Fast Focused Crawling with URL Significance and Dynamic Backtracking

Sai Ganesh Sitharaman
Department of Computer Science
Texas A&M University
College Station, TX-77843.
ssai@cs.tamu.edu

Richard Furuta
Department of Computer Science
Texas A&M University
College Station, TX-77840.
furuta@cs.tamu.edu

ABSTRACT

World Wide Web (WWW) is one of the primary vehicles of information transfer and sharing among the people today. Using simple HTTP protocol and a powerful hypertext markup language HTML, WWW has effectively turned itself into a large decentralized digital library that is highly dynamic and constantly undergoes creation, deletion, and modification. Experimental observations show that more than 100,000 pages are updated or changed everyday around the globe. Manual collection, organization and finding relevant pages are thus out of question without the help of powerful collection schemes, crawlers and search techniques. In this article, we propose an state-of-the-art accelerated focused crawling technique that achieves quick crawling and convergence in several less CPU cycles while utilizing URL-based measures for finding the effective distance between the current document and the document to be fetched. We propose 5 new measures for making the focused crawl far more effective than today's techniques. These may be used in addition to today's context analysis techniques such as similarity distance measure of current document with the new one. Our goal is to develop techniques that preempt systems from fetching a large number of unrelated documents and thus making the system converge faster. We propose 5 major metrics towards making a crawl effective and justify the usefulness of each of them using live experiment results performed with 5 major online portal sites. Our measures include: (1) Relative frequency ranking of URL to be fetched from the current document *u* to the potential document *v* – **RFU(*u*, *v*)**, (2) *Basename* frequency of URL occurrence in the current document *u* – **BFU(*u*)**, (3) In-degree or the current document hop number from original seed document – **ID(*u*)**, (4) Associated Relevance Feedback to back-associate and change URL significance with respect to the fetched document *v* and adjustment – **ARF(*v*, *u*)**, and (5) Web server roundtrip-delay estimate from the current server to the intended URL – **RTT(*u*, *v*)**. In our proposal, we present some of the pitfalls of the current measures and demonstrate investigations to prove why our techniques can lead to a powerful yet fast-focused crawler.

1 INTRODUCTION

World Wide Web (WWW) is used in wide variety of information exchange from a simple file transfer to complicated multimedia audio and video applications. It has become one of the foremost medium to transfer and share information among people across all the globe. WWW has also turned out to be a vast collection of digital information archive. Users requiring finding topic-specific information typically start with few known popular starting points and explore from there to seek new relevant pages closer to the topic. WWW has effectively turned itself into a large decentralized digital library that is highly dynamic today and is constantly created, deleted, and modification.

Statistics gathered from over two million web pages specified by over 25,000 users of a web clipping service. Experimental observations show that more than 100,000 pages are updated or changed everyday around the globe [1]. The data indicates that the time between modifications and the rate of such changes are exponentially distributed and can even be modeled in a memory less model. Today's web search and classification mechanisms can be roughly divided into three major categorizes: Manual human intervened filters, Automatic classification using tools and powerful classification engines, and Natural Languages Processing (NLP) machines that uses the content found in the web page to intelligently analyze and fetch relevant pages of interest.

In this article, we propose a fast-track accelerated focussed crawling technique that achieves quick crawling and convergence in several less CPU cycles at the same time utilizing URL-based measures for effectively finding the distance between the current document and the document to be retrieved. We make use of several useful measures in order to evaluate the URL distance measure between the given current document *u* and the potential document *v* to be fetched.

2 PROBLEM DEFINITION AND MOTIVATION

Existing techniques on content analysis measures a vector distance between the frequently occurring words in the target document with respect to some document to be fetched. Just as the analytical distance between two points on a 2D geometrical grid is calculated, the distance metric for such a document is calculated through functions such as *Inverse Document Frequency (IDF).* IDF is a primary discriminatory factor showing the appearance of words in one document against another. Since we have not seen the entire collection yet, the crawling process estimates the IDF factors from the pages that have been crawled, or from some reference IDF terms computed at some other time. This makes the crawling process really slow and takes numerous hops to converge from the given document to a target.

Techniques have been proposed that makes use of URLs and links to find the relative distance between document to fetch. Such techniques are called *link analysis* techniques. Link analysis is based on the theory that a document pointed to by one document is likely to be similar to it, or that two documents linked to by the same document might be related to each other. The proof for the claim lies in the probabilistic argument that two un-correlated pages are far more *distant* metric-wise as compared to the links that are directly available from the current document. That is, hyperlinks are more similar than two randomly chosen pages. Our main question in the investigation is whether it is possible to advance the link analysis with further enhancements and to introduce practical HTML-based analysis techniques to the model.

