Automatic Evaluation of Search Engines with Social Relevancy Rank Factoring

December 10, 2009

1. Abstract

In the past precision of information retrieval systems has been evaluated based on general subject queries and some specific query domains both manually and automatically. Manually evaluating the effectiveness of information retrieval systems, in terms of relevance, requires a large amount of human effort and time. Automatic evaluation is much better in adapting to the fast changing web and search engines, as well as the large amount of information on the Web. Many relevance scoring methods have been developed to evaluate the relevance of hits returned by search engines.The future of search almost certainly involves social networks, social graphs, or social filtering in some capacity[12]. The key question is how to organize the data in social media to be used as a factor in relevance scoring. We propose to develop a new measure for evaluating information retrieval systems called the Social Relevancy Rank (SRR). Using the Social Relevancy Rank in relevance scoring, search results will be reordered based on social relevancy to improve retrieval performance. We evaluate this approach throughly using 25 student evaluators from diverse backgrounds for a total of 1250 queries for recall and precision metrics.We believe that the Social Relevancy Rank can be as prominent as Page Rank in a few years. By factoring the social relevancy rank during evaluation of search engines we are contributing towards more efficient and effective real time search and live search in the Web Search Industry. Also, the distinction between real time search,semantic search and social search will diminish and become meaningless. All will together play a role in contextualising and personalizing search for the users and would holistically improve what we think of today as search.

2. Introduction

The information retrieval field has been using retrieval experiments on test collections to advance the state of art[2]. The basic frame work of research in information retrieval is quite straightforward. A user poses a query that represents the information need; the retrieval system uses a matching algorithm to identify the documents that are likely to satisfy this need; and the user reads the re-turned documents to find the answers to this query.

Based on this framework, an information retrieval system can be measured with respect to a test collection[2][3][4].A test collection for information retrieval requires three components:

- set of documents,
- a set of queries and
- a set of relevance judgments.

In a common experimental scenario, a particular retrieval system configuration is used to run a set of queries against a set of documents and the results are obtained in terms of precision and recall.

Evaluating the effectiveness of information retrieval systems, in terms of relevance, requires a large amount of human effort. Many environments, such as the World Wide Web, grow and change too rapidly for a single evaluation to carry meaning for any extended period.

Test Collection is static data.The web is live data continuously changing. Thus there is a need for an evaluation methodology that can practically and repeatedly be applied to evaluating search services on the live web. One of the key advantages of an automated approach is that it enables the authors to run thousands of queries where a manual approach is generally limited to a handful of queries.Many relevance scoring methods have been developed to evaluate the relevance of hits returned by the search engines[20].

We look forward to further automating the process of test collection and evaluation by taking into consideration more dynamic and voluminous human organized information which keeps pace with the fast evolving live web. Large manually built directories which are present on the web like ODP(Open Directory Project) and Looksmart directories have opened doors to completely new evaluation procedures[19].

An open question is how can social media complement traditional relevance scoring methods? Open directory was a means for the internet to organize itself with a little help from the humans.On leveraging the power of social media we are focussing on user-curated data, or people-powered search.As a first step towards this we are evaluating the possibility of calculation of a social relevancy rank which can be factored into the currently used relevance scoring methodologies of retrieval systems.

3. Background

The accuracy of an information retrieval system is measured in terms of precision, the proportion of retrieved documents that are relevant.The coverage of a system is measured by recall.

For a large corpus, it is infeasible to construct an ideal test collection where each document will be judged as relevant or not relevant with respect to each query, as prohibitive effort would be required to examine and judge every document.

