2012 International Conference on Information and Knowledge Management (ICIKM 2012)
IPCSIT vol.45 (2012) © (2012) IACSIT Press, Singapore

Focused Web Crawler

Ayoub Mohamed H. Elyasir[1], Kalaiarasi Sonai Muthu Anbananthen[2]

Multimedia University, Melaka, Malaysia
[1]Email: ayoub_it@msn.com, [2]Email: kalaiarasi@mmu.edu.my

Abstract. Web crawler is a program that traverses the internet based on automated manner to download the web pages or partial pages' contents according to the user requirements. Small amount of work on web crawler has been published for various reasons. Web crawling and its techniques are still in the shadow and possess many secrets due to its involvement in the giant search engine applications where they tend to obscure it, as it is the secret recipe for their success. There is also secrecy involved to protect against search spamming and ranking functions so web crawling methods are rarely published or publically announced. Web crawler is as an important and fragile component for many applications, including business competitive intelligence, advertisements, marketing and internet usage statistics. In this work, we compare between the two main types of web crawlers: standard and focused to choose one of them and apply it in our latter framework for opinion mining in the education domain.

Keywords: Web Crawling, Focused Crawler, Search Engine, Uniform Resource Locator, Canonicalization

1. Introduction

Over the last decade, the World Wide Web has evolved from a number of pages to billions of diverse objects. In order to harvest this enormous data repository, search engines download parts of the existing web and offer Internet users access to this database through keyword search. One of the main components of search engines is web crawler. Web crawler is a web service that assists users in their web navigation by automating the task of link traversal, creating a searchable index of the web, and fulfilling searchers' queries from the index. That is, a web crawler automatically discovers and collects resources in an orderly fashion from the internet according to the user requirements. Different researchers and programmers use different terms to refer to the web crawlers like aggregators, agents and intelligent agents, spiders, due to the analogy of how spiders and crawlers traverses through the networks, or the term (robots) where the web crawlers traverses the web using automated manner.

To date various applications for web crawler have been introduced and developed to perform particular objectives. Some of these applications are malicious in a way that penetrates the users' privacy by collecting information without their permission. However, web crawlers have applications with significant impact on the market as it is mainly involved in the search engines, business competitive intelligence and internet usage statistics. Unfortunately, web crawling is still in the shadow and possess many secrets due to its involvement in the giant search engines applications where they tend to obscure it, as it is the secret recipe for their success. There is also secrecy involved to protect against search spamming and ranking functions so web crawling methods are rarely published or publically announced.

2. Web Crawling

Searching is the most prominent function all over the web, internet user tend to look into various topics and interests every time he/she surfs the web. Web crawling is the technical synonym for internet searching which giant search engines provide nowadays to the users at no cost. No client side elements needed outside the browser to crawl through the web, crawling consists of two main logistics parts: crawling, the process of finding documents and constructing the index; and serving, the process of receiving queries from searchers and using the index to determine the relevant results.

We crawling is the means by which crawler collects pages from the Web. The result of crawling is a collection of Web pages at a central or distributed location. Given the continuous expansion of the Web, this crawled collection guaranteed to be a subset of the Web and, indeed, it may be far smaller than the total size of the Web. By design, web crawler aims for a small, manageable collection that is representative of the entire Web.

Web crawlers may differ from each other in the way they crawl web pages. This is mainly related to the final application that the web crawling system will serve. Crawlers classified based on their functionality to standard and focused. Standard crawler has a random behavior for collecting web pages while focused crawler has a guided way to do the traversal process. Figure one below shows that standard crawler branches generally through the nodes (web pages) regardless of the node domain, while focused crawler traverses deeper and narrower toward a specific node domain. Another remark in Figure 1 is the starting node (root) which is same for both standard and focused crawler.

A focused crawler ideally would like to download only web pages that are relevant to a particular topic and avoid downloading all others. It predicts the probability that a link to a particular page is relevant before actually downloading the page. A possible predictor is the anchor text of links. In another approach, the relevance of a page is determined after downloading its content. Relevant pages sent to content indexing and their contained URLs added to the crawl frontier; pages that fall below a relevance threshold are discarded

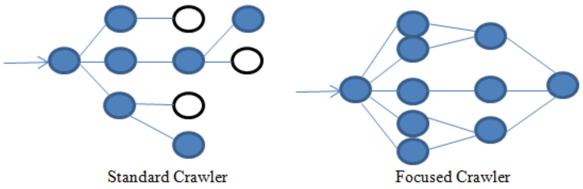

Figure 1: Standard versus Focused Crawler

2.1 Comparison

Table 1 below shows the difference between standard and focused web crawlers:

Table 1: Comparison between standard and focused web crawler

Crawler Type	**Standard Web Crawler**	**Focused Web Crawler**
Synonym	No-selection web crawler	Topical web crawler Note: Topical web crawler may refer to a focused web crawler that does not use a classifier but a simple guided technique instead
Introduced by	Various contributions	Chakrabarti, Berg and Dom (1999)
Definition	Traverses the internet in an automated and pre-defined manner with random web pages collection	Traverses the same way as standard web crawler, but it only collects pages similar to each other based on the domain, application, inserted query, etc…
Path Searching	Random search and may lose its way while traversing the web	Narrowed searching path with steady performance
Web Pages	Not necessarily related or linked to each other	Must be related to a particular criteria

