FOCUSED CRAWLING TECHNIQUES

Pankaj Mishra
C.Sc. Dept.
S.R.Govt. College (W), Amritsar

ABSTRACT

The need for more and more specific reply to a web search query has prompted researchers to work on focused web crawling techniques for web spiders. Variety of lexical and link based approaches of focused web crawling are introduced in the paper highlighting important aspects of each.

General Terms

Focused Web Crawling, Algorithms, Crawling Techniques.

Keywords

Focused, Crawler, Lexical, Link, Similarity, HITS, ARC, DOM, Graph, Fish Search, Shark Search

INTRODUCTION

Crawler periodically traverses the web and collects information about web documents [6] for search engine to be added to its database and indexed. Examples of Crawler: WebCrawler [2], Mercator [3], WebFountain [4], UbiCrawler [5]. A crawler may employ Breath First(begins at a particular web page and then explores all pages that it can reach by using only one hyperlink from the original page. Once it has exhausted all web pages at that one level, it explores all of the web pages that can be reached by following only one hyperlink from any page that was discovered at level one.) or Depth First(A depth-first search proceeds by following a chain of hyperlinks down as far as possible. In contrast to breadth-first search, hyperlinks on a given page are not fully exhausted before the crawler goes to the next level page.) methods to search the web for new pages.A crawler identifies the location of a document by its URL. The crawler maintains a list of unvisited URLs called the *frontier*. Each *crawling loop* (see Fig.1) involves picking the next URL to crawl from the frontier, fetching the page corresponding to the URL through HTTP, parsing the retrieved page to extract the URLs and application specific information, and finally adding the unvisited URLs to the frontier.

(Fig: 1) Data Structures required for the algorithm

```
        p0 is a valid web URL hyperlink
        Q is a queue of valid hyperlinks
        P is a set of web pages
        H is a set of hyperlinks
Algorithm
$Q \leftarrow P0$        {insert P0 into the queue Q}
while $|Q| \neq \emptyset$ do
$p \leftarrow Q$   {get head of queue Q}
retrieve web page $p$
$P \leftarrow P \cup p$
extract URL hyperlinks contained in $p$ into $H$
   for all $h \in H,\ h \notin Q$ do
      $Q \leftarrow h$
   end for
end while
```

The basic crawler has the following components [20]:

- Frontier
- History and Page Repository
- Fetching
- Parsing
  - *URL Extraction and Canonicalization*
  - *Stoplisting and Stemming*

The first generation of crawlers on which most of the web search engines are based rely heavily on traditional graph algorithms, such as breadth-first or depth-first traversal, to index the web(see Fig. 2). A focused crawler explores the web using a best-first search according to a specific topic; i.e. it downloads only topic-relevant documents in its path (see Fig. 3) instead of downloading all links as in case of a general crawler.

There are two main issues regarding the focused crawling discussed as follows:

- The crawlers need to identify from a list of unvisited URLs the ones most likely to contain relevant information.
- The crawlers should avoid irrelevant or bad quality documents by determining the quality and reputation of each document.

Fig: 2 General Web Crawling

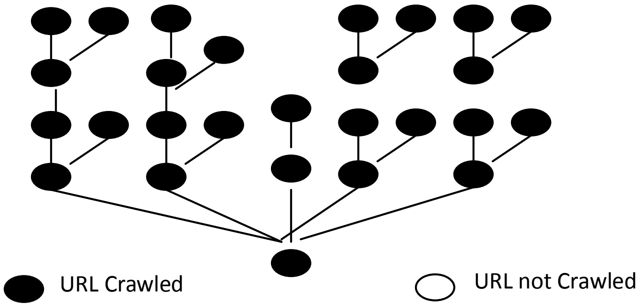
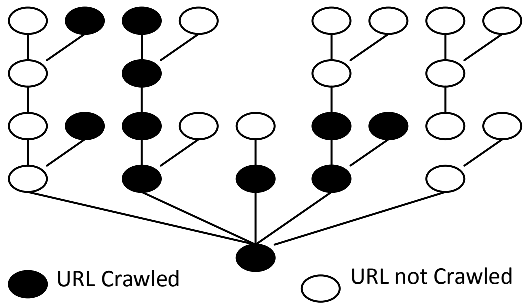
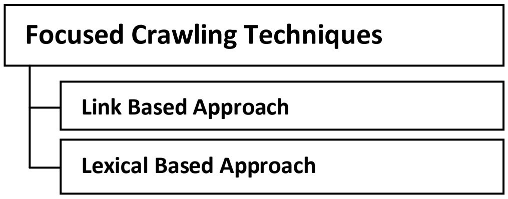


