Size doesn’t always matter: exploiting pageRank for query routing in distributed IR

Josiane Xavier Parreira and Sebastian Michel
and Matthias Bender

2006
Size Doesn’t Always Matter: Exploiting PageRank for Query Routing in Distributed IR

Josiane Xavier Parreira, Sebastian Michel, Matthias Bender
Max-Planck Institut für Informatik
Saarbrücken, Germany
jparreir, smichel, mbender}@mpi-inf.mpg.de

ABSTRACT
PageRank authority scores have proven to be a powerful ingredient to local document scoring. Since query routing, i.e., carefully selecting a small subset of promising peers for a particular query from a large network, bears a close resemblance to local document scoring, it suggests itself that authority scores could also be beneficial for query routing, which is one of the biggest challenges in P2P Web search. For that purpose, we showcase JXP, an authority score measure that converges to true global PageRank scores in a distributed environment. Subsequently, we present several possible strategies to incorporate authority scores into query routing, including a hybrid strategy that combines authority scores with other existing measures. Preliminary experimental results support our hypothesis that authority scores can be highly beneficial to query routing.

Categories and Subject Descriptors
H.4 [Information Systems Applications]: Miscellaneous; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—selection process, information filtering; H.3.4 [Information Storage and Retrieval]: Systems and Software—Distributed Systems

General Terms
Algorithms, Design, Experimentation

Keywords
Peer-to-Peer information systems, distributed PageRank, JXP, query routing

1. INTRODUCTION

1.1 Motivation
The PageRank authority measure [7] is a powerful technique to improve the result ranking quality in centralized Web search. Its intuitive approach to favor documents from high-authority sources has been proven a pivotal ingredient to sophisticated document scoring models, beyond popular statistical information describing the documents, such as tf*idf or BM25 [22].

However, the computation of PageRank authority scores has traditionally required complete knowledge of the underlying Web graph, i.e., knowledge of all documents (nodes) and interconnecting hyperlinks (edges). While there exist a number of approaches to compute global PageRank authority scores in an environment with the Web graph partitioned disjointly over the peers [25, 30, 14, 24], we envision a distributed system of autonomous peers that independently crawl the Web, effectively leading to (typically overlapping) partitions of the Web graph being stored at the peers. We have recently developed the first solution to efficiently compute authority scores that provably converge to global PageRank authority scores in such an environment in a completely decentralized manner [19].

On top of such a peer collaboration, P2P Web search engines try to leverage from the distributed knowledge by evaluating IR-style top-k queries over the combined document corpus of all peers. The peers maintain a local search index built from the pages they have crawled and offer to evaluate queries based on their local index on behalf of other peers. One of the key challenges in such a system is query routing, i.e., judiciously selecting a small subset of remote peers that are expected to be good sources of information for a specific query from an a-priori unlimited number of peers. Popular techniques for query routing, such as CORI [9, 8], tend to prefer larger peers (i.e., peers containing more documents) over smaller peers, as larger peers are expected to have a higher probability of containing high-quality query results. But: size doesn’t always matter!

For example, consider a general purpose news portal, like cnn.com, and a highly specialized portal for financial statements. While the general purpose portal may have more documents for a company like Enron, the authority of the financial portal may be even higher than the (already high) authority of cnn.com.

Conceptually, the query routing process closely resembles the local document scoring process when evaluating a query. If we visualize each peer as one large document (the combination of all its local documents) like query routing strategies based on statistical language models [26] do, query routing boils down to finding the top-k “documents” in the network. As PageRank authority scores have been shown to greatly improve this process, it suggests itself to
exploit global PageRank scores for query routing purposes, so to base query routing on an a scoring model that better captures the expected result quality of a peer.

1.2 Contribution

The key idea is to improve the query routing process by preferring peers that have a high local PageRank score mass. We present different approaches that take into account the total PageRank score mass of a collection, query-specific portions of the PageRank score mass, and a hybrid framework that combines the authority score mass with other existing scoring models for query routing, such as CORI. We discuss the impact that both ingredients have on query routing and provide evidence gained from preliminary experiments that shows the potential impact on query result quality.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 illustrates a generalized system architecture for P2P Web search. One possible (though fully orthogonal) way to compute global authority scores and its integration into the system architecture is introduced in Section 4. Section 5 discusses several approaches to exploit PageRank authority scores for the query routing process in detail. Section 6 develops a hybrid strategy that combines PageRank authority scores and CORI-style scores into a single framework for query routing. The results of a preliminary experimental evaluation are illustrated in Section 7, before Section 8 concludes this work and points at future research directions.