A more efficient approach is to examine and judge only a subset of the documents, provided the subset can be selected to include all the relevant documents. This practice is represented by pooling methodology in which only the first k documents from a each of a number of sources are judged [Cormack et al.][2].These methods continue to be

used at theTREC conference series, which support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. An important outcome of these workshops is a set of large test collections that are now widely used by the retrieval community. A variety of organizations -including many that do not participate in the TREC workshops themselves use these collections to develop their own retrieval strategies. Since a lot of research is being conducted on the collections, it is important that the collections reliably reflect the relative merit of different retrieval strategies. Kazuko Kuriyama et al. [6] verified the efficiency and the effectiveness of the pooling method, the exhaustiveness of the relevance assessments, and the reliability of the evaluation using the test collection based on the pooling method. It has been studied using empirical investigations that results based on the relevance judgments formed from a limited depth pool are reliable-if the pool is sufficiently deep-both for systems that contributed to the pool and for new systems.

Zobel[3] proposed a new pooling strategy that increases the number of relevant documents found for a given judgment effort. He also proposed that simple regression on the per query number of new relevant documents found at each pool depth, although highly approximate, is a good basis for choice of queries for further judgment effort.

More effective test collections requiring fewer judgments were developed by Cormack et al. [2] with the introduction of methods like Move-to-Front pooling and Interactive Searching and Judging. This was further developed by Aslam et al. [7] who presented a unified model for meta search, pooling and system evaluation based on the Hedge algorithm for on-line learning where the proposed system learns which documents are likely to be relevant from a sequence of on-line relevance judgments.

New Evaluation Measures It is understandable that building substantially larger test collections with essentially complete relevance judgments through pooling is not likely to be possible due to the amount of assessor time and the diversity of retrieval runs that would be required. Chris Buckley et al.[8] looked at the effect,relaxing the completeness assumption, has on the Cranfield evaluation methodology. The authors introduced a new measure, named **bpref** which is inversely related to the fraction of judged non relevant documents that are retrieved before relevant documents.The measure **bpref** is also more resilient to change than other measures when used on dynamic collections and in an embedded collection environment, where a test collection with known judgments is embedded in a much larger collection of similar documents with no judgments.

Carterette et al.[9] developed an algorithm that can be used to incrementally design retrieval systems by simultaneously comparing sets of systems where the number of additional judgments needed after each incremental design change decreases at a rate reciprocal to the number of systems being compared.A lot of work has been extended in the area of Minimal test collections by Carterette et al.[10] for retrieval evaluation

and dynamic test collections by Ian Soboroff [11] almost simultaneously by using a new perspective on average precision leading to an algorithm for selecting documents that should be judged to evaluate retrieval systems in minimal time and a methodology of collection maintenance which supports measuring search performance both for a single system and between systems run at different points in time respectively.

Further evaluation on reducing the assessor effort[13] and reusability of judgments were conducted by Carterette et al. [12][14] who presented a method for augmenting a set of relevance judgments with relevance estimates that requires no additional assessor effort.With as few as five judgments per topic taken from only two systems, the method demonstrated that we could reliably evaluate a larger set of ten systems.Using cluster hypothesis that closely associated documents tend to be relevant to the same request, the authors Carterette et al. [12] use inter-document similarity to provide more accurate and robust evaluation methodology to accurately rank retrieval systems with upto 99 percent fewer relevance judgments than ever before laying the foundation for semi automated evaluation of retrieval systems.

4. Current Limitations and Key Open Issues

Users of the World-Wide Web are not only confronted by an immense overabundance of information, but also by a plethora of tools for searching for the web pages that suit their information needs. Web search engines differ widely in interface, features, coverage of the web, ranking methods, delivery of advertising, and more.

Most of the evaluation methodology of search engines is based on TREC methodology of using static test collection and manual relevance judgments to evaluate systems. When we use TREC style of evaluation, it creates a large number of problems.

The web is too large to perform deep manual relevance judgments of enough queries to be able to measure recall in any reasonable way. Evaluating the effectiveness of information retrieval systems, in terms of relevance, requires a large amount of human effort. Many environments, such as the World Wide Web, grow and change too rapidly for a single evaluation to carry meaning for any extended period. Changes in their document collection, query population, and set of search services demand the repetition of evaluations over time.

