Web Crawling

Web Crawling

Christopher Olston

Yahoo! Research
701 First Avenue
Sunnyvale, CA, 94089
USA

olston@yahoo-inc.com

Marc Najork

Microsoft Research
1065 La Avenida
Mountain View, CA, 94043
USA

najork@microsoft.com

the essence of knowledge

Boston – Delft

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is C. Olston and M. Najork, Web Crawling, Foundation and Trends® in Information Retrieval, vol 4, no 3, pp 175–246, 2010

ISBN: 978-1-60198-322-0

Foundations and Trends® in Information Retrieval
Volume 4 Issue 3, 2010

Editorial Scope

Foundations and Trends® in Information Retrieval will publish survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2010, Volume 4, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Foundations and Trends® in
Information Retrieval
Vol. 4, No. 3 (2010) 175–246

DOI: 10.1561/1500000017

Web Crawling

Christopher Olston[1] and Marc Najork[2]

[1] *Yahoo! Research, 701 First Avenue, Sunnyvale, CA, 94089, USA*
olston@yahoo-inc.com
[2] *Microsoft Research, 1065 La Avenida, Mountain View, CA, 94043, USA*
najork@microsoft.com

Abstract

This is a survey of the science and practice of web crawling. While at first glance web crawling may appear to be merely an application of breadth-first-search, the truth is that there are many challenges ranging from systems concerns such as managing very large data structures to theoretical questions such as how often to revisit evolving content sources. This survey outlines the fundamental challenges and describes the state-of-the-art models and solutions. It also highlights avenues for future work.

Contents

1

Introduction

A *web crawler* (also known as a *robot* or a *spider*) is a system for the bulk downloading of web pages. Web crawlers are used for a variety of purposes. Most prominently, they are one of the main components of web search engines, systems that assemble a corpus of web pages, index them, and allow users to issue queries against the index and find the web pages that match the queries. A related use is web archiving (a service provided by e.g., the Internet archive [77]), where large sets of web pages are periodically collected and archived for posterity. A third use is web data mining, where web pages are analyzed for statistical properties, or where data analytics is performed on them (an example would be Attributor [7], a company that monitors the web for copyright and trademark infringements). Finally, web monitoring services allow their clients to submit standing queries, or *triggers*, and they continuously crawl the web and notify clients of pages that match those queries (an example would be GigaAlert [64]).

The raison d'être for web crawlers lies in the fact that the web is not a centrally managed repository of information, but rather consists of hundreds of millions of independent web content providers, each one providing their own services, and many competing with one another. In other words, the web can be viewed as a federated information repository, held together by a set of agreed-upon protocols and data formats, such as the Transmission Control Protocol (TCP), the Domain Name Service (DNS), the Hypertext Transfer Protocol (HTTP), the Hypertext Markup Language (HTML) and the robots exclusion protocol. So, content aggregators (such as search engines or web data miners) have two choices: They can either adopt a pull model where they will proactively scour the web for new or updated information, or they could try to establish a convention and a set of protocols enabling content providers to push content of interest to the aggregators. Indeed, the Harvest system [24], one of the earliest search services, adopted such a push model. However, this approach did not succeed, and virtually all content aggregators adopted the pull approach, with a few provisos to allow content providers to exclude all or part of their content from being crawled (the robots exclusion protocol) and to provide hints about their content, its importance and its rate of change (the Sitemaps protocol [110]).

There are several reasons why the push model did not become the primary means of acquiring content for search engines and other content aggregators: The fact that web servers are highly autonomous means that the barrier of entry to becoming a content provider is quite low, and the fact that the web protocols were at least initially extremely simple lowered the barrier even further — in fact, this simplicity is viewed by many as the reason why the web succeeded where earlier hypertext systems had failed. Adding push protocols would have complicated the set of web protocols and thus raised the barrier of entry for content providers, while the pull model does not require any extra protocols. By the same token, the pull model lowers the barrier of entry for content aggregators as well: Launching a crawler does not require any a priori buy-in from content providers, and indeed there are over 1,500 operating crawlers [47], extending far beyond the systems employed by the big search engines. Finally, the push model requires a trust relationship between content provider and content aggregator, something that is not given on the web at large — indeed, the relationship between content providers and search engines is characterized by both mutual dependence and adversarial dynamics (see Section 6).