We have 2 primary motivations to revisit and study link analysis technique for better crawling. Firstly, we are interested in studying link analysis technique to seek the relevant documents in as fewer number of steps as possible and secondly, we are intrigued by the nature of the links present in the current document themselves that can help identify the content of the potential document to be fetched in advance, preventing pre-fetching all the documents. To demonstrate, we show later using sample analysis that pre-fetching can indeed be significantly detrimental and several CPU cycles can indeed be saved by simpler smarter analysis techniques. Furthermore, we are not aware of articles that studied appropriate feedback system from the fetched page back to the original page in order to dynamically change the ranking process itself in the original page. In our proposal, we call this effective backtracking to signify the change of path that may be taken since the last calculated path from the original document.

We arrived at several new techniques in our proposal through systematic study of existing methods. Following are some of the primary motivations to study practical link analysis techniques for focussed crawling: (1) Dynamism of the web and rate of changes, (2) Crawl learning techniques and relevant feedback, (3) Measure of hits and success rate such as harvest rate, efficiency, time bound and precision of fetched pages, (4) Reduction in the overhead of fetching all pages as opposed to focussed fetches, and (5) Diversity of the large URL measures between documents. Using each such metric, we found that newer practical techniques can be introduced that addresses each such issue. We describe each one of them in section 4 proposal.

Our proposal is organized as follows. Section 3 describes some of the main motivating articles in this area that helped us analyze differently. Section 4 introduces the various performance measures, the crux of our measures and provides rationale behind each of our ideas based on live simulation experiments performed with 5 major portal sites. We conclude with future research and tasks that need to be worked on in section 5.

3 RELATED WORK

Chakrabarti *et al.* [1] introduce a new focused crawling scheme that selectively seeks out pages that are relevant to a specific topic. Classification and distillation is done and compared against such exemplary documents. They introduce a goal-oriented crawling with two major components including a classifier and a distiller. The basic idea is to collect a large number of several access points to many relevant links in the beginning of the crawl process.

In yet another novel article, Chakrabarti *et al.* [3] describe an accelerated feedback technique to make use of the HTML anchor HREF tag tree structure in the Dynamic Object Model (DOM) to speed up the focused crawler. Their model includes a form of relevant feedback from the trainer within the crawler system. The supervised training module takes online feedback to adapt itself and learn about the URL ordering which in turn is fed to the crawler to perform focused crawling. We partially take motivation from this article that analyzes the various component of a crawler and that takes appropriate feedback. However, our backtracking feedback is much different in nature as will be explained in the next section.

Cho *et al.* [4] present yet another useful crawl accelerating technique to efficiently find significant URLs. They present 5 URL ordering techniques and important metrics to obtain important pages significantly faster. Their techniques uses page similarity, page ranks, back-link count and location metric to measure a value that represents the worthiness of the page to be fetched. The central goal is to devise a function that scores high for the relevant links and hence fetches them early.

Fast focused crawling with url significance and backtracking feedback mechanism – a proposal

Earlier sections described the need and possibilities of newer analysis techniques to accelerate focused crawling while still maintaining the importance and relevance of each link accessed. As each path is traversed, the relevance is evaluated and most importantly, if the same relevance holds in future is of primary importance.

We propose 5 novel measures for making the focused crawl far more effective than today's techniques. These may be used in conjunction to today's context analysis techniques such as similarity distance measure of current document with the new one. Our goal is to develop techniques that preempt systems from fetching a large number of unrelated documents to make the system converge faster. We introduce a new measure namely a *Guided Path Measure* (GPR) that uses the associated relevance feedback from those fetched pages only to dynamically change the route taken (the shortest error route) from a given document *u* to a target document *v*.

Our five major metric towards effective crawl include: (1) Relative frequency ranking of URL to be fetched from the current document *u* to the potential document *v* – **RFU(*u*, *v*)**, (2) *Basename* frequency of URL occurrence in the current document *u* – **BFU(*u*)**, (3) In-degree or the current document hop number from original seed document – **ID(*u*)**, (4) Associated Relevance Feedback to back-associate and change URL significance with respect to the fetched document *v* and adjustment – **ARF(*v*, *u*)**, and (5) Web server roundtrip-delay estimate from the current server to the intended URL – **RTT(*u*, *v*)**. We present some of the pitfalls of the current measures and demonstrate investigations to prove why our techniques together add up a powerful yet fast-focused crawler

Section 4.1 describes the performance measures such as Harvest rate (HR), Efficiency, and Guide Path Measure (GPR). Sections 4.2 to 4.6 describe in detail each of these measure and their relative merits using experiment results.