2. RELATED WORK

2.1 Query Routing

Many approaches have been proposed for collection selection in distributed IR, most notably, CORI [9, 8], the decision-theoretic framework [11], the GLOSS method presented in [12], and methods based on statistical language models [26]. [18] presents a comparison of popular query routing strategies.

Those techniques have been used for query routing in P2P systems, but fail to address the overlap between collections. An overlap-aware query routing strategy to overcome this problem has been developed in [17].

2.2 P2P Systems

The peer-to-peer (P2P) approach has become popular in the context of file-sharing systems like Gnutella or KaZaA. These early P2P systems used message flooding to distribute information over there unstructured network overlays.

The ample recent work on structured P2P networks, such as Chord [27], CAN [20], Pastry [23], or P-Grid [2], has developed scalable routing protocols for single-dimensional key-based requests only. How to map multidimensional data onto distributed hash tables (DHTs) and other overlay networks [4, 29] has not received enough attention, and these approaches do not work for the very-high-dimensional data spaces formed by text keywords and they do not provide any support for ranked retrieval either.

Recent research on P2P search, including our MINERVA project [17], addresses these problems from different directions [3, 16, 5, 10, 31, 13, 21, 28].

2.3 Authority Scores

The basic idea of the most popular measure of authority, PageRank, is the assumption that a link from document $p$ to document $q$ represents an implicit endorsement of $q$, which adds to $q$'s authority. How much $p$ contributes to the authority of $q$ is proportional to the importance of $p$ itself.

This recursive definition of authority is captured by the stationary distribution of a Markov chain that describes a random walk over the Web graph, starting at an arbitrary document and following a random outgoing link from the current page at every step. To ensure the ergodicity of this Markov chain (i.e., the existence of stationary page-visit probabilities), additional random jumps to uniformly chosen target pages are allowed with a small probability ($1 - \varepsilon$). Formally, the PageRank of a page $q$ is defined as:

$$PR(q) = \varepsilon \times \sum_{p | p \rightarrow q} PR(p)/out(p) + (1 - \varepsilon) \times 1/N$$

where $N$ is the total number of pages in the Web graph, $PR(p)$ is the PageRank score of the page $p$, $out(p)$ is the outdegree of $p$, the sum ranges over all link predecessors of $q$, and $(1 - \varepsilon)$ is the random jump probability, with $0 < \varepsilon < 1$ and usually set to a value like 0.85.

PageRank scores are usually computed by initializing a PageRank score vector with uniform values $1/N$, and subsequently applying power iterations with the previous iteration's values on the right-hand side of the above equation to recompute the left-hand side. This iteration process is repeated until sufficient convergence, i.e., until the PageRank scores of exhibit only minor changes.

Several similar approaches have been proposed as authority measures such as HITS [15]. For distributed environments, [25, 30, 14, 24] compute authority scores in the presence of disjointly partitioned collections. [32] uses a layered Markov model for Web rank computation.

3. SYSTEM ARCHITECTURE

The operational environment is a P2P-based Web search engine as follows. Every data peer builds local collections and local per-key indexes from autonomous crawls, according to the user's thematic profile, e.g., by using a focused crawler. In our context, a key typically refers to a single term, but can also refer to more sophisticated features like sets of correlated terms [6, 1]. The local per-key indexes consist of inverted lists of documents relevant to the corresponding key, with documents represented by a globally unique docID identifier. With the peers crawling the Web autonomously and independently, a considerable overlap in the document collections (e.g. the set of documents known to individual peers) is expected to occur.

A DHT-style overlay is used to connect all peers. Directory peers are responsible for maintaining very compact, aggregated meta-information about the peers' local indexes, which are continuously disseminated by all peers to the extent that the individual data peers are willing to share\(^1\). More specifically, each directory peer is responsible for the meta-information from all over the network regarding a random subset of all keys. For performance and fault tolerance reasons, the directory metadata for a key may be replicated

\(^1\)Usually, each peer in the DHT is expected to fulfill both roles of a data peer and a directory peer.
across multiple directory peers. The DHT’s lookup method can be used to determine the peer responsible for a particular key, in order to send or retrieve metadata about a key.