1.1 Challenges

The basic web crawling algorithm is simple: Given a set of seed Uniform Resource Locators (URLs), a crawler downloads all the web pages addressed by the URLs, extracts the hyperlinks contained in the pages, and iteratively downloads the web pages addressed by these hyperlinks. Despite the apparent simplicity of this basic algorithm, web crawling has many inherent challenges:

- **Scale.** The web is very large and continually evolving. Crawlers that seek broad coverage and good freshness must achieve extremely high throughput, which poses many difficult engineering problems. Modern search engine companies employ thousands of computers and dozens of high-speed network links.
- **Content selection tradeoffs.** Even the highest-throughput crawlers do not purport to crawl the whole web, or keep up with all the changes. Instead, crawling is performed selectively and in a carefully controlled order. The goals are to acquire high-value content quickly, ensure eventual coverage of all reasonable content, and bypass low-quality, irrelevant, redundant, and malicious content. The crawler must balance competing objectives such as coverage and freshness, while obeying constraints such as per-site rate limitations. A balance must also be struck between exploration of potentially useful content, and exploitation of content already known to be useful.
- **Social obligations.** Crawlers should be "good citizens" of the web, i.e., not impose too much of a burden on the web sites they crawl. In fact, without the right safety mechanisms a high-throughput crawler can inadvertently carry out a denial-of-service attack.
- **Adversaries.** Some content providers seek to inject useless or misleading content into the corpus assembled by

the crawler. Such behavior is often motivated by financial incentives, for example (mis)directing traffic to commercial web sites.

1.2 Outline

Web crawling is a many-faceted topic, and as with most interesting topics it cannot be split into fully orthogonal subtopics. Bearing that in mind, we structure the survey according to five relatively distinct lines of work that occur in the literature:

- Building an efficient, robust and scalable crawler (Section 2).
- Selecting a traversal order of the web graph, assuming content is well-behaved and is interconnected via HTML hyperlinks (Section 4).
- Scheduling revisitation of previously crawled content (Section 5).
- Avoiding problematic and undesirable content (Section 6).
- Crawling so-called "deep web" content, which must be accessed via HTML forms rather than hyperlinks (Section 7).

Section 3 introduces the theoretical crawl ordering problem studied in Sections 4 and 5, and describes structural and evolutionary properties of the web that influence crawl ordering. Section 8 gives a list of open problems.

References

[1] S. Abiteboul, M. Preda, and G. Cobena, "Adaptive on-line page importance computation," in *Proceedings of the 12th International World Wide Web Conference*, 2003.

[2] E. Adar, J. Teevan, S. T. Dumais, and J. L. Elsas, "The web changes everything: Understanding the dynamics of web content," in *Proceedings of the 2nd International Conference on Web Search and Data Mining*, 2009.

[3] Advanced Triage (medical term), http://en.wikipedia.org/wiki/Triage#Advanced_triage.

[4] A. Agarwal, H. S. Koppula, K. P. Leela, K. P. Chitrapura, S. Garg, P. K. GM, C. Haty, A. Roy, and A. Sasturkar, "URL normalization for de-duplication of web pages," in *Proceedings of the 18th Conference on Information and Knowledge Management*, 2009.

[5] C. C. Aggarwal, F. Al-Garawi, and P. S. Yu, "Intelligent crawling on the world wide web with arbitrary predicates," in *Proceedings of the 10th International World Wide Web Conference*, 2001.

[6] D. Ahlers and S. Boll, "Adaptive geospatially focused crawling," in *Proceedings of the 18th Conference on Information and Knowledge Management*, 2009.