Fig: 3 Focused Web Crawling



The classification of crawling techniques to retrieve relevant, high quality web pages is:

(Fig: 4) Focused Crawling Techniques



1. LEXICAL BASED APPROACH

In lexical-based approach, the actual HTML content of a Web page is analyzed to induce information about the page. Some of the focused crawlers based on this technique have been investigated below.

- **Fish-Search System:** [15], one of the initial crawler, whose key principle is: "*It takes as input a seed URL and a search query, and dynamically builds a priority list (initialized to the seed URL) of the next URLs (hereafter called nodes) to be explored.*"
- **Shark Search System:** Mapuccino [16] is another crawler that uses shark-search algorithm. Sharksearch algorithm is an improved algorithm of fish-search algorithm. It overcomes some limitations of fish-search. One improvement is using a fuzzy score to replace the binary (relevant/irrelevant) evaluation of document relevance, i.e., a score *between 0* and *1* (*0* for no similarity, *1* for perfect "conceptual" match) rather than a binary value.
- **Focused Crawler based on Category Taxonomy:** A Focused Crawler (first put forward by S. Chakrabarti) [17] was designed which is a web resource discovery system based on canonical topic taxonomy with examples. The focused crawler has three main components: a classifier, distiller and a crawler. The focused crawler downloads the seed web pages, computes the weight of each words in the pages based on TF.IDF (Term frequency / Inverse Document Frequency) weighting scheme [18] and generates a set of topic keywords based on the top N highest weight keywords in these seed pages. The crawler continues to follow the out-links until the queue is empty or an user-specified limit is reached.

2. LINK BASED APPROACH

Link-based approaches [19-27] have drawn increasing attention in recent years. Usually, the larger the number of in-links, the better a page is considered to be. The rationale is that a page referenced by more people is likely to be more important than a page that is seldom referenced. Some of the focused crawling techniques / systems based on this approach have been investigated below.

- **Similarity based Crawling System :** A similarity-based crawler is the one that orders URLs having target keyword in anchor text or URL [19]takes into account various measures of importance for a page such as similarity to a driving query, number of pages pointing to this page, page rank, location, etc..
- **HITS Algorithm based Crawling System:** Klienberg [20] developed HITS (Hyper-link-induced- topic search) algorithm in a way that, an authority page is defined as a high quality page related to a particular topic or search query and a hub page is one that provides pointers to other authority pages. Based upon this, a web page is associated with an *Authority Score* and a *Hub Score* that is calculated to identify the web page context. The basic principle here is the following mutually reinforcing relationship between hubs and authorities. A good *hub page* points to many good *authority pages*. A good authority page is pointed to by many good hub pages.
- **Automatic Resource Compiler (ARC) based Crawling System:** The goal of ARC [21] is to automatically compile a resource list on any topic that is broad and well-represented on the web. The algorithm has three phases: a **search-and-growth** phase, a **weighting** phase, and an **iteration-and-reporting** phase. The construction of resource lists in ARC is considerably simpler and more efficient. Moreover, the results [21] suggest that it is possible to automate most of the process of compiling resource lists on the web through the combination of link and text analysis.
- **DOM (Document Object Model) based Focused Crawling :** According to S. Chakrabart et al [22,23] DOM parser creates a tree structure in memory that contains the document's data [24]. The DOM tree, which is already available in the memory, can be used in retrieving the tags and their text portions quickly.
- **Context Graph based Focused Crawling:** Diligenti et al [26] present a focused crawling algorithm that builds a model for the context within which topically relevant pages occur on the web. This context model can capture link hierarchies within which valuable pages occur, as well as model content in documents that frequently co-occur with relevant pages. There are two phases in the algorithm
  - **Initialization Phase:** The first phase (initialization) aims to extract the context within which target pages are typically found, and encodes this information in a context graph. It constructs a set of context graphs and associated classifiers for each of the seed documents.
  - **Crawling Phase:** The second phase is the crawling phase. It uses the classifiers to guide the search, and performs online updating of the context graphs. Their test results show that the best results are achieved when the topic is so that the target content can be reliably co-located with pages from a different category, and where common hierarchies do not exist or are not implemented uniformly across different web-sites.

