A Distributed P2P Link Analysis Based Ranking System

CS 8803 Advanced Internet Application Development Project Proposal
Bhuvan Bamba

Sponsors – Prof. Ling Liu, James Caverlee, Mudhakar Srivatsa, Aameek Singh

Abstract

Link Based approaches are among the most popular ranking approaches employed by search engines. They make use of the inherent linkage based structure of World Wide Web documents assigning each document an importance score. This importance score is based on the incoming links for a document; a document which is pointed to by many high quality documents should have a higher importance score. Googles' highly popular search technology [1] exemplifies the success of the link based ranking algorithms in identifying important pages. However, such link analysis based algorithms suffer from some drawbacks. Googles' PageRank algorithm has an update time extending into months which is not feasible for frequent updating of the system. Secondly, the algorithm is susceptible to manipulation by malicious *Web Spammers*, who manipulate the link based analysis to favor their fake websites. The problem commonly termed as *Web Spam* can seriously hurts the performance of PageRank algorithm, leading the algorithm into providing unjustified high PageRank to spam web pages. In [2], the authors propose the SourceRank approach for enhancing PageRank through source-based link analysis, which can potentially help combat the problem of web spam. In this project, we propose to implement the SourceRank technique on top a P2P crawler Apoidea [3]. We plan to perform experimental analysis using the SourceRank technique on the www.gatech.edu domain and analyse the results to verify the claims made in [2]. In addition, we analyze the results obtained by the above experimentation to enable us to answer an important question raised in [2] – *How can we identify a collection of pages which constitutes a single source?*

Motivation

PageRank-based algorithms suffer from several critical problems, like a very long update time and the problem of web spam. The first problem is a direct consequence of the huge size of the World Wide Web and the inherent expense associated with PageRanks' eigen vector based iterative algorithm. The World Wide Web comprises of Terabytes of data; search engines are capable of crawling and indexing a few billion pages. A web graph comprising of these billions of pages needs to be analyzed by the PageRank algorithm. The end result of the attempts to perform a link-based analysis results in weeks and months spent in updating the PageRanks of the web pages. As the size of the World Wide Web continues to grow there are no clear cut solutions available to deal with the complexity associated with the PageRank computation. An obvious solution is to somehow reduce the size of the graph. The size of the graph can indeed be reduced by grouping together pages according to some criteria. In [2], the authors propose the idea of using a source- based link analysis to reduce the complexity of these operations. A source can be loosely identified as a collection of pages, such as pages belonging to the same domain. Distributed processing of the web graph can also help speed up the processing time for calculating pagerank. A logical division of the web pages into collections is required for such an algorithm to work. SourceRank provides us with a methodology for dividing the web graph into separate collections.

The web graph can be split into different sources and a distributed system can be used to calculate the PageRank of pages within the source and the SourceRank of the particular source in relation to other sources. We plan to implement the SourceRank (and PageRank) algorithm on top of Apoidea [3], a decentralized P2P crawler for the World Wide Web.

SourceRank can also help deal with the problem of web spam. Spammers manipulate PageRanks' link based propagation model to promote their own pages. A user can increase the pagerank of a target page by using "link farms" of pages that point to this particular target page. "Hijacking" of other pages to add links to the target page is another way of promoting the pagerank for a page. Such link based manipulation of web pages is intended to fool the search engine into misjudging the quality of a particular page. Recent results suggest that the amount of Web spam is a significant portion of all Web content – estimates range from 8.1% of pages [4] to 18% of sites [5].

An implementation of SourceRank on top of Apoidea will solve multiple problems. Firstly, it provides a distributed setup for calculation of PageRank (as well as SourceRank). Secondly, SourceRank can also be used to rank the nodes in the P2P system in terms of the content they are holding. This will prove to be highly beneficial when searching across a large number of peers. Each file in the P2P system can now be given a SourceRank and a LocalRank and the overall rank of this file will be a combination of the two factors.