3.1 PERFORMANCE MEASURES

Several performance and evaluation measures are provided to evaluate the effectiveness of the crawl. Some of these include Harvest Rate (HR) and Crawl efficiency. We introduce yet another measure namely, the Guided Path Measure (GPR) that gives a measure of the number of times a link is successfully traversed towards the target and the number of times it has backtracked (and hence took an incorrect decision).

Harvest rate (HR) or precision is a ratio of relevant pages collected to the total pages collected in the process of the crawl. Efficiency is the measure of highly quickly relevant documents is downloaded early in the crawling process itself. Usually highly correlated links are attempted first and hence should get a significant result in the beginning but this is not a must.

Guided Path Measure (GPR) uses the Associated Relevance Feedback (ARF) measure from those fetched pages only to dynamically change the route taken (the shortest error route) from a given document *u* to a target document *v*. The goal of the GPR is to prevent the crawling process to not get misguided by other page ranking analysis techniques and page analysis details. Simple form of GPR function takes into account the number of successful hits to the relevant pages but significantly reduces the GPR value if uncorrelated pages are fetched in the process.

3.2 RELATIVE FREQUENCY RANKING (RFU)

To demonstrate our main theme, we show the following live statistics from some of the primary important site deployed in today's Internet. We poll some of the useful information from the following sites and give an estimate as what the relative frequency ranking means. Following are the main sites polled (1) Yahoo! Movies, (2) Yahoo! Health, (3) Google Computers, (4) Google Science, and (5) BBC Society and Culture. RFU unit of each of these are packed in the following figure 1 below in that order.

We used a simple URL crawler that starts from the given base URL (in this case, the five of the above listed) and prints out the relative and absolute URL to be traversed from that document. Our experimental tool ran on Windows XP with crawler code using Microsoft MSHTML interface. MSHTML Internet Client SDK comes as a part of the Win32 licensing agreement and allows developers to interface with COM and browser APIs. MSHTML contains code for HTML parsing, rendering and exposes Dynamic HTML Object Model (DOM) to the system.

Our measure of RFU is calculated as follows. RFU always lies between 0 and 1.

$$\boldsymbol{RFU} = \frac{\text{URL with same domain}}{\text{Total URL of all domains}} * \frac{1}{\text{Number of unique URL domains}}$$

We observe at least two interesting aspects from the plot. First, the higher the measure of RFU, the better the page is found to be organized and hence enables faster focused crawling. For instance, both of Google Computers and Science scored (RFU = 0.163, 0.165 correspondingly) significantly higher RFU units as compared to Yahoo! sites. Secondly, it also shows that there is a vast majority of the links that may not be completely relevant and hence add extra overhead in terms of CPU processing. One such instance is the lowest scored BBC Society and Culture (RFU = 0.005). When analyzed the reason for this lower value, we found that BBC had as many as 20 URL domains (such as advertisements, shopping, entertainment, etc.) that were not directly related to society or culture. Moreover, the active number of Society & Culture-related URLs was significantly lower compared to other unrelated URLs. It thus appears that our RFU is a good measure of the relative significance of relevant URLs from that page.

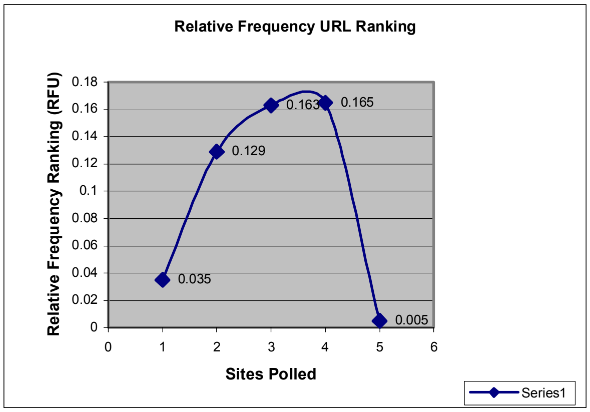

Fig 1: Relative Frequency Ranking (RFU) of sites polled (12/05/2002)

3.3 Basename Frequency URL (BFU)

Basename URLs are relatively simple measures that potentially point to *similar* or *equivalent* pages from a given document. While RFU units measure the average number of correlated URL domains, BFU count the number of same or similar URLs in the document. We use the same set of poll sites and look for the number of such basename strings. Figure 2 below plots the measure for various sites polled.