[7] Attributor. http://www.attributor.com.

[8] R. Baeza-Yates and C. Castillo, "Crawling the infinite web," *Journal of Web Engineering*, vol. 6, no. 1, pp. 49–72, 2007.

[9] R. Baeza-Yates, C. Castillo, M. Marin, and A. Rodriguez, "Crawling a country: Better strategies than breadth-first for web page ordering," in *Proceedings of the 14th International World Wide Web Conference*, 2005.

[10] B. Bamba, L. Liu, J. Caverlee, V. Padliya, M. Srivatsa, T. Bansal, M. Palekar, J. Patrao, S. Li, and A. Singh, "DSphere: A source-centric approach to

crawling, indexing and searching the world wide web," in *Proceedings of the 23rd International Conference on Data Engineering*, 2007.

[11] Z. Bar-Yossef and M. Gurevich, "Random sampling from a search engine's index," in *Proceedings of the 15th International World Wide Web Conference*, 2006.

[12] Z. Bar-Yossef, I. Keidar, and U. Schonfeld, "Do not crawl in the DUST: Different URLs with similar text," in *Proceedings of the 16th International World Wide Web Conference*, 2007.

[13] L. Barbosa and J. Freire, "Siphoning hidden-web data through keyword-based interfaces," in *Proceedings of the 19th Brazilian Symposium on Databases SBBD*, 2004.

[14] L. Barbosa and J. Freire, "An adaptive crawler for locating hidden-web entry points," in *Proceedings of the 16th International World Wide Web Conference*, 2007.

[15] L. Barbosa, A. C. Salgado, F. de Carvalho, J. Robin, and J. Freire, "Looking at both the present and the past to efficiently update replicas of web content," in *Proceedings of the ACM International Workshop on Web Information and Data Management*, 2005.

[16] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates, "Link-based characterization and detection of web spam," in *Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web*, 2006.

[17] A. Benczúr, K. Csalogány, T. Sarlós, and M. Uher, "Spamrank — fully automatic link spam detection," in *Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web*, 2005.

[18] K. Bharat and A. Broder, "Mirror, mirror on the web: A study of host pairs with replicated content," in *Proceedings of the 8th International World Wide Web Conference*, 1999.

[19] K. Bharat, A. Broder, J. Dean, and M. Henzinger, "A comparison of techniques to find mirrored hosts on the WWW," *Journal of the American Society for Information Science*, vol. 51, no. 12, pp. 1114–1122, 2000.

[20] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Communications of the ACM*, vol. 13, no. 7, pp. 422–426, 1970.

[21] P. Boldi, B. Codenotti, , M. Santini, and S. Vigna, "UbiCrawler: A scalable fully distributed web crawler," *Software — Practice & Experience*, vol. 34, no. 8, pp. 711–726, 2004.

[22] P. Boldi, B. Codenotti, M. Santini, and S. Vigna, "Structural properties of the African web," in *Poster Proceedings of the 11th International World Wide Web Conference*, 2002.

[23] P. Boldi, M. Santini, and S. Vigna, "Paradoxical effects in pagerank incremental computations," *Internet Mathematics*, vol. 2, no. 3, pp. 387–404, 2005.

[24] C. M. Bowman, P. B. Danzig, D. R. Hardy, U. Manber, and M. F. Schwartz, "The Harvest information discovery and access system," in *Proceedings of the 2nd International World Wide Web Conference*, 1994.

[25] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," in *Proceedings of the 7th International World Wide Web Conference*, 1998.

[26] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the web," in *Proceedings of the 9th International World Wide Web Conference*, 2000.

[27] A. Broder, M. Najork, and J. Wiener, "Efficient URL caching for World Wide Web crawling," in *Proceedings of the 12th International World Wide Web Conference*, 2003.

[28] A. Z. Broder, S. C. Glassman, and M. S. Manasse, "Syntactic clustering of the web," in *Proceedings of the 6th International World Wide Web Conference*, 1997.