Related Work

Our system builds on top of a web based crawler, Apoidea, developed at the College of Computing at Georgia Institute of Technology. Apoidea is a decentralized P2P crawler for the World Wide Web and provides an ideal setup for implementing a distributed PageRank system. Apoidea is self managing and uses geographical proximity of the web resources to the peers for a better and faster crawl. It uses DHT based protocols to perform URL duplicate and content-duplicate tests. Content-duplication may be associated with the web spam problem which this system attempts to solve.

The SourceRank algorithm attempts to solve the problem of web spam by treating the web as a collection of sources rather than as a collection of pages. SourceRank emphasizes the quality of the source a page belongs to, which can be extremely useful to judge instances of manipulated PageRanks. A page which has a high PageRank but a very low SourceRank has a high probability of being associated with a spamming operation.

Proposed Work

In this project we propose to develop a complete implementation of a distributed P2P Ranking System. Apoidea provides an ideal platform for implementing such a system. This can help speed up PageRank calculations. Secondly, it helps deal with the problem of web spam as described above. Lastly, it also provides a system for ranking files in a P2P system based on the importance of the node on which the file is located as well as the local rank of the file.

We will crawl the www.gatech.edu domain to analyze the performance of the system on this particular domain. Initially we will use a single machine on which we implement the crawler and the SourceRank algorithm. This setup can be easily extended to a distributed setting by using the P2P communication mechanism for Apoidea (and using more than one machine!!). We explore the ability of the system to identify important pages within this domain if we limit the impact of links between pages to a particular source.

[2] fails to address the concept of a source. We will try to analyze our results for the www.gatech.edu domain to refine the concept of a source in the World Wide Web. How does one identify a collection of pages as a source? Since SourceRank tries to prevent manipulation of link based analysis it seems obvious that a source should be identified in a manner which prevents this manipulation by malicious users. The basic concept is to identify sources that are authored by the same person or organization in order to reduce their linkage impact for ranking via SourceRank. Can we identify a relationship between two sites to determine if they are authored by the same person? We list some linkage structures that SourceRank attempts to identify and nullify the affect of (provided by James Caverlee!).

(a) CNNSI and ESPN are related in content, but are clear competitors. They don't link to each other. Linkage patterns are different (internal and external). These sources are related, but not the product of a common author.

(b) Some casino sites run a bunch of sites that are all interconnected. Perhaps there is a linkage pattern to identify this.

(c) Some site runs multiple topically distinct sites that all interconnect. Content linkage is low, but linkage is high.

(d) Some web sites may be laid out according to a common template. So linkage structure may be similar, but they aren't really connected. Need to filter these out.

The system architecture of a single Apoidea peer is shown in Figure 1. We implement the PageRank and SourceRank algorithms on top of each Apoidea peer.

Plan of Action

We plan to develop a C++/Java implementation of the system described above. The implementations for the two systems are available. We need to integrate the code for Apoidea with the SourceRank/PageRank code. This might involve a new implementation of either of the two packages and analyzing the performance of Apoidea and link based ranking. We clearly identify the goals for completion of this project on a weekly basis. The milestones can be listed as below.