Below(see Fig:5) is a tabulation of various approaches for focused web crawling or focused web searching.

Fig: 5 Main Characteristics Of Various Focused Crawling Techniques

System		Main Characteristics	References
Lexical based Approach	Fish Search System	Crawling algorithm based on depth-first search. Heuristics determine the selection of the documents that are to be retrieved.	[15]
	Shark Search System	Based on a fuzzy score to replace the binary relevant/irrelevant) evaluation of document relevance, i.e., a score *between 0* and *1* (*0* for no similarity whatsoever, *1* for perfect "conceptual" match) rather than a binary value.	[16]
	Focused Crawler based on Category Taxonomy	Based on a category tree based document taxonomy and seed documents which build a model for classification of retrieved pages into categories. Uses semi-supervised learning	[17]
	Focused Crawler based on Similarity Computing Engine	Retrieves information related to a user-specified topic from the Web using TF.IDF weighting scheme to find the topic keywords set to represent the topic, and uses vector space model to determine whether web pages are relevant to the topic or not.	[18]
Link Based Approach	Similarity based Crawler	Calculates the Page rank score on the graph induced by pages downloaded so far and then using this score as a priority of URLs extracted from a page.	[17]
	HITS	Identifies the web page context based on Authority Score and Hub score associated with the web page. The basic principle here is the following mutually reinforcing relationship between hubs and authorities. A good hub page points to many good authority pages. A good authority page is pointed to by many good hub pages.	[10]
	ARC	Automatically compiles a resource list on any topic that is broad and well-represented on the web. Based on modified weighted Authority Score and Hub score. Uses Anchor window wherenin the system looks on either side of the href for a window of *B* bytes, to increase the authority weight.	[20]
	DOM based Focused Crawler	Locates regions and subtrees of pages using DOM tree of a web page, gets favorable treatment in propagating link-based popularity and implicitly suppressing propagation of popularity to regions with noisy links. Identifies and extract hub regions relevant to a topic and guides the hub and authority reinforcement to work on a selected, highly relevant subgraph of the web.	[21], [22]
	Context Graph based Focused Crawler	Uses a context graph to train a set of classifiers to assign documents to different categories based on their expected link distance to the target. Naïve Bayes classifier is used for each layer of the graph.	[26]