[29] M. Burner, "Crawling towards eternity: Building an archive of the world wide web," *Web Techniques Magazine*, vol. 2, no. 5, pp. 37–40, 1997.

[30] J. Callan, "Distributed information retrieval," in *Advances in Information Retrieval*, (W. B. Croft, ed.), pp. 127–150, Kluwer Academic Publishers, 2000.

[31] J. Callan and M. Connell, "Query-based sampling of text databases," *ACM Transactions on Information Systems*, vol. 19, no. 2, pp. 97–130, 2001.

[32] J. P. Callan, Z. Lu, and W. B. Croft, "Searching distributed collections with inference networks," in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1995.

[33] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg, "Automatic resource compilation by analyzing hyperlink structure and associated text," in *Proceedings of the 7th International World Wide Web Conference*, 1998.

[34] S. Chakrabarti, M. van den Berg, and B. Dom, "Focused crawling: A new approach to topic-specific web resource discovery," in *Proceedings of the 8th International World Wide Web Conference*, 1999.

[35] K. C.-C. Chang, B. He, C. Li, M. Patel, and Z. Zhang, "Structured databases on the web: Observations and implications," *ACM SIGMOD Record*, vol. 33, no. 3, pp. 61–70, 2004.

[36] K. Chellapilla and A. Maykov, "A taxonomy of JavaScript redirection spam," in *Proceedings of the 16th International World Wide Web Conference*, 2007.

[37] H. Chen, M. Ramsey, and C. Yang, "A smart itsy bitsy spider for the web," *Journal of the American Society for Information Science*, vol. 49, no. 7, pp. 604–618, 1998.

[38] S. Chien, C. Dwork, R. Kumar, D. R. Simon, and D. Sivakumar, "Link evolution: Analysis and algorithms," *Internet Mathematics*, vol. 1, no. 3, pp. 277–304, 2003.

[39] J. Cho and H. García-Molina, "The evolution of the web and implications for an incremental crawler," in *Proceedings of the 26th International Conference on Very Large Data Bases*, 2000.

[40] J. Cho and H. García-Molina, "Parallel crawlers," in *Proceedings of the 11th International World Wide Web Conference*, 2002.

[41] J. Cho and H. García-Molina, "Effective page refresh policies for web crawlers," *ACM Transactions on Database Systems*, vol. 28, no. 4, pp. 390–426, 2003.

[42] J. Cho and H. García-Molina, "Estimating frequency of change," *ACM Transactions on Internet Technology*, vol. 3, no. 3, pp. 256–290, 2003.

[43] J. Cho, J. García-Molina, and L. Page, "Efficient crawling through URL ordering," in *Proceedings of the 7th International World Wide Web Conference*, 1998.

[44] J. Cho and A. Ntoulas, "Effective change detection using sampling," in *Proceedings of the 28th International Conference on Very Large Data Bases*, 2002.

[45] J. Cho and U. Schonfeld, "RankMass crawler: A crawler with high personalized PageRank coverage guarantee," in *Proceedings of the 33rd International Conference on Very Large Data Bases*, 2007.

[46] E. G. Coffman, Z. Liu, and R. R. Weber, "Optimal robot scheduling for web search engines," *Journal of Scheduling*, vol. 1, no. 1, 1998.

[47] CrawlTrack, "List of spiders and crawlers," http://www.crawltrack.net/crawlerlist.php.

[48] A. Dasgupta, A. Ghosh, R. Kumar, C. Olston, S. Pandey, and A. Tomkins, "The discoverability of the web," in *Proceedings of the 16th International World Wide Web Conference*, 2007.

[49] A. Dasgupta, R. Kumar, and A. Sasturkar, "De-duping URLs via rewrite rules," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.

[50] B. Davison, "Recognizing nepotistic links on the web," in *Proceedings of the AAAI-2000 Workshop on Artificial Intelligence for Web Search*, 2000.

