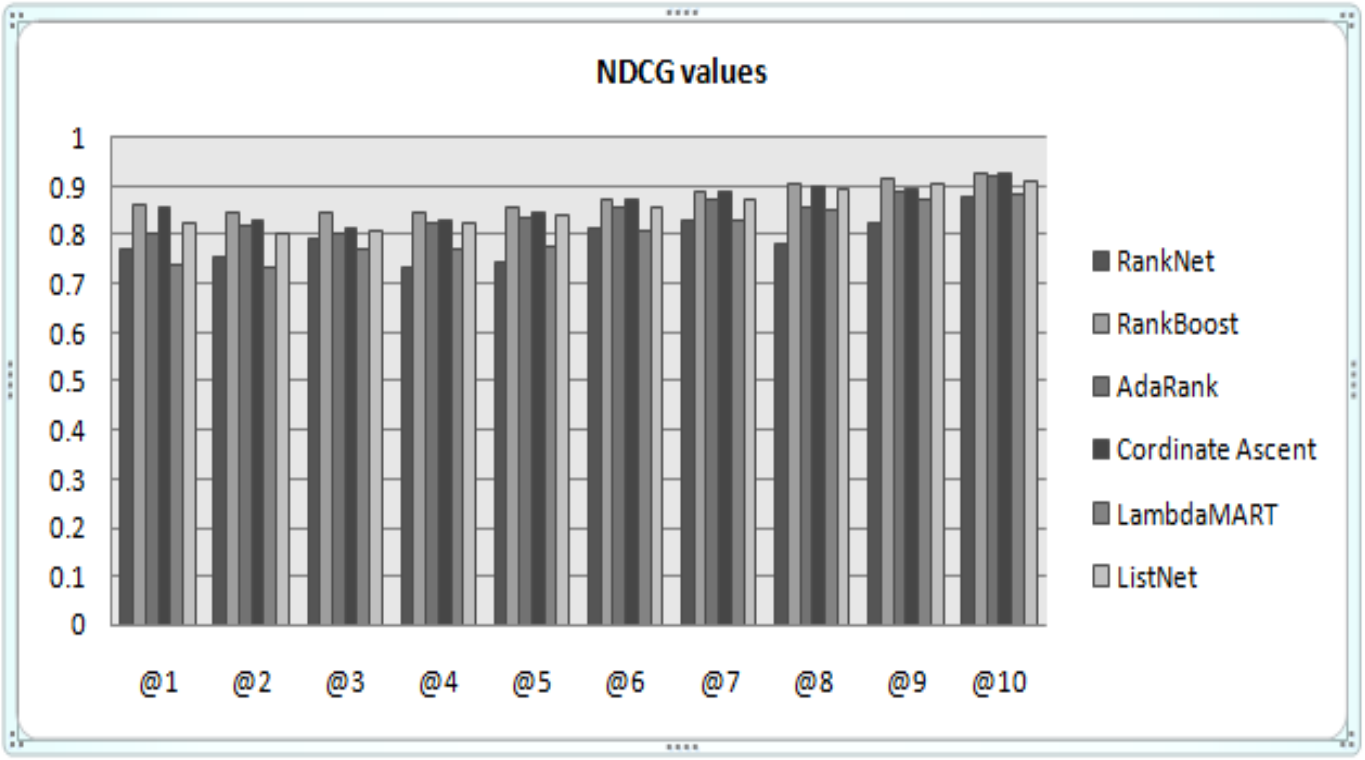


Fig. 8: NDCG values at different cut-offs.

6. REFERENCES

[1] Lewandowski, D., 2013. Evaluating the retrieval effectiveness of web search engines using a representative query sample. Journal of the Association for Information Science and Technology, Vol 66, issue 9, pages-1763-1775.

[2] Can, F., Nuray, R. and Sevdic, A. B., 2003. Automatic performance evaluation of Web search engines. Information Processing and Management 40 (2004) 495–514.

[3] Lewandowski, D., (2008) "The retrieval effectiveness of web search engines: considering results descriptions", Journal of Documentation, Vol. 64 Issue: 6, pp.915-937, https://doi.org/10.1108/00220410810912451

[4] Ali, R and Beg, M. M. S., 2011. An overview of Web search evaluation methods. Computers and Electrical Engineering. 37 (2011) 835–848.

