Using PageRank to Infer User Preferences

Praveen Chandar and Ben Carterette
{pcr,carteret}@udel.edu
Department of Computer and Information Sciences
University of Delaware
Newark, DE, USA 19716

ABSTRACT

Recently, researchers have shown interest in the use of preference judgments for evaluation in IR literature. Although preference judgments have several advantages over absolute judgment, one of the major disadvantages is that the number of judgments needed increases polynomially as the number of documents in the pool increases. We propose a novel method using PageRank to minimize the number of judgments required to evaluate systems using preference judgments. We test the proposed hypotheses using the TREC 2004 to 2006 Terabyte dataset to show that it is possible to reduce the evaluation cost considerably. Further, we study the susceptibility of the methods due to assessor errors.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]

Keywords: preference judgments, PageRank

1. INTRODUCTION

The idea of pairwise preference judgments is relatively new in IR literature. In the case of preference judgments, an assessor looks at a *pair* of documents and expresses a preference for one over the other instead of assigning a relevance label to a document. Comparison studies between absolute and preference judgments show that preference judgments have various advantages such as reducing the complexity of the task and increasing inter-assessor agreement [3]. Preference judgments tend to help assessors make finer distinctions between documents [5]. Although it is possible to use graded absolute judgments, it is difficult to determine the specifics of the grades, and the burden on assessors is likely to increase with an increase in the number of grades.

A major drawback of using pairwise preference judgments is that as the number of documents in the pool increases, the number of judgments increases polynomially. If there are n documents in the pool, then $n(n-1)/2$ preference judgments would be necessary to ensure that all pairs of documents are judged. This enormous increase in the number of pairwise judgments not only increases evaluation cost but also gives rise to user fatigue and boredom, possibly leading to poor quality of judgments.

In this work, we describe a novel technique that could be used to reduce the number of pairwise judgments. Firstly, we show that pairwise judgments for each query can be represented in the form of a directed graph on which various graph algorithms can be applied. We study the behavior of PageRank and show that the number of pairwise judgments could be reduced considerably. Further, we show that the PageRank method, although susceptible to assessor error, does significantly better than the majority vote approach that is considered as the baseline.

2. PREFERENCE FRAMEWORK

In this section we describe a novel technique to collect preference judgments at a reduced cost. The technique consists of two steps: graphical representation of pairwise judgments and scoring of each node in the graph. Scoring of nodes can be done using various graph algorithms such as PageRank, HITS, etc.

We describe a typical user interface for collecting preference judgments as follows: the assessor would be shown two documents and a statement of an information need (a topic); the assessor would have to pick the most preferred document using the *prefer left* or *prefer right* buttons. Additionally, both documents could be judged *not relevant* to indicate that all other documents should be preferred to them. The criteria for preference is task dependent and is often explained to the assessor in the form of guidelines.

The collected pairwise judgments are then represented in the form of a directed graph. Each unique document is represented as a vertex in the graph and the edges between vertices corresponds to the pairwise judgments. While there are several ways to do this conversion, we describe a way which worked for us. Let's say there are three judgments a user can make for a pair of documents (Doc_A, Doc_B): prefer Doc_A, both non relevant, and prefer Doc_B. When prefer Doc_A is selected for a pair, this translates to a directed link from Doc_B to Doc_A, similarly a link from Doc_A to Doc_B when Doc_B is preferred and there exists no link between Doc_A and Doc_B in the case of both non relevant.

Now that the pairwise judgments have been represented in the form of a graph, there are several graph algorithms such as PageRank, HITS, etc. that could be used to assign a score. The score represents the relevance of a document and documents are ranked based on these scores. In this work we compare the performance of PageRank with the a baseline majority vote method.

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
Copyright 2012 ACM 978-1-4503-1472-5/12/08 ...$10.00.

PageRank The PageRank algorithm proposed by Brin and Page [2] is a graph algorithm that assigns a numerical weighting to each vertex in a graph, with the purpose of measuring its relative importance within the graph. We compute the page rank scores for each node in the preference graph and hypothesize that the page rank scores correlate to the degree of relevance of the document.

Majority Vote A common technique used to produce a ranking from pairwise judgments is to sort documents by the number of times the document was preferred. The count of the number of preferences can be obtained from the preference graph by computing the in-degree for each node. We use this method as our baseline.

Both algorithms produce the same ranking when all pairwise judgments are used, *i.e.* if $n(n-1)/2$ pairwise judgments are used to judge n documents. But their behavior changes if some pairwise judgments are removed from the sample of $n(n-1)/2$ pairwise judgments. In order to study this behavior, we simulate preference judgments from graded judgments and evaluate the performance of rankings produced by randomly sampling 1%, 5%, 10%, ..., 100% of $n(n-1)/2$ pairwise judgments. We use the nDCG measure proposed by Jarvelin and Kekalainen [4] to evaluate the performance of each ranking against original graded judgments.