[51] G. T. de Assis, A. H. F. Laender, M. A. Gonçalves, and A. S. da Silva, "A genre-aware approach to focused crawling," *World Wide Web*, vol. 12, no. 3, pp. 285–319, 2009.

[52] J. Dean and M. Henzinger, "Finding related pages in the world wide web," in *Proceedings of the 8th International World Wide Web Conference*, 1999.

[53] P. DeBra and R. Post, "Information retrieval in the world wide web: Making client-based searching feasible," in *Proceedings of the 1st International World Wide Web Conference*, 1994.

[54] M. Diligenti, F. M. Coetzee, S. Lawrence, C. L. Giles, and M. Gori, "Focused crawling using context graphs," in *Proceedings of the 26th International Conference on Very Large Data Bases*, 2000.

[55] C. Duda, G. Frey, D. Kossmann, and C. Zhou, "AJAXSearch: Crawling, indexing and searching web 2.0 applications," in *Proceedings of the 34th International Conference on Very Large Data Bases*, 2008.

[56] J. Edwards, K. S. McCurley, and J. A. Tomlin, "An adaptive model for optimizing performance of an incremental web crawler," in *Proceedings of the 10th International World Wide Web Conference*, 2001.

[57] D. Eichmann, "The RBSE spider — Balancing effective search against web load," in *Proceedings of the 1st International World Wide Web Conference*, 1994.

[58] D. Fetterly, N. Craswell, and V. Vinay, "The impact of crawl policy on web search effectiveness," in *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2009.

[59] D. Fetterly, M. Manasse, and M. Najork, "Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages," in *Proceedings of the 7th International Workshop on the Web and Databases*, 2004.
[60] D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener, "A large-scale study of the evolution of web pages," in *Proceedings of the 12th International World Wide Web Conference*, 2003.
[61] R. Fielding, "Maintaining distributed hypertext infostructures: Welcome to MOMspider's web," in *Proceedings of the 1st International World Wide Web Conference*, 1994.
[62] A. S. Foundation, "Welcome to Nutch!," http://lucene.apache.org/nutch/.
[63] W. Gao, H. C. Lee, and Y. Miao, "Geographically focused collaborative crawling," in *Proceedings of the 15th International World Wide Web Conference*, 2006.
[64] GigaAlert, http://www.gigaalert.com.
[65] D. Gomes and M. J. Silva, "Characterizing a national community web," *ACM Transactions on Internet Technology*, vol. 5, no. 3, pp. 508–531, 2005.
[66] L. Gravano, H. García-Molina, and A. Tomasic, "The effectiveness of GlOSS for the text database discovery problem," in *Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data*, 1994.
[67] M. Gray, "Internet growth and statistics: Credits and background," http://www.mit.edu/people/mkgray/net/background.html.
[68] D. Gruhl, L. Chavet, D. Gibson, J. Meyer, P. Pattanayak, A. Tomkins, and J. Zien, "How to build a WebFountain: An architecture for very large-scale text analytics," *IBM Systems Journal*, vol. 43, no. 1, pp. 64–77, 2004.
[69] Z. Gyöngyi and H. García-Molina, "Web Spam Taxonomy," in *Proceedings of the 1st International Workshop on Adversarial Information Retrieval*, 2005.
[70] Y. Hafri and C. Djeraba, "High performance crawling system," in *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2004.
[71] M. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork, "Measuring index quality using random walks on the web," in *Proceedings of the 8th International World Wide Web Conference*, 1999.
[72] M. Henzinger, A. Heydon, M. Mitzenmacher, and M. Najork, "On near-uniform URL sampling," in *Proceedings of the 9th International World Wide Web Conference*, 2000.
[73] M. R. Henzinger, R. Motwani, and C. Silverstein, "Challenges in web search engines," *SIGIR Forum*, vol. 36, no. 2, pp. 11–22, 2002.
[74] M. Hersovici, M. Jacovi, Y. S. Maarek, D. Pelleg, M. Shtalhaim, and S. Ur, "The shark-search algorithm — An application: Tailored web site mapping," in *Proceedings of the 7th International World Wide Web Conference*, 1998.
[75] A. Heydon and M. Najork, "Mercator: A scalable, extensible web crawler," *World Wide Web*, vol. 2, no. 4, pp. 219–229, 1999.
[76] *International Workshop Series on Adversarial Information Retrieval on the Web*, 2005–.
[77] Internet Archive, http://archive.org/.
[78] Internet Archive, "Heritrix home page," http://crawler.archive.org/.