Test Collection is static data. Static test collections become outdated too quickly and require too much effort to reconstruct. Rather, practitioners often compare a small number of live engines by judging every result retrieved at a shallow depth without system pooling. Also the number of queries necessary for such an evaluation to be reliable must be determined.

Web is live data continuously changing. The collection on web is constantly changing, i.e. any evaluation is not reproducible in the future.

Almost less than half of queries are informational in nature. Singhal et al. evaluated the search tasks of web users and proposed that navigational queries were more

significant to web search evaluation than traditional TREC ad-hoc information gathering. Even though importance of navigational queries led to TREC incorporate know item evaluations as a part of web track, it is still essentially a static test collection.

Thus there is a need for new evaluation methodology that can practically repeatedly be applied to evaluating search services on the live web. One of the key advantages of an automated approach is that it enables the authors to run thousands of queries where a manual approach is generally limited to a handful of queries.

The need for manually assessed relevance judgments complicates the Information Retrieval System evaluation. However large manually built directories which are present on the web like ODP(Open Directory Project) and Looksmart directories open the door to completely new evaluation procedures.

Beitzel et al. showed that by assuming that web pages are the known relevant items for queries that exactly match their title, they use the ODP (Open Directory Project) and Looksmart directories for the Evaluation of Known-Item Retrieval. The research involved testing this approach with a sample from a log of ten million web queries and show that such an evaluation is :

- Unbiased in terms of the directory used,
- Stable with respect to the query set selected and
- Correlated with a reasonably large manual evaluation.

Jenson et al. also showed that by augmenting manual judgments with pseudo-relevance judgments mined from Web taxonomies reduces both the chances of missing a correct pair wise conclusion, and those of finding an errant conclusion, by approximately 50

Today ODP powers the core directory services for many of worlds largest search engines like Netscape Search,AOL Search,Google and Alexa. Overture announced a third party search combining Yahoo! Directory search results with ODP titles, descriptions and category metadata.

5. Proposed Novel Approaches

The proposed novel approach is to take this research one step higher by taking into consideration a more dynamic and voluminous human organized information which keeps pace with the fast evolving live web.We tried to identify the next big attempt where humans organize content on the web or the place where listings of similar topics are getting grouped together. The most underutilized resource with a lot of potential is that of social media data predominantly data from blogs, twitter, facebook, networking forums etc. Social applications are the fastest growing segment of the web. They establish new forums for content creation, allow people to connect to each other and share information, and permit novel applications at the intersection of people and information.Social media has been primarily popular for connecting people, not for finding information.

How can social media complement traditional web search is the question that leads

us to a natural conclusion to look for ways to utilize abundant data getting churned on an everyday basis. Open directory was a means for internet to organize itself with a little help from the humans, the data from social media if utilized in a proper way can go a long way in ensuring high quality relevant content.

To use this data for evaluation purposes, we propose to develop a social relevancy ranking just like the page rank, not essentially the same but similar in concept to be used indirectly or partially in the evaluation process. Also another approach would be to enable a scenario where in when a user searches streams of activity the results will be ordered not chronologically but by how relevant each is to the user on a social graph. This will change the face of general web searches in time. Today the results are automatically ranked by relevancy and freshness. Once a social relevancy rank is factored in, search results are reordered based on social relevancy. When a user is looking for useful information the secret is selection and trust (i.e. filtered information).This is where the power of network and social media becomes unignorable. It is indeed essential for us to take this i.e. social relevancy rank into account for evaluation purposes.

