Identification of Requirements for Focused Crawlers in Technology Intelligence

Günther Schuh, André Bräkling, Katharina Apfel
Fraunhofer Institute for Production Technology IPT, Aachen, Germany

Abstract--**The fast and high availability of knowledge is at first seen as a benefit for knowledge workers in the information age. On closer examination the outcome of this is a big challenge: The amount of data that is available these days has to be reasonably structured and conditioned. Only the US Library of Congress collected 235 terabyte of data on its own by April 2011. Technology intelligence as a fundamental component of technology management is expected to monitor these data, so technology managers are able to respond to new developments and trends just in time. Possible tools to meet this challenge in an efficient way are the focused crawlers. These are programs, which explore data collections independently to identify material related to the current working context.**

To implement such a tool, there exist a multitude of different approaches within the field of information retrieval, but they have to be used and combined on an individual basis to fit the requirements of a particular task. Hence, before a focused crawler can make the processes of technology intelligence more efficient, the dedicated requirements have to be identified. In this paper we develop a requirements model to close this gap.

I. INTRODUCTION

Technology intelligence is a fundamental element of technology management and hence an important part of a company's business intelligence. Its purpose is the systematic identification of technological prospective chances but also threats to a company [1]. To this end it uses three basic activities called technology scanning, monitoring and scouting.

Technology scanning is a constant and undirected search process, which gives an overview about unknown or new technology-related information. Based on a scanning's results, a technology monitoring can be done, which examines particular or long-term relevant technology fields over a longer period of time. However a scouting is an event driven and detailed fast supply with information and information sources which are related to a specific technology topic. These scoutings could be used to perform a diversification of applications and markets. As shown in Fig. 1, the technology intelligence process consists of the same four steps independent of the inquired activity [1, 2]:

1. *Determination of information needs:* First of all, the field of observation and the required level of detail has to be identified and defined. The decision about the actual information needs establishes a basic orientation guide for the next steps.
2. *Information search:* After the information needs are defined, a decision about the information sources has to be taken. Afterwards a sufficient amount of information has to be discovered using various sources of information, e.g., literature, databases, websites as well as networks inside and outside of the company. It is not only important to get an overview about activities outside the company, but also to keep an eye on the internal developments to identify possible synergies. Also a well-balanced examination of the current topics, e.g., by watching trends of the market, and future topics, e.g., by analyzing research results and patents, is necessary for reliable forecasts.
3. *Information assessment:* The assessment of the discovered information is divided into three parts: The selection to reduce the results of step 2 according to priority and relevance, the analysis regarding the current intention and the prediction concerning prospective developments. The information is not expediently usable until the assessment is done whereby the comprehensive search results are structured in a manageable way.
4. *Communication of information:* Concluding the findings of step 3 have to be prepared and communicated to the concerned departments. This will allow, e.g., a company's management to make important decisions on the future orientation and investments in technology topics.

Regarding these steps, the process always comprises information search and processing. So the fast and high availability of information is at first a benefit for technology intelligence, but on closer examination the outcome of this is also a big challenge: The amount of data that is available these days has to be reasonably structured and conditioned. Only the US Library of Congress collected 235 terabyte of data on its own by April 2011 [3]. Obviously this amount of data, which could contain the Complete Works of Shakespeare provided by Project Gutenberg almost 50 million times, is more than a human can handle without computerized support.

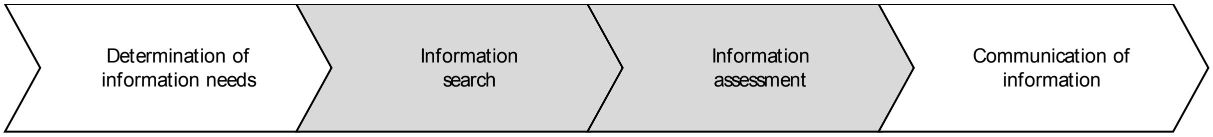
Determination of information needs
Information search
Information assessment
Communication of information

Fig. 1. Typical technology intelligence process.

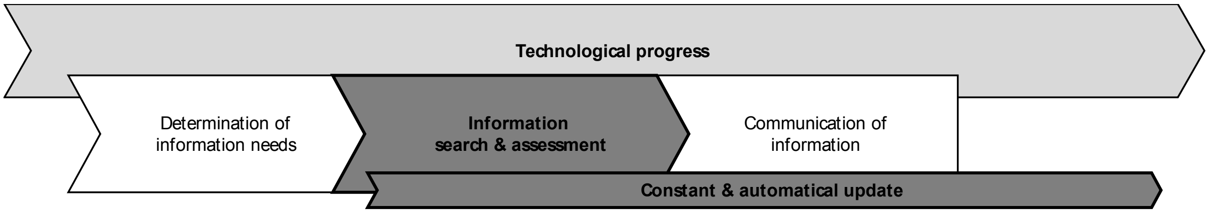
Technological progress
Determination of information needs
Information search & assessment
Communication of information
Constant & automatical update

Fig. 2. Technology intelligence process improved by information retrieval technologies.

Because of a growing diversification and the increasing *cross selection character* of technologies, technology intelligence is an important guidepost in a company's strategic management, whereas the available resources for technology intelligence processes are still limited. Therefore prospective technology intelligence processes have to meet the challenges of information explosion by utilizing information retrieval approaches given by the computer science [4, 5]. In addition, the process itself can be reformed to a continuous observation process as shown in Fig. 2, constantly and automatically done by the contemplated computerized solution.