Figure 1 illustrates the distributed directory infrastructure. N17, N45, and N76 are directory nodes responsible for keys sports, politics, comedy and music. They maintain the information describing data peers P. The mapping from keys to directory nodes is given by the DHT’s lookup() method that, e.g., maps “sport” to the directory node with nodelist 17.

![Figure 1: System Architecture](image1)

A peer that initiates a query first retrieves all appropriate metadata (i.e., the metadata for all keys that appear in the query) from the distributed directory. It combines this information to identify a judiciously chosen small subset of promising peers for the current query. This step is commonly referred to as resource selection or query routing. Referring back to Figure 1, the initiator of a query containing the key sport can turn to N17 and select promising data peers based on their metadata. The query initiator forwards the query to those peers for evaluation based on their local indexes and eventually merges all incoming query results.

4. DISTRIBUTED AUTHORITY SCORE COMPUTATION

This section illustrates how to incorporate global authority score computation into our generalized system architecture of a distributed collaboration of peers. We will show that the communication overhead of adding JXP authority score computation is negligible, as the necessary information exchange can largely be piggybacked onto existing communication.

While there exist a number of techniques to compute peer authority scores for disjoint collections [30, 14], we highly advocate solutions that compute the actual PageRank scores for all documents individually. Knowledge of global PageRank scores for individual documents can be of great additional benefit for the local document scoring.

The exact algorithm to compute global PageRank scores for our environment, where the Web graph is partitioned over autonomous peers, is orthogonal to this work, as we simply exploit PageRank scores for query routing purposes. We showcase JXP [19], an algorithm to compute global authority scores in a decentralized manner. [19] gives a mathematical proof of the convergence of JXP scores to the global PageRank authority scores, i.e., the scores that would be obtained by a PageRank computation on a hypothetically centralized combined Web graph over all peers.

![Figure 2: Local Web graph, augmented by world node](image2)

Running at each peer, JXP combines standard PageRank computations on the local portion of the Web graph with condensed knowledge on the rest of the network continuously being refined by meetings with other peers. The knowledge about the non-local partition of the Web graph is collapsed into a single dedicated node that is added to the local Web graph, the so-called world node3. It conceptually represents all non-local documents of the Web graph. As such, all documents of the local Web graph that point to non-local documents will create an edge to the world node (cf. Figure 2).

Meetings with other peers in the network are used to exchange local knowledge and to improve the local approximation of global authority scores, illustrated in Figure 3. As a peer learns about non-local documents pointing at a local document, a corresponding edge from the world node to that local document is inserted into the local Web graph4. Each peer locally maintains a list of scores for external documents that point to a local document. The weight of an edge from the world node to a local document reflects the authority score mass that is transferred from the non-local document; if this edge already exists, its weight is updated with the maximum of both scores. The world node contains an additional self-loop link, representing links within non-local pages. The JXP authority score of the world node itself reflects the JXP score mass of all non-local pages. Locally, each peer recomputes its local JXP scores by a standard PageRank power iteration on the local Web graph augmented by the world node.

The JXP algorithm is scalable, as the PageRank power iteration computation is always performed on small local graphs, regardless of the number of peers in the network. The local storage requirements at each peer are independent from the number of remote peers they have previously met and the size of the remote (or even the complete) Web graph, i.e., the size of the local Web graph only reflects the local crawl. The autonomy of peers is fully preserved by the asynchronous nature of communication and computation.

Our system architecture illustrated in the above section involves manifold steps of mutual peer communication that

---

3This is an application of the state-lumping techniques used in the analysis of large Markov models.

4Note that such a meeting does not increase the number of nodes of a peer’s local Web graph.
can naturally be exploited for the JXP authority score computation without arranging additional “meetings”:
when disseminating local metadata to directory peers, when retrieving directory data on the occasion of a global query execution, and when forwarding queries to the peers selected at query routing. The dissemination of metadata, for example, is a well-suited candidate for JXP meetings, as the communication endpoint is determined pseudo-randomly by the keys which are hashed to directory peers. The data that has to be exchanged for updating local JXP authority scores can easily be piggybacked onto this communication, avoiding extra network messages. The additionally incurred bandwidth requirements are also manageable.