[5] Leighton, H. V. and Srivastava, J., 1999. First 20 Precision among World Wide Web Search Services (Search Engines). Journal of the American Society for Information Science. 50(10):870–881.

[6] Clough, P. And Sanderson, M., 2013. Evaluating the performance of information retrieval systems using test collections. Information research. Vol 18(2).

[7] Vaughan, L., 2004. New measurements for search engine evaluation proposed and tested. Information Processing and Management 40 (2004) 677–691.

[8] Ali, R. and Beg, M., M. S., 2009. Modified rough set based aggregation for effective evaluation of web search engines. Fuzzy Information Processing Society, 2009. NAFIPS 2009. Annual Meeting of the North American.

[9] Bokhari, M. U. and Adhami, M. K., 2015. Article: A New Criterion for Evaluating News Search Systems. *Communications on Applied Electronics* 2(7):28-35, August 2015. Published by Foundation of Computer Science (FCS), NY, USA.

[10] Bokhari, M. U. and Adhami, M. K., 2016. How well they retrieve fresh news items: News search engine perspective. Perspective in Science. Elsevier, Volume 8, September 2016, Pages 469-471.

[11] Liu., K. L., Meng, W., Qiu, J., Yu, C., Raghavan, V., Wu., Z., Lu, Y., He, H. and Zhao, H., 2007. AllInOneNews: Development and Evaluation of a Large-Scale News Metasearch Engine. SIGMOD'07 Proceedings of the 2007 ACM SIGMOD international conference on Management of data Pages 1017-1028.

[12] Lewandowski, D. and Sunkler, S., 2013. Designinh search engine retrieval effectiveness tests with RAT. Information Services & Use 33 (2013) 53–59.

[13] Ali, R. and Naim, I., 2011. Neural Network based Supervised Rank Aggregation. In international Conference on Multimedia, Signal Processing and Communication Technologies, pages 72-75, IEEE 978-1-4577-1105-3.

[14] Freund, Y., Iyer, R., Schapire, R. E. and m Singer , Y. 2003. An Efficient Boosting Algorithm for Combining Preferences. Journal of Machine Learning Research 4 (2003) 933-969.

[15] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N. and Hullender, G., 2005. Learning to rank using gradient descent. In *ICML'05* Proceedings of the 22nd international conference on Machine learning, Pages 89 – 96, Bonn, Germany.

[16] Xu, J. and Li, Hang., 2007. AdaRank: a boosting algorithm for information retrieval. In *SIGIR'07,* Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pages 391-398.

[17] Schapire, R., E. and Singer, Y., 1999. Improved Boosting Algorithms Using Confidence-rated Predictions. Machine Learning, Volume 37, issue 3, pages 297-336.

[18] Metzler, D. and Croft, W. B., 2006. Linear feature-based models for information retrieval. In Information Retrieval, Kluwer Academic Publishers, Netherlands.

[19] Wu, Q., Burges, C. J., Svore, K. M. and Gao, J., 2010. Adapting boosting for information retrieval measures. Information Retrieval, Volume 13, issue 3, pages 254-270.

[20] Burges, C. J., Ragno, R. and Le, Q. V., 2006. Learning to rank with non-smooth cost functions. In Advances in Neural Information Processing Systems 18, 2006.

[21] Cao, Z., Qin, T., Liu, T. Y., Tsai, M. F. and Li, H., 2007. Learning to rank from pairwise approach to listwise approach. In *ICML'07* Proceedings of International Conference on Machine Learning.

[22] Kristofer, T., 2015. Learning to rank, a supervised approach for ranking of documents. Master Thesis in Computer Science - Algorithms, Languages and Logic Chalmers University of Technology, Sweden.

[23] Dong, A., Chang, Y., Zheng, Z., Mishne, G., Bai, J., Zhang, R., Buchner, K., Liao, C. and Diaz, F., 2010. Towars recency ranking in web search. In *WSDM'10* Proceedings of the third ACM international conference on Web search and data mining, pages 11-20, New York, USA.