We need to keep in mind that people make sophisticated decisions about who to ask for information, and this varies according to the nature of the task. For example, there's no guarantee that you'll ask the same person to recommend both movies and financial products however well one knows them.These two examples depend to varying degrees on taste overlap and the expertise of the information source. It's this dynamics we'll need to understand in greater detail to develop a good algorithm. The key question is how to organize the data in social media to be used as a factor in relevance scoring.Some of the ways we propose include:

- Ranking by users friends and people users know, follow on twitter like social networking sites, blogs etc.
- Data based on links or semantic analysis of social data of a user over a time or even comparing people based on the links and semantics of their links and tweets. This looks like a costly calculation but it is not a difficult problem if examined over the time.
- Forming a social biography of a person based on the information present about him in the social media. This can be created by either directly using the information fed by the user augmented by his activities, collaborations and usage streams present in the social media.
- The problem with just taking into account users friends and people he knows and follows is that of sparse data.Since we need more data, we can incorporate other sources that we trust, i.e. , a social graph. Search results could rank data taking into account people not just the people directly connected to the user but those also indirect e.g., in Twitter scenario, it could be people who are followed by people you follow. Basically, the assumption being friends of friends contributions will be as valuable.

- Taste neighbors include people with similar taste, and this approach already works great with vertical social networking searches such as last.fm, flixter and good read. This enables us to get an idea about which people other than just your friends are just like you.
- Aggregating semantics of the crowd: Using the number of people following a person's social media behavior should also be given a small weight age as someone who is being followed by 1000s and 10000s of people is probably going to be more relevant to a user than just someone the user doesnt know at all.

6. **Evaluation**

The evaluation strategy would be to use a large set of manual relevance judgments to compare our automatic evaluation with social relevancy rank (SRR) versus the automatic evaluation method.(without taking into account the SRR).The evaluation will be performed for Recall and Precision metrics.

Student Evaluators with Diverse Backgrounds Using student evaluators would be a good option due to the easy access to students. We plan to use 25-30 student evaluators 5 from Information Retrieval course, 5 graduate Computer Science students,5 undergraduate Computer Science students, 5 students from arts, 5 students from Biological sciences and 5 from Physics and Chemical Engineering streams. This will ensure a good mix of students from varied backgrounds and not cause any kind of bias towards students who have knowledge about search engine rankings and methodology.

Evaluation Methodology

Students would be asked to evaluate two lists of 5 documents for each query. One will pull out information without the social relevancy rank factored in and one with social relevancy rank taken into account. We will make the students make a decision for both simultaneously on adjacent windows for the top 5 results on both the windows. There will be 10 such queries in each set. Every evaluator has to mandatorily evaluate at least 5 sets, and more if desired. A single set of ten queries has to be evaluated in a single session and if a user wishes to leave it incomplete he has the flexibility to do so. The next time the user starts a new session he would be given a new set of ten queries. This not only gives flexibility to the evaluator but also reduces the threats to validity of the results which are characteristics of such evaluation processes.

Evaluation Interface

The evaluation interface would be simple with just two windows with query and top 5 results for both the evaluation techniques. Each of the top 5 results will have three buttons

(a) Relevant

(b) Not Relevant

(c) Duplicate

Evaluation Metrics

We will then find the improved precision and recall displayed by the new evaluation methodology with social relevancy factor

taken into consideration over the original one or vice versa as per the results generated.

The future evaluations can also involve comparing the social relevancy factored evaluation process to a completely manual evaluation process and calculate the Pearson Correlation. However this might not be possible due to time and resource constraints in the first level evaluations.

Discussion It would be ideal to not just use students for evaluation and use general user data for evaluation purposes. However,students are the most easily available resource, so by considering students from diverse backgrounds, we are reducing the bias to the maximum possible extent.

The evaluation process gives us an idea for each set how much better does Social Relevancy Factor perform as compared to the one without it for the top 5 results for each query in both the cases. Here we are making an assumption based on generic human behavior of usually not going beyond the first 5 queries or first page for information needs.

7. Related work

The most closely related work is Carterette et al.s approach to semi-automate the evaluation of retrieval systems using document similarities[14] and on rank correlation and distance between rankings[15]. Our technique looks forward to augment closely on the work carried out by these authors with the social relevancy rank factored in.