5. EXPLOITING PAGERANK FOR QUERY ROUTING

This section introduces two ways to exploit PageRank authority scores for query routing purposes.

Total PR Mass

One way is to use the total PageRank score mass of a peer (i.e., the sum over all PageRank scores accumulated by the local documents), possibly normalized by the total size of the local collection. Given a local PageRank computation continuously running on a local peer, the collection score $s_i$ is simply calculated as follows:

$$s_i = \sum_{d \in i} PR_d$$

where $PR_d$ is the PageRank score of a document $d$. It is easy to disseminate this value along with the previous metadata to the distributed directory, most naively included in every metadata object.

However, the total PageRank score mass does not well reflect a peer’s authority for a particular query; instead, the very same peers would be chosen regardless of the actual query, if they have posted any appropriate metadata. For example, consider the Web page of a computer scientist that has crawled the publications of leading CS university departments. While his local Web graph might have accumulated an above-average PageRank score mass, the peer is ill-suited to evaluate queries about sport, cars, or movies. So the total PageRank score pass of a collection is an inappropriate measure for the expected result quality for a particular query. We need a more expressive way of using authority scores for query routing sensible for concrete queries.

Term-specific PR Mass

In order to become query-specific, we want to describe the PageRank score mass of a peer in more detail. A query consists of several terms, and only those local documents that contain at least one of the query terms can be potentially relevant to this query. So the key idea is to compute term-specific PageRank score masses: for each term individually, we sum up the PageRank score masses of only those documents that contain this term — just as CORI computes term-specific collection subscore.

Query routing based on such term-specific PageRank score masses is straightforward, as the collection score $s_i$ for peer $i$ are computed as:

$$s_i = \sum_{t \in Q, d \in i : t} PR_d$$

where $PR_d$ is the PageRank score of a document $d$, and the summation is over all documents that contain at least one query term. Note that this scheme does not require a separate PageRank computation for every term, but simply sums up the the regular PageRank values for the documents that contain the particular term at query time.

Also remember that our system architecture assumes the dissemination of term-specific metadata to a conceptually global, but physically distributed directory. It is straightforward to include the term-specific PageRank score masses with the appropriate piece of metadata for subsequent use in the query routing process.

While the existence of term-specific quality estimators allows for a fine-grained query routing approach by summing up only potentially relevant portions of the PageRank mass, it postulates term independence, as high score masses w.r.t. terms $a$ and $b$ alone don’t guarantee a single high authority document for the combined query $(a, b)$.

6. COMBINING AUTHORITY AND QUALITY

There exist several popular strategies for query routing based on various kinds of statistical models. We will briefly introduce CORI, one of the most popular strategies and subsequently present a hybrid framework that can benefit from the best of both worlds by combining quality and authority scores for query routing.

6.1 CORI

CORI [9, 8] computes the collection score $s_i$ of the $i$-th peer with regard to a query $Q = \{t_1, t_2, ..., t_n\}$ as $s_i = \sum_{t \in Q} \frac{s_{i,t}}{|t|}$ where $s_{i,t} = \alpha \cdot (1 - \alpha) \cdot T_{i,t} \cdot I_{i,t}$.

The computations of $T_{i,t}$ and $I_{i,t}$ use the number of peers in the system, denoted $np$, and the document frequency (cdf) of term $t$ in collection $i$ for any term $t$ in collection $i$:

$$T_{i,t} = \frac{cdf_{i,t}}{cdf_{i,t} + 50 + 150 \cdot \frac{|V_i|}{np}}$$
where the collection frequency $c_f$ is the number of peers that contain the term $t$. The value $\alpha$ is chosen as $\alpha = 0.4$ \cite{9}.

CORI considers the size $|V_i|$ of the term space of a peer (i.e., the total number of distinct terms that the peer holds in its local index) and the average term space size $|V_i^{avg}|$ over all peers that contain term $t$. Note that, in the absence of global knowledge, we replace $|V_i^{avg}|$ (as defined in \cite{8}) by the average term space size over all peers that contain term $t$ as this information is available from the distributed metadata directory during query routing. In practice, it is difficult to compute the average term space size over all peers in the system (regardless of whether they contain query term $t$ or not). We approximate this value by the average over the metadata in the distributed directory.