Our measure of BFU is calculated as follows. BFU always lies between 0 and 1.

$$BFU = \frac{\text{URL with same basename}}{\text{Total URL of all domains}} * \frac{1}{\text{Number of Unique Basenames}} \qquad (2)$$

We observe two more interesting aspects of the properties of the pages and how links are organized based on the BFU measure. We maintain that BFU can range between 0 and 1 and the higher the BFU, the better it is suitable for enhancing focused and accurate crawling. We find that Yahoo! movies and BBC Society & Culture are one of worst developed. While the number of unique basename are poor, there is a wide distribution of (of about 22 unique basename) basename in Yahoo! movies. Similarly, BBC Society & Culture has a very low number of unique society or culture-related basename. The observation here is that several of them are not related to the topic that we are interested in and hence these links should fall low in our queue (for instance, advertisements, shopping, and entertainment under BBC Society & Culture).

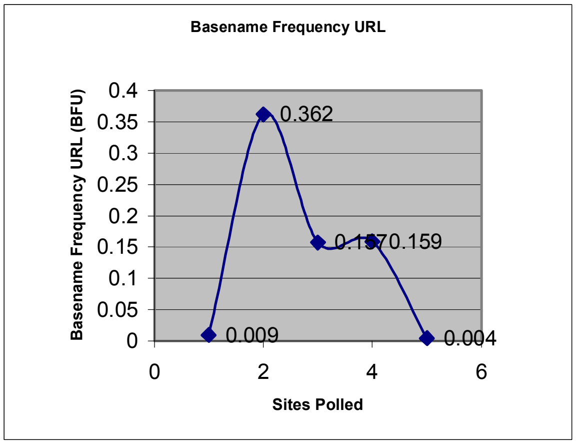

Fig 2: Basename Frequency URL (BFU) of sites polled (12/05/2002)

However, we do maintain that both Google Computers and Science are still well organized and the BFU figures are comparable to RFU. BFU measures for both Computers and Science are relatively the same for both and are high in the ladder making it a better organized information site.

3.4 In-Degree (ID)

In-degree of a document refers to the number of times this document is accessed with reference to another document. In-degree (ID) count best represents a simple popularity measure of that page. Such a popular page is well fit to be potentially retrieved next and hence moves up the queue. ID is related to BFU but not entirely. While ID requires URL to be unique, BFU only considers the necessary base URL string to make a decision. Thus ID requires counting the number of *same* URL pages (the consequence of clicking on the page has to be a valid HTML document!) from the current document. We argue that even though this factor is already taken into consideration in BFU, unique URLs represent something more interesting and popularity that should be separately taken into consideration.

Our primary motivation of this factor came through the maintenance of Google Page Rank (PR) measure. Google's PR calculation of a web page is based on external and internal linking of a site, as well as on-page criteria of the web page being linked to as well as the web page being linked from. PR is in some ways related to link popularity, but the calculation is dependant on the quality and strength of the links, not just the number of links

3.5 ASSOCIATED RELEVANCE FEEDBACK (ARF)

Chakrabarti *et al.* [3] develop a relevance feedback model between a learner that constantly receives feedback from fetching various pages and a crawler that takes instructions from the learner. The baseline learner provides feedback through an apprentice so that it can improve on the job of accelerated crawling. Other neural networks based training techniques have also evolved and applied to this kind of crawling process. The simple linear preceptor model, training happens with delta rule and simple linear additions. Convergence in such models is slow. The fundamental idea in such models is to provide some form of positive or negative feedback based on the current path that is taken and how such a path converges.