3. IMPLEMENTATION AND RESULTS

Our experiments are conducted on data simulated using the TREC 2004 to 2006 Terabyte dataset. The TREC Terabyte consists of a total of 150 topics with relevance graded judgments on a three-point scale. The corpus is a collection of Web data crawled from websites in the GOV domain during early 2004. The absolute graded judgments are converted to preference judgments by generating all possible pairs for the documents in the qrels file for each topic; the document with a higher grade is preferred in each pair. Aslam et al. [1] presented a meta-search approach known as meta-AP which is a function of the document's position in a set of ranked lists, with higher rankings contributing more to the metaAP score. We employ metaAP scores to resolve ties when the grades of both the documents in a pair are equal (document with higher meta score is chosen). Document pairs containing two non-relevant documents are considered as *both non-relevant* (no link in preference graph).

Figure 1 shows the nDCG scores at rank 20 and 1000 for the PageRank method compared with the majority votes method. The figure includes a run each for various sample sizes of the pairwise judgments and three *Down Sample* runs represented as *DS*. The *Down Samples* were generated by pairing each document in the pool of n documents with another document k times, *i.e.* each document in the pool of n documents is paired with *another* document for *DS1* and *two* other documents for *DS2* and so on. Note that there would be n pairwise judgments in *DS1* and $2n$ pairwise judgments in *DS2*. The nDCG computed from PageRank-based judgments is much higher than that computed from majority vote. Clearly, the PageRank method requires fewer number of pairs of judgments compared to the majority votes method on average and the results suggest that it is possible to collect preference judgments by judging just 5% of the pairs with minimal loss of performance.

Figure 2 shows the performance of majority votes and PageRank methods with error. We added error for each run by sampling 1%, 5% and 10% of the pairs and changing the judgments. The judgments were changed randomly either by deleting the edge or changing the direction of the edge. The PageRank method dominates the majority vote method when more preferences are given (5% and greater), but majority vote tends to do better for smaller samples and when there are a constant number of preference judgments for each document.

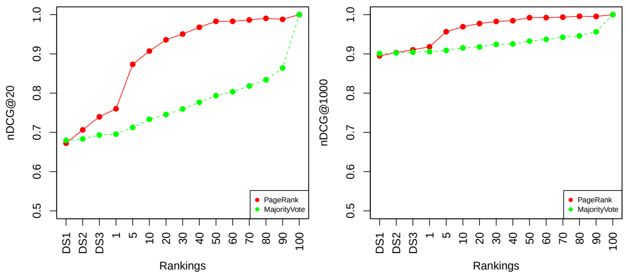

Figure 1: nDCG@20 and nDCG@1000 scores for the PageRank and majority votes method for various sample sizes. DS denotes *Down Sample*.

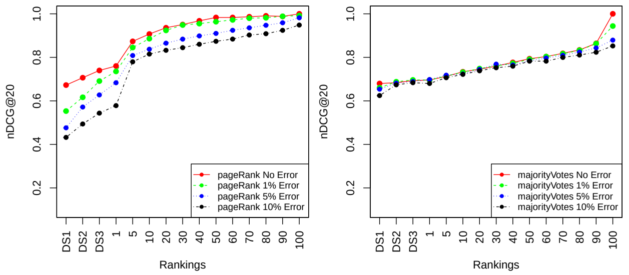

Figure 2: nDCG@20 scores for the PageRank and majority votes method for various sample sizes with assessor errors. DS denotes *Down Sample*.

4. CONCLUSION AND FUTURE WORK

We have presented a novel method to reduce evaluation cost using PageRank. While it is common to use majority vote to score and obtain a ranking from preferences, we have shown that using PageRank might be cost effective. Further, the PageRank method outperforms the baseline while assessor errors are added. Future directions for our work includes studying various other graph methods, picking documents intelligently for pairwise judging.

5. REFERENCES

[1] J. A. Aslam, V. Pavlu, and E. Yilmaz. Measure-based metasearch. In *Proceedings of SIGIR*, SIGIR '05, pages 571–572, NY, USA, 2005. ACM.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of WWW*, pages 107–117, 1998.

[3] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais. Here or there: preference judgments for relevance. In *Proceedings of the ECIR*, pages 16–27, 2008.

[4] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, pages 422–446, October 2002.

[5] M. E. Rorvig. The simple scalability of documents. *JASIS*, 41(8):590–598, 1990.