It can easily be seen that CORI mainly focuses on the document frequency of the query terms to select the most promising peers for a query, but it does not take into account the quality of these documents. To overcome this problem, we present hybrid approaches to combine CORI-style quality measures with PageRank-style authority scores for query routing.

### 6.2 Hybrid Approaches

We ignore the query-insensitive approaches that are based on the total PageRank authority score mass of a peer and, instead, focus on term-specific PageRank authority score masses. We propose the following linear combination to compute $s_i$, the hybrid collection score of the $i$-th peer:

$$s_i = \sum_{t \in Q} \beta \cdot CORI_{i,t} + (1 - \beta) \cdot PR_{i,t}$$

where $CORI_{i,t}$ and $PR_{i,t}$ are the (both term-specific) CORI and PageRank subscores of peer $i$ for term $t$, respectively.

This approach most closely resembles the approach taken in local query execution, where also statistical quality measures describing a document (like $tf\cdotidf$ or BM25) are coupled with a document’s PageRank score for a total document score. As extreme cases, $\beta = 1$ results in standard CORI-based query routing, while $\beta = 0$ results in query routing solely based on term-specific PageRank score masses.

Please note that for multi-term queries, conceptually, the PageRank score mass for documents containing both terms is accounted for twice. With standard CORI query routing conceptually suffering from the same problem, we are not aware of literature that has identified this as a serious problem. We expect that we will not experience such problems with the PageRank score masses either. Also, we currently do not weight the PageRank score masses of different terms differently, according to some term-specific importance measure, like the idf part does for $tf\cdotidf$ or like the $I_{i,t}$ part does for CORI. While preliminary experiments have not shown a major impact when the PageRank score mass was additionally weighted with idf values, we leave a more systematic study of this issue for future work.

In order to account for the different absolute subscore values yielded by CORI and PageRank, we previously apply the following normalization to all values of $CORI_{i,t}$ and $PR_{i,t}$, generalized to score:

$$I_{i,t} = \frac{\log(\frac{\text{np} + 0.5}{\text{cf}_t})}{\log(\text{np} + 1)}$$

where $\text{np}$ is the number of peers.

7. EXPERIMENTS

7.1 Preliminary Thoughts

The effectiveness of any query routing strategy builds on the hypothesis that peers are mutually discriminative, i.e., the quality measure of choice is not uniformly distributed over all peers. For example, consider an extreme case where the full Web graph is randomly and uniformly distributed over all peers. In this case, the metadata between all peers will only show minor variations, because all terms as well as all relevant or high-authority documents are expected to be uniformly partitioned.
both portals publish different stories and/or the same stories with different importance, the expected quality of articles on Israel is higher on cnn.com, the expected quality of articles about the German Social Security System might be higher for the German newspaper. Needless to say that, for both examples, the PageRank score mass is higher for cnn.com. Similarly, with users of different skill-levels and different prior knowledge on their fields of interest starting Web crawls from different crawl seeds, the distribution of PageRank authority score mass over the peers is not expected to be uniform — not for the total PageRank score mass, and even less for the term-specific PageRank score mass.

To appropriately model this behavior, we have partitioned the data of a Web crawl we had conducted in late 2004 / early 2005 for a total of 10 topics. Several prior experiments have shown that identifying those peers with the “right” topic (for a specific query) is easily done by all query routing strategies, i.e., separating the sport peers from the politics peers for a query about soccer is not a challenging task. However, to find the best order within the peers of the right topic is a challenging task.

For this purpose, we have restricted ourselves to exactly one topic, movies, and distributed only the documents related to movies over a total of 10 peers. With a number of queries related to movies, we want to study the behavior of different strategies regarding the hard task to discriminate peers that share the same topic.

Figure 4 plots the total PageRank score mass for each of the 10 resulting peers, ordered by their PageRank mass, together with their respective collection sizes. For example, Peer 2 has a total PageRank score mass of approximately 0.01 and contains approximately 2900 documents. It can clearly be seen that the total PageRank score masses accumulated at each peer differ highly and do not correspond to the respective collection sizes. Analogously, Figure 5 shows that a small fraction of peers contains a large fraction of the total accumulated PageRank authority score mass. We expect to see this kind of distribution in real-world settings, backing up our hypotheses that size alone does not matter and that the PageRank authority score masses are a discriminative factor even for similar-sized collections. So there is hope that a powerful query routing based on authority scores can have a remarkable impact on the result quality.