[79] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002.

[80] J. Johnson, K. Tsioutsiouliklis, and C. L. Giles, "Evolving strategies for focused web crawling," in *Proceedings of the 20th International Conference on Machine Learning*, 2003.

[81] R. Khare, D. Cutting, K. Sitakar, and A. Rifkin, "Nutch: A flexible and scalable open-source web search engine," Technical Report, CommerceNet Labs, 2004.

[82] J. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[83] M. Koster, "A standard for robot exclusion," http://www.robotstxt.org/orig.html, 1994.

[84] H.-T. Lee, D. Leonard, X. Wang, and D. Loguinov, "IRLbot: Scaling to 6 billion pages and beyond," in *Proceedings of the 17th International World Wide Web Conference*, 2008.

[85] L. Lim, M. Wang, S. Padmanabhan, J. S. Vitter, and R. C. Agarwal, "Characterizing web document change," in *Proceedings of the International Conference on Advances in Web-Age Information Management*, 2001.

[86] L. Liu, C. Pu, W. Tang, and W. Han, "CONQUER: A continual query system for update monitoring in the WWW," *International Journal of Computer Systems, Science and Engineering*, vol. 14, no. 2, 1999.

[87] B. T. Loo, O. Cooper, and S. Krishnamurthy, "Distributed web crawling over DHTs," UC Berkeley Technical Report CSD-04-1305, 2004.

[88] J. Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, and A. Halevy, "Google's deep-web crawl," in *Proceedings of the 34th International Conference on Very Large Data Bases*, 2008.

[89] M. Mauldin, "Lycos: Design choices in an internet search service," *IEEE Expert*, vol. 12, no. 1, pp. 8–11, 1997.

[90] O. A. McBryan, "GENVL and WWWW: Tools for taming the web," in *Proceedings of the 1st International World Wide Web Conference*, 1994.

[91] F. Menczer and R. K. Belew, "Adaptive retrieval agents: Internalizing local context and scaling up to the web," *Machine Learning*, vol. 39, pp. 203–242, 2000.

[92] F. Menczer, G. Pant, and P. Srinivasan, "Topical web crawlers: Evaluating adaptive algorithms," *ACM Transactions on Internet Technology*, vol. 4, no. 4, pp. 378–419, 2004.

[93] G. Mohr, M. Stack, I. Ranitovic, D. Avery, and M. Kimpton, "An introduction to Heritrix, an open source archival quality web crawler," in *Proceedings of the 4th International Web Archiving Workshop*, 2004.

[94] M. Najork and A. Heydon, "High-performance web crawling," Technical report, Compaq SRC Research Report 173, 2001.

[95] M. Najork and J. L. Wiener, "Breadth-first search crawling yields high-quality pages," in *Proceedings of the 10th International World Wide Web Conference*, 2001.

[96] A. Ntoulas, J. Cho, and C. Olston, "What's new on the web? The evolution of the web from a search engine perspective," in *Proceedings of the 13th International World Wide Web Conference*, 2004.

[97] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis," in *Proceedings of the 15th International World Wide Web Conference*, 2006.

[98] A. Ntoulas, P. Zerfos, and J. Cho, "Downloading textual hidden web content through keyword queries," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 2005.

[99] C. Olston and S. Pandey, "Recrawl scheduling based on information longevity," in *Proceedings of the 17th International World Wide Web Conference*, 2008.