Our proposal develops an Associated Relevant Feedback (ARF) model in which we take advantage of the details of those pages fetched so far back to the original fetch schedule process such that the crawling is accelerated and focus maintained. We are motivated to look at this aspect of the ARF due to two reasons: firstly, to ensure that our RFU and BFU measures indeed are developing the shortest least error route between the current document *u* and the target resource *v*, and secondly, to develop a feedback system to guide appropriately if the pages are not well organized. We justify the latter using some of the statistics show earlier. We found that sites such as BBC Society & Culture contain a wide distribution of irrelevant links and at the same time has less significant URLs to traverse. If the crawler does not make use of the RFU and BFU measures of such a site, it is possible that many of the irrelevant links are traversed naturally due to their larger existence (an apparent popularity). Thus, the crawler is misguided in evaluating the feedback. To prevent this from happening, we associate a feedback measure with those fetched pages (for instance, BBC shopping from BBC Society & Culture). Using page analysis technique of the fetched page, we provide a negative feedback to the original crawler (since we are not interested in shopping links from BBC Society). A negative feedback value to the crawler immediately senses the urgency of stopping further crawling of similar links from the page.

3.6 ROUND-TRIP DELAY ESTIMATE (RTT)

The idea behind using the round trip delay estimate in fetching web pages is not new. But we are not aware of articles that use this technique for better crawling and on an average reducing the number of CPU cycles required to complete the crawling process. Several studies do understand that document download speed is one of the critical bottlenecks towards faster fetching. These techniques provide alternative techniques for massive crawl. It is more useful to perform a single massive crawl instead of individual crawls and analyzing at the same time. In our proposal, we thus include the usage of such a delay estimate. The estimate is particularly useful when several unique domains are yet to be crawled but that can potentially be delayed until further, if their page relevance is not high.

Figure 3 demonstrates and compares the average round trip delays of our 5 sites polled namely *movies.yahoo.com, health.yahoo.com, www.google.com,* and *www.bbc.co.uk*. Once the average delays are known, the idea is to use this metric to push those sites with higher page relevance up in the queue to be fetched before others.

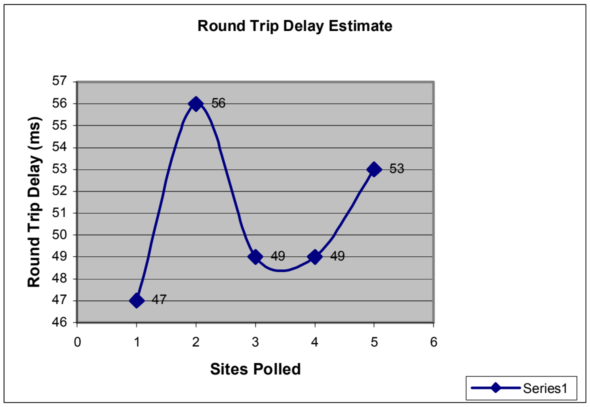

Fig 3: Round Trip Delay Estimate for 5 polled sites (12/05/2002)

4 CONCLUSION AND FUTURE WORK

We introduced 5 new measures and demonstrated experimentally that there can be value-added crawling that can perform well than techniques available today. Our techniques can be applied in conjunction with page analysis such that they together perform well. We also introduced feedback measures that help in further guiding along the path.

Our contribution in this article is many fold: first, we introduced several newer techniques applied to accelerate crawling and develop importance measure of the page. We introduce a Guided Path Measure (GPR) to show that the measures indeed work. Finally, we demonstrated the usefulness and uniqueness of the approach using live simulations from 5 major sites often visited.

Several future work can be undertaken starting for this simple proposal. We first have to develop a complete empirical model to show the form of the function is correct for RFU and BFU measures.

Determining ID for pages is also a must and can be correlated to BFU to a large extent. Furthermore, it is possible to define a Fetch Relevance function **FR(*u, v*)** that is a function of all the above metrics suitably constructed that represents the overall score for potentially fetching the document *v* from the current *u*. Finally, we are yet to demonstrate that the average Harvest Rate and overall crawl efficiency remains consistent and better than others for our scheme.

5 REFERENCES

[1] Brian E. Brewington , George Cybenko, How dynamic is the Web?, Computer Networks: The International Journal of Computer and Telecommunications Networking, v.33 n.1-6, p.257-276, June 2000.

[2] S. Chakrabarti and B. Dom, and M. van den Berg. Focused Crawling: A New Approach for Topic-Specific Resource Discovery, Proceedings of the 8th World Wide Web conference, 1999.

[3] S Chakrabarti, K Punera, and M Subramanyam. Accelerated focused crawling through online relevance feedback. In Proc. 11th International World Wide Web Conference. ACM Press, 2002.

[4] Cho, J., Garcia-Molina, H. and Page, L. Efficient Crawling Through URL Ordering, in Proceedings of the 7th World Wide Web Conference (Brisbane, Australia, Apr 1998).