### 7.2 Results

We have created a hypothetical combined collection of the 10 movie peers that serves as a reference collection. As a query workload, we have chosen queries from Google’s Zeitgeist archive\(^5\) that match the movies topic, at the time of and slightly prior to the Web crawl. Table 1 shows those queries. Preliminary experiments have shown that all query routing strategies managed to separate these 10 peers for movie-specific queries from the rest of the peers; so we focus on the hard task of ordering within the topic-specific peers. We report on relative recall as the average portion of relevant documents in each collection and query routing was based on an ascending order of relevant documents. Both the hybrid strategy and our strategy based on term-specific PageRank authority score masses outperform the baseline, CORI, in terms of relative recall, in particular for a small number of peers. This is crucial, because the ultimate goal of query routing is to achieve good recall with a very small number of peers. The fact that quality-unaware query routing based on PageRank authority scores only performs as good as our hybrid strategy is an artifact of our small-scale experimental setup; preliminary results for larger-scale environments support this hypothesis. Even though this is a small-scale experiment, this gives first evidence of proof for our hypothesis that authority score masses can be a helpful ingredient in discriminating peers for query routing.

### 8. CONCLUSION / FUTURE WORK

This paper has introduced a framework to integrate knowledge of PageRank authority scores into the query routing process, which is one of the key issues of P2P Web search. Preliminary experiments have shown the potential of this approach, outperforming CORI as a popular baseline algorithm: the number of peers that have to be involved in order to achieve a certain level of relative recall is decreased substantially. This is an important step towards making P2P Web search scalable and feasible.

---

\(^5\)http://www.google.com/press/zeitgeist.html

<table>
<thead>
<tr>
<th>Queries</th>
<th>Document Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>superbowl commercials</td>
<td>earthquake</td>
</tr>
<tr>
<td>harry potter</td>
<td>christopher reeve</td>
</tr>
<tr>
<td>julia roberts</td>
<td>angelina jolie</td>
</tr>
<tr>
<td>desperate housewives</td>
<td>golden globes</td>
</tr>
<tr>
<td>jennifer aniston</td>
<td>academy awards</td>
</tr>
<tr>
<td>blockbuster</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Queries

was used on the reference collection, based on standard tf*idf document scores.

We compare the following instances of our hybrid framework:

- \(\beta = 1\): standard CORI
- \(\beta = 0.5, \beta = 0.1\): hybrid strategies
- \(\beta = 0\): term-specific PageRank masses only

We do not show the experimental results based on the non-query-specific total PageRank authority score mass of a peer, as we have motivated that this can go arbitrarily wrong. Nevertheless we would like to point out that, for our specific setting, basing query routing solely on the total PageRank authority score mass worked remarkably well. This is due to the fact that we take a limited focus on peers with documents related to movies only, so that for our related queries the total PageRank score mass is already a good indicator for the expected result quality. There are no peers in our setting that accumulate a high total PageRank score mass on topics other than movies, which would break the query routing solely based on the total PageRank authority score mass.

Figure 6 plots the relative recall for an increasing number of peers selected by the different query routing strategies. The optimal curve shows a theoretical result where, for each query, we precomputed the relevant documents in each collection and query routing was based on an ascending order of relevant documents. Both the hybrid strategy and our strategy based on term-specific PageRank authority score masses outperform the baseline, CORI, in terms of relative recall, in particular for a small number of peers. This is crucial, because the ultimate goal of query routing is to achieve good recall with a very small number of peers. The fact that quality-unaware query routing based on PageRank authority scores only performs as good as our hybrid strategy is an artifact of our small-scale experimental setup; preliminary results for larger-scale environments support this hypothesis. Even though this is a small-scale experiment, this gives first evidence of proof for our hypothesis that authority score masses can be a helpful ingredient in discriminating peers for query routing.
We are currently conducting more systematic experimental studies on substantially larger document corpora in order to further validate our hypotheses. To support our underlying assumptions on the skewness of the PageRank authority score mass distribution, we are investigating several collections harvested in real-world settings. For the experimental evaluation, we plan to rely on manual user relevance assessments instead of relative recall to more precisely quantify the user-perceived improvements on result quality. Last but not least, we are integrating overlap-awareness into this approach to fully leverage from our research on query routing.

9. REFERENCES


Figure 6: Relative Recall Performance