There is also a move towards forming test collections with so system pooling as expressed by Mark Sanderson et al[16],Cormack et al[17]. who have developed approaches other than pooling like even random sampling in an attempt to achieve results comparable to pooling methods. These have not yet achieved results which can be compared to pooling methodologies. William Webber et al[15]. have proposed methodologies leading to score adjustment for correction of pooling bias as expressed by the authors Cormack et al.[17]

Other interesting but distinctly different work being carried out is in the area of building an information retrieval test collection for spontaneous conversational speech by Douglas W. Oard et al[18].

8. Contributions

Social Applications are the fastest growing segment of the web. By factoring the social relevancy rank during evaluation of search engines we are contributing towards more efficient and effective real time search and live search in the Web Search Industry. Also, the distinction between real time search,semantic search and social search will diminish and become meaningless.All will together play a role in contextualising and personalizing search for the users and would holistically improve what we think of today as search.

9. Acknowledgements

I would like to thank Brynn M.Evans student at UC San Diego working on role of social interactions during search and Alex Iskold, feature writer for Read Write Web and CEO of Adaptive Blue for their ideas on Social Search and Future of Search Engines without which I not have been able to come up with this proposal.

10. References

1. Ellen M. Voores. Variations in Relevance Judgments and Measurement of Retrieval E ectiveness,SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval,1998.

2. Gordon v. Cormack, Christopher R. Palmer,Charles L.A.Clarke.Efficient Construction of Large Test Collec-tions,SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval,1998.

3. Justin Zobel. How reliable are the results of Large scale Information Retrieval Experiments,SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval,1998.

4. Gordon V.Cormack, Ondrej Lho-tak, Christopher Palmer, Estimating precision by random sampling,SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval,1999.

5. Mark Sanderson, Forming Test Collections with no system pooling,SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval,2004.

6. Kuriyama,Kando,Noriko,Nozue,Toshihiko and Eguchi, Koji, Pooling for a Large-Scale Test Collection: An Analysis of the Search Results from the first NTCIR Workshop.

7. Javed A. Aslam, Virgiliu Pavlu,Robert Savell, A unified model for meta-search, pooling, and system evaluation,CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management,2003.

8. Chris Buckley,Ellen M. Voorhees, Retrieval Evaluation with Incomplete Information,SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval,2004.

9. Ben Carterette and James Allan, Incremental Test Collections,CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management,2005.

10. Ben Carterette, James Allan, Ramesh Sitaraman, Minimal Test Collections for Retrieval Evaluation, SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval2006.

11. Ian Soboroff ,Dynamic test collections: measuring search effectiveness on the live web, SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval,2006.

12. Ben Carterette,James Allan, Semiautomatic Evaluation of Retrieval Systems Using Document Similarities,CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management,2007.

13. Ben Carterette,Mark D. Smucker, Hypothesis testing with incomplete relevance judgments,CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management,2007.

14. Ben Carterette, Robust Test Collections for Retrieval Evaluation,SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval,2007.

15. William Webber, Laurence A. F. Park, Score adjustment for correction of pooling bias, SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval,2009.

16. Mark Sanderson, Hideo Joho, Forming Test Collections with No System Pooling, SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval,2004.

17. Gordon V. Cormack, Thomas R. Lynam, Power and bias of subset pooling strategies,SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval,2007.

18. Douglas W. Oard, Dagobert Soergel, David Doermann, Xiaoli Huang, G. Craig Murray, Jianqiang Wang, Bhuvana Ramabhadran, Martin Franz, Samuel Gustman, Building an Information Retrieval Test Collection for Spontaneous Conversational Speech,2004.

19. Beitzel, Steven M. and Jensen, Eric C. and Chowdhury, Abdur and Grossman, David and Frieder, Ophir , Using manually-built web directories for automatic evaluation of known-item retrieval,SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval,2003.

20. Longzhuang Li, Yi Shang, and Wei Zhang,Relevance Evaluation of Search Engines Query Results,World Wide Web,Kluwer Academic Publishers,2000.