[100] V. J. Padliya and L. Liu, "Peercrawl: A decentralized peer-to-peer architecture for crawling the world wide web," Georgia Institute of Technology Technical Report, 2006.

[101] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," Technical Report, Stanford University, 1998.

[102] S. Pandey, K. Dhamdhere, and C. Olston, "WIC: A general-purpose algorithm for monitoring web information sources," in *Proceedings of the 30th International Conference on Very Large Data Bases*, 2004.

[103] S. Pandey and C. Olston, "User-centric web crawling," in *Proceedings of the 14th International World Wide Web Conference*, 2005.

[104] S. Pandey and C. Olston, "Crawl ordering by search impact," in *Proceedings of the 1st International Conference on Web Search and Data Mining*, 2008.

[105] S. Pandey, K. Ramamritham, and S. Chakrabarti, "Monitoring the dynamic web to respond to continuous queries," in *Proceedings of the 12th International World Wide Web Conference*, 2003.

[106] G. Pant and P. Srinivasan, "Learning to crawl: Comparing classification schemes," *ACM Transactions on Information Systems*, vol. 23, no. 4, pp. 430–462, 2005.

[107] G. Pant and P. Srinivasan, "Link contexts in classifier-guided topical crawlers," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 107–122, 2006.

[108] B. Pinkerton, "Finding what people want: Experiences with the WebCrawler," in *Proceedings of the 2nd International World Wide Web Conference*, 1994.

[109] S. Raghavan and H. García-Molina, "Crawling the hidden web," in *Proceedings of the 27th International Conference on Very Large Data Bases*, 2001.

[110] U. Schonfeld and N. Shivakumar, "Sitemaps: Above and beyond the crawl of duty," in *Proceedings of the 18th International World Wide Web Conference*, 2009.

[111] V. Shkapenyuk and T. Suel, "Design and implementation of a high-performance distributed web crawler," in *Proceedings of the 18th International Conference on Data Engineering*, 2002.

[112] A. Singh, M. Srivatsa, L. Liu, and T. Miller, "Apoidea: A decentralized peer-to-peer architecture for crawling the world wide web," in *SIGIR Workshop on Distributed Information Retrieval*, 2003.

[113] Q. Tan, Z. Zhuang, P. Mitra, and C. L. Giles, "A clustering-based sampling approach for refreshing search engine's database," in *Proceedings of the 10th International Workshop on the Web and Databases*, 2007.

[114] T. Urvoy, T. Lavergne, and P. Filoche, "Tracking web spam with hidden style similarity," in *Proceedings of the 2nd International Workshop on Adversarial Information Retrieval on the Web*, 2006.

[115] J. L. Wolf, M. S. Squillante, P. S. Yu, J. Sethuraman, and L. Ozsen, "Optimal crawling strategies for web search engines," in *Proceedings of the 11th International World Wide Web Conference*, 2002.

[116] B. Wu and B. Davison, "Identifying link farm spam pages," in *Proceedings of the 14th International World Wide Web Conference*, 2005.

[117] P. Wu, J.-R. Wen, H. Liu, and W.-Y. Ma, "Query selection techniques for efficient crawling of structured web sources," in *Proceedings of the 22nd International Conference on Data Engineering*, 2006.

[118] Yahoo! Research Barcelona, "Datasets for web spam detection," http://www.yr-bcn.es/webspam/datasets.

[119] J.-M. Yang, R. Cai, C. Wang, H. Huang, L. Zhang, and W.-Y. Ma, "Incorporating site-level knowledge for incremental crawling of web forums: A list-wise strategy," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009.

[120] S. Zheng, P. Dmitriev, and C. L. Giles, "Graph-based seed selection for web-scale crawlers," in *Proceedings of the 18th Conference on Information and Knowledge Management*, 2009.

[121] K. Zhu, Z. Xu, X. Wang, and Y. Zhao, "A full distributed web crawler based on structured network," in *Asia Information Retrieval Symposium*, 2008.