Starting Seed	Root seed	Root seed with dependency on the web search engine to provide the starting point
Ending Seed	Some random seed	Relevant to the traversed seeds
Robustness	Prone to URL distortions	Robust against any distortions because it follows a relevant URL path
Discovery	Wide radius but less relevant web pages	Narrow radius with relevant web pages
Resource Consumption	Less resource consumption because of the basic path traversing algorithms	High resource usage especially with distributed focused crawlers that run on multiple workstations
Page weight	Assigns value to the web page for priority reasons	Assigns value to the web page for priority and relativity (credits) reasons
Performance Dependency	Crawling is independent	Crawling is dependent on the link richness within a specific domain
Flexibility	Customizable with lots of options	Less flexible due to its dependency
Classifier	No classification involve but rely heavily on traditional graph algorithms like depth-first traversal or breadth-first	Classify to relevant or not relevant pages using Naïve Bayesian, Decision Trees, Breadth-First, Neural Network or Support Vector Machine (SVM) which outperforms the other methods especially when it is applied on page contents and link context
Overall	Less resource consumption and performance	Higher resource consumption and performance with high quality collections of web pages

From the comparison, we find that focused crawler is a better choice for traversing through the internet. The ability to narrow the search radius with specific and guided path makes the focused crawler quality wise in terms of web page collection, in which attempts to identify the most related links, and skips the off-topic links. Malformed URL that causes a false direction in the crawling path easily distorts standard crawler, because it follows each link using breadth first algorithm and downloads them all on its crawling way. Resources consumption is less in standard crawling, focused are still better choice though, due to the availability of the computing resources nowadays under reasonable prices. Focused crawler is not as customizable as standard crawler but the first has the ability to classify the results based on page contents and link context. Additionally, commercial applications prefer focused crawler because of the domain dependency and restriction where some crawl through topics and others crawl based on regions and locations.

3. Review on Focused Web Crawler

Search engine web sites are the most visited in the internet worldwide due to their importance in our daily life. Web crawler is the dominant function or module in the entire World Wide Web (WWW) as it is the heart of any search engine. Standard crawler is a powerful technique for traversing the web, but it is noisy in terms of resource usage on both client and server. Thus, most of the researchers focus on the architecture of the algorithms that are able to collect the most relevant pages with the corresponding topic of interest. The term focused crawling was originally introduced by (Chakrabarti, Berg, & Dom, 1999) which indicates the crawl of topic-specific web pages. In order to save hardware and network resources, a focused web crawler analyzes the crawled pages to find links that are likely to be most relevant for the crawl and ignore the irrelevant clusters of the web.

Chakrabarti, Berg and Dom (1999) descriped a focused web crawler with three components, a classifier to evaluate the web page relevance to the chosen topic, a distiller to identify the relevant nodes using few link layers, and a reconfigurable crawler that is governed by the classifier and distiller. They try to impose various features on the designed classifier and distiller: Explore links in terms of their sociology, extract specified web pages based on the given query, and explore mining communities (training) to improve the crawling ability with high quality and less relevant web pages.

Web page credtis problem was addressed by (Diligenti, Coetzee, Lawrence, Giles and Gori, 2000), in which the crawl paths chosen based on the number of pages and their values. They use context graph to capture the link hierarchies within which valuable pages occur and provide reverse crawling capabilities for more exhaustive search. They also concluded that focused crawling is the future and replacement of standard crawling as long as large machine resources are available.

Suel and Shkapenyuk (2002) described the architecture and implementation of optimized distributed web crawler which runs on multiple work stations. Their crawler is crash resistant and capable of scaling up to hundreds of pages per second by increasing the number of participating nodes.

CROSSMARC approach was introduced by (Karkaletsis, Stamatakis, Horlock, Grover and Curran, 2003). CROSSMARC employs language techniques and machine learning for multi-lingual information extraction and consists of three main components: site navigator to traverse web pages and forward the collected information to (Page filtering) and (Link scoring). Page filtering is to filter the information based on the given queries and link scoring sets the threshold likelihood of the crawled links.

Baeza-Yates (2005) highlighted that crawlers in the search engine are responsible for generating the structured data and they are able to optimize the retrieving process using focused web crawler for better search results. Castillo (2005) Designed a new model for web crawler, which was integrated with the search engine project (WIRE) and provided an access to metdata that enables the web crawling process. He emphasized on how to catpure the most relevant pages as there are infinite number of web pages in the internet with weak association and relationship. He also stated that traversing only five layers from the home page is enough to get overview snapshot of the corressponding web site, hence it saves more bandwidth and avoid network congestion.

Rungsawang and Angkawattanawit (2005) attempt to enhance the crawling process by involving knowledge bases to build the experience of learnable focused web crawlers. They show results of an optimized focused web crawler that learn from the information collected by the knowledge base within one domain or category. They have proposed three kinds of knowledge bases to help in collecting as many relevant web pages and recognize the keywords related to the topic of interest.