3. REFERENCES

1) www.worldwidewebsize.com, 15 March, 2008
2) Pinkerton, B., "Finding what people want: Experiences with the WebCrawler", Proc. of the First World Wide Web Conference, Geneva, Switzerland, 1994
3) Heydon, A. and Najork, M., "Mercator: A scalable, extensible Web crawler", World Wide Web, 2(4):219–229, 1999
4) Edwards, J., McCurley, K. S., and Tomlin, J. A., "An adaptive model for optimizing performance of an incremental web crawler", Proc. of the Tenth Conference on World Wide Web: 106-113. Hong Kong: Elsevier Science, 2001
5) Boldi, P., Santini, M., and Vigna, S., "Do your worst to make the best: Paradoxical effects in pagerank incremental computations", Proc. of the third Workshop on Web Graphs (WAW), volume 3243 of Lecture Notes in Computer Science, pages 168-180, Rome, Italy. Springer, 2004
6) Junghoo Cho, Hector Garcia-Molina, " Parallel Crawlers", Proceedings of the 11th International World Wide Web Conference, Technical Report, UCLA Computer Science, 2002.
7) http://www.metacrawler.com
8) http://www.dogpile.com
9) S. RaviKumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal, " Stochastic models for the Web graph" In *FOCS*, pages 57-65, Nov. 2000.
10) K. Bharat, A. Border, "A technique for measuring the relative size and overlap of public web search engines", Proc. of 7th WWW conference, 1998
11) S. Lawrence, C.L. Giles, "Searching the world wide web", Science, 280:98-100, April 1998
12) Martin Ester, Matthias Grob, Hans-Peter Kriegel, "Focused Web Crawling: A Generic Framework for specifying the user interest and for Adaptive crawling strategies", Proc. of 27th International Conference on Very Large databases(VLDB '01), 2001
13) Naresh Chauhan, A.K. Sharma, "A Comparative Analysis of Focused Crawling Techniques", Proceedings of National Conference on IT, Panipat, March 2006
14) Naresh Chauhan, A.K. Sharma, "Focused Web Crawling: Techniques and Issues", *Proceedings* of National Conference on Information Technology: Emerging Engineering Perspectives & Practices (ITEEPP'07), Patiala, April 2007
15) P. De Bra, G.-J. Houben, Y. Kornatzky, R. Post, "Information Retrieval in distributed hypertexts", Proc. of RIAO'94, Intelligent Multimdia, Information Retrieval systems and management, New York, 1994
16) M. Hersovici, M. Jacovi, Y. Maarek, D. Pelleg, M. Shtalheim, " The Shark-Search Algorithm – an application: tailored web site mapping", Computer Networks and ISDN systems, Special Issue on 7th WWW conference, Brisbane, Australia, 30(1-7), 1998
17) S. Chakrabarti, M. Van Den Berg, B. Dom, "Focused Crawling: A New Approach to Topic specific web resource discovery", Proc. Of 8th International WWW conference, Toronto, Canada, May,1999
18) Yang Yongsheng, Wang Hui, "Implementation of Focused Crawler", COMP630D Course Project Report
19) Junghoo Cho, Hector Garcia-Molina, L.Page, "Efficient crawling through URL ordering", Proc. of 7th International WWW conference, Brisbane, Australia, April, 1998
20) J.Kleinberg, "Authoritative Sources in a Hyperlinked Environment", Proceedings of the 9th ACMSIAM Symposium on Discrete Algorithms, San Francisco, California, USA, 1998.
21) S. Chakrabarti, B. Dom, D. Gibson, J. Klienberg, P. Raghvan, S. Rajagopalan, "Automatic Resource Compilation by analyzing hyperlink structure and associated text", 7th WWW Conference, 1998
22) S. Chakrabarti, Mukul Joshi, Vivek Tawade, " Enhance Topic Distillation using text, markup tags, and hyperlinks", Proc. of SIGIR'01, Sep. 2001
23) S. Chakrabarti, Kunal Punera, Mallela Subramanyam, "Accelerated Focused Crawling through Online relevance feedback", Proc. of WWW conference, 2002
24) Deitel & Deitel, Goldberg, "Internet and World Wide Web: How to Program", Pearson Education, Inc., 2004
25) Filippo Menczer, "Lexical and Semantic clustering by web links", Journal of the American Society for Information Science and Technology, 55(14):1261-1269, 2004
26) Michelangelo Diligenti, Frans Coetzee, Steve Lawrence, C. Lee Giles, and Marco Gori, "Focused crawling using context graphs", VLDB '00: Proceedings of the 26th International Conference on Very Large Data Bases (San Francisco, CA, USA), Morgan Kaufmann Publishers Inc., pp. 527–534, 2000.
27) Gautam pant, Padmini Srinivasan, "Link Contexts in Classifier-Guided Topical Crawlers", IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No.1, Jan. 2006.