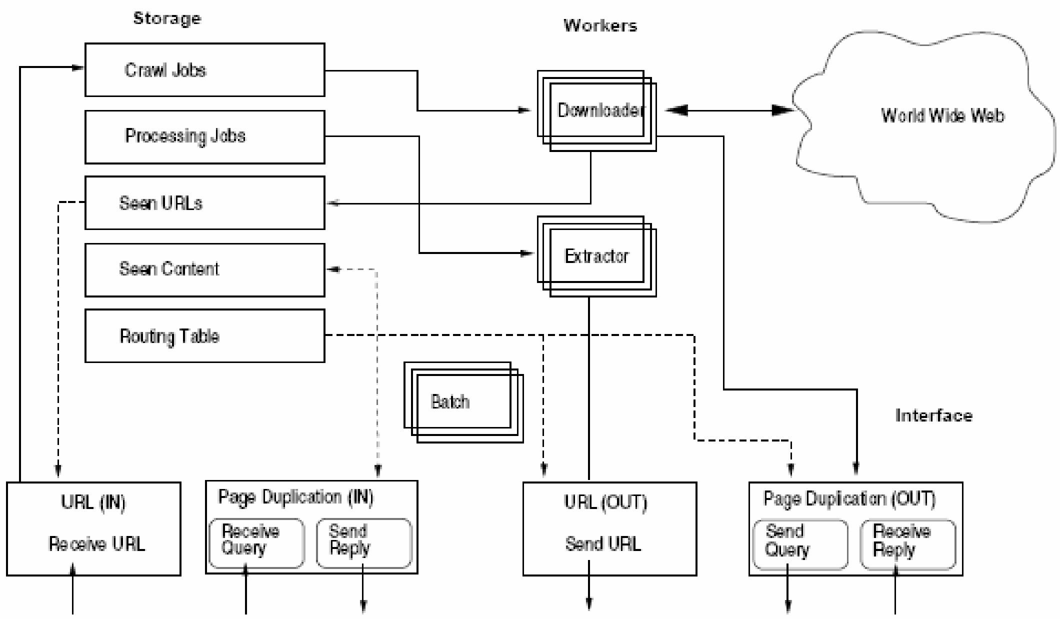

Fig 1. System Architecture of a Single Apoidea Peer

1. Identification of tasks for the project. Completion of project proposal. **(Feb. 12 – Feb.18)**

2. Finalizing the tasks for the project. Acquiring Source Code for Apoidea and PageRank/SourceRank. Identifying a machine for implementation of the system. Deciding on the platform, language to be used for implementation. **(Feb. 19 – Feb. Feb. 25)**

3. System implementation with PageRank/SourceRank implemented on top of Apoidea. **(Feb. 26 – March 18)**

4. Experimentation on the www.gatech.edu domain. Start working on final report. **(March 19 – March 25)**

5. Analyzing results for the experiment and working on source identification problem. Work on project report. **(March 26 – April 8)**

6. Complete project report and work on source identification problem. **(April 9 – April 22)**

Evaluation and Testing Method

1. The primary deliverable in this project is the integrated software package for Apoidea crawler enhanced with the SourceRank/PageRank ranking system.

2. The system evaluation is primarily based on the experimentation and analysis of the www.gatech.edu domain. We analyze the importance scores assigned by our algorithm to the pages in this domain to see if important pages are being ranked higher. We can compare the results with those achieved by the other major search engines.

3. Another parameter for evaluation will be the performance of the SourceRank algorithm for updating the rankings. A comparison of the PageRank and SourceRank algorithms for this particular domain will be provided in order to emphasize the advantages of SourceRank over PageRank.

4. Progress towards identifying the concept of sources (source identification) can also be used as an evaluation parameter.

References

[1] Larry Page, Sergey Brin, R. Motwani, T. Winograd, "*The PageRank Citation Ranking: Bringing Order to the Web*", Technical report, Stanford University, 1998

[2] James Caverlee and Ling Liu, "*Enhancing PageRank through Source-Based Link Analysis*", Manuscript under preparation.

[3] Aameek Singh, Mudhakar Srivatsa, Ling Liu, Todd Miller, "*Apoidea: A Decentralized Peer-to-Peer Architecture for Crawling the World Wide Web* ", Proceedings of the SIGIR 2003 Workshop on Distributed Information Retrieval, Lecture Notes in Computer Science, Volume 2924

[4] D. Fetterly, M. Manasse, and M. Najork, "*Spam, damn spam, and statistics*", WebDB, 2004.

[5] Z. Gy¨ongyi, H. Garcia-Molina, and J. Pedersen, "*Combating Web spam with TrustRank*", VLDB, 2004.