Liu, Milios and Korba (2008) presented a framework for focused web crawler based on Maximum Entropy Markov Models (MEMMs) that enhanced the working mechanism of the crawler to become among the best Best-First on web data mining based on two metrics, precision and maximum average similarity. Using MEMMs, they were able to exploit multiple overlapping and correlated features, including anchor text and the keywords embedded in the URL. Through experiments, using MEMMs and combination of all features in the focused web crawler performs better than using Viterbi algorithm and dependent only on restricted number of features.

Batsakis, Petrakis and Milios (2009) evaluated various existing approaches to web crawling such as Breadth-First, Best-First and Hidden Markov Model (HMM) crawlers. They proposed focused web crawler based on HMM for learning, leading to relevant pages paths. They combined classic focused web crawler attributes with the ideas from document clustering to result in optimized relevant path analysis.

Liu and Milios (2010) developed their previous framework (Liu, Milios and Korba, Exploiting Multiple Features with MEMMs for Focused Web Crawling 2008), in which they proposed two probabilistic models to build a focused crawler, MEMMs and Linear-chain Conditional Random Field (CRF) as shown in Figure 2. Their experiments show improvements on the focused crawling and gave advantage over context graph (Diligenti, et al. 2000) and their previous model.

We provided an explanatory literature review only on focused crawler because of its popularity among the researchers community. Various methods and algorithms embedded in the focused crawler to boost the traversing performance and produce quality results such as, context graph, statistical classifiers, machine learning techniques, information theory and entropy. Other techniques are in use by the giant search engines to enhance their crawling oriented services based on various criteria like locations and regions, inserted search query, language, user browsing history and page ranks.

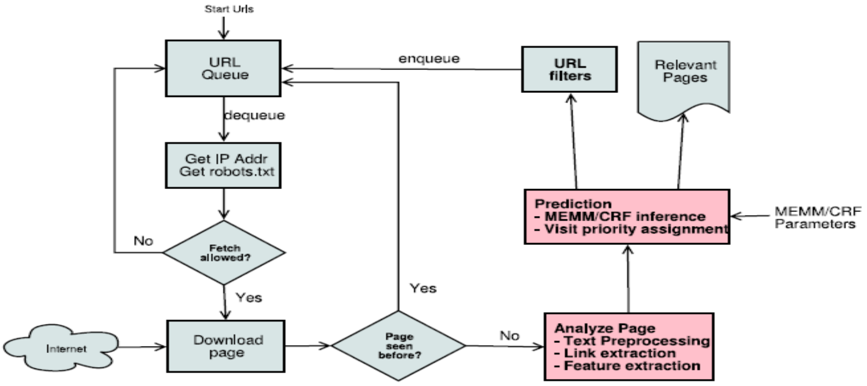

Figure 2: Focused Crawling using MEMM/CRF Models (Liu and Milios, Probabilistic Models for Focused Web Crawling 2010)

4. Conclusion

Web crawling is an initial component in many applications including search engines and opinion mining frameworks. We compared between standard and focused web crawlers to understand which one is better and apply it in our opinion mining framework in a future work.

5. Acknowledgement

This work was supported by Project “Mining Opinions Using Combination of Word Sense and Bag of words approach for Educational Environments”, funded by the Fundamental Research Grant Scheme under the 2010–2012.

6. References

[1] Baeza-Yates, Ricardo. "Applications of Web Query Mining." Springer, 2005: 7-22.

[2] Batsakis, Sotiris, Euripides Petrakis, and Evangelos Milios. "Improving the performance of focused web crawlers." ELSEVIER, 2009.

[3] Castillo, Carlos. "EffectiveWeb Crawling." ACM, 2005 .

[4] Chakrabarti, Soumen, Martin van den Berg, and Byron Dom. "Focused crawling: a new approach to topic-specific Web resource discovery." Elsevier, 1999.

[5] Diligenti, Coetzee, Lawrence, Giles, and Gori. "Focused Crawling Using Context Graphs." 26th International Conference on Very Large Databases, VLDB 2000. Cairo, Egypt, 2000. 527–534.

[6] Karkaletsis, Vangelis, Konstantinos Stamatakis, James Horlock, Claire Grover, and James R. Curran. "Domain-SpecificWeb Site Identification: The CROSSMARC Focused Web Crawler." Proceedings of the 2nd International Workshop on Web Document Analysis (WDA2003). Edinburgh, UK, 2003.

[7] Liu, Hongyu, and Evangelos Milios. "Probabilistic Models for Focused Web Crawling." Computational Intelligence, 2010.

[8] Liu, Hongyu, Evangelos Milios, and Larry Korba. "Exploiting Multiple Features with MEMMs for Focused Web Crawling." NRC, 2008.

[9] Rungsawang, Arnon, and Niran Angkawattanawit. "Learnable topic-specific web crawler." Science Direct, 2005: 97–114.

[10] Suel, Torsten, and Vladislav Shkapenyuk. "Design and Implementation of a High-Performance Distributed Web Crawler." Proceedings of the IEEE International Conference on Data Engineering. 2002.