A promising approach is the focused crawling which was described in [6] as a system which *seeks, acquires, indexes, and maintains pages on a specific set of topics that represent a relatively narrow segment of the web*. The fish-search algorithm which was presented 1994 in [7] can be considered as a form of a pioneer algorithm that behaves like a focused crawler: Starting at a defined web page and configured with relevance criteria (e.g. keywords that have to be found on a page), it creates a list of all URLs found at this page. Afterwards, it checks each page listed on the frontier (i.e. the list of URLs referencing unexplored pages), calculates its relevance index based on the relevance criteria and decides depending on this index how many URLs at the current page where added to the URL list this time, see Fig. 3. Following the metaphor, a school of fish that does not find food (= relevant pages) will starve, a school of fish finding sufficient food keeps growing.

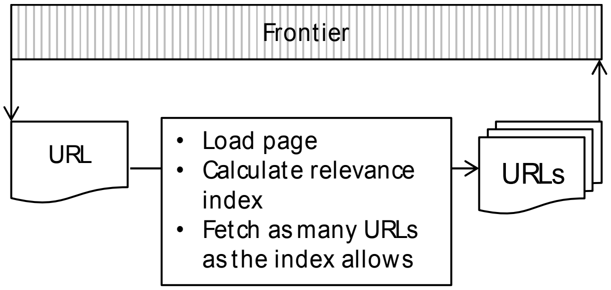
Frontier
URL
• Load page
• Calculate relevance index
• Fetch as many URLs as the index allows
URLs

Fig. 3. A simple focused crawling algorithm like fish-search.

Based upon this algorithm a lot of new approaches and also implementations were developed during the last 20 years [8]. A general high level structure of a focused (or topical) crawler is sketched in Fig. 4 and contains four layers as described in [9]:

1. *Networking:* The networking layer handles the technical process to request, download and store the crawled pages. It also decides whether a cached page can be reused or has to be downloaded again to consider updated contents. The results will be delivered to the layers above.
2. *Parsing and Extraction:* This layer parses the page delivered by the networking layer and extracts all necessary data, e.g., URLs plus their context.
3. *Representation:* The representation layer converts the extracted heterogeneous data into a standardized formal representation that serves homogeneous raw data for the following computerized utilization.
4. *Intelligence:* The final layer assigns a relevance index or a similar score to the extracted URLs and adds them to the frontier for subsequent processing.

Furthermore, the figure shows the frontier, a history and page repository to keep track of earlier results and the inputs which defines the initial position. Consequently, the intelligence layer depends not only on the results of the preceding layers but also on the inputs which are called important or even the most crucial aspects for the configuration of the crawler [10, 11]. More complex architectures are briefly described in [11, 12, 13, 14].

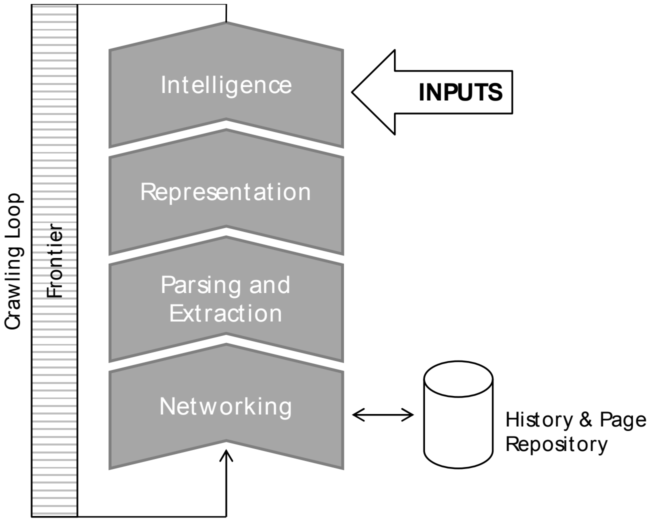

Fig. 4. A focused crawling infrastructure according to [9].

The basic principle of focused crawling seems to face the requirements of information search and assessment extended by an architecture that considers the needs of continuous monitoring in technology intelligence processes in companies. Hence it is examined as a reasonable tool for competitive and accordingly business intelligence [15, 16] and should be worth to be tested in technology intelligence as well.

II. AIM OF THE PAPER

As indicated in the paper's introduction, a focused crawler consists of different components with different characteristics. While it is clear *what* the layers of a crawler do, it is important to decide *how* they perform their tasks. The width of the differences in effectiveness and efficiency of varying crawler implementations is considerably shown in [9, 11]. To score the effectiveness and efficiency of crawlers regarding to application fields in technology intelligence, initially the dedicated requirements have to be identified.

Therefore this paper will first discuss the activities of information search and assessment. Second it will identify the links between these activities and a general focused crawling workflow. By examining existing focused crawler implementations or approaches and by comparing the results of earlier research related to these crawlers' pros and cons, a brief overview of basic crawler properties and their influence on the results will be given. Weighted by the needs of technology intelligence in companies an initial requirements model for focused crawlers in technology intelligence will be introduced as the aim of the paper.

This requirements model should be a basis for further research and experiments regarding focused crawling in technology intelligence processes. A focused crawler will enable companies to analyze a larger amount of information in less time and for lower effort, to set up continuous monitoring processes and to get in-time notifications about important events. Supplemental the focused approach also spares IT capacities, because only relevant information will be stored and processed. Knowledge workers like technology intelligence experts in companies will be able to spend their limited time on challenging interpretation and prediction tasks instead of interminable inquiry tasks.

III. ACTIVITY OF INFORMATION SEARCH AND ASSESSMENT

To keep the information assessment manageable, the information search has to evaluate incoming data. The information needs were determined already in the first step, so the following parameters remain to adjust the information search at the beginning according to [1]:

1. Who is responsible for the actual technology intelligence process? Depending on the executing department and its business segment the *value* of different information may vary.
2. Which sources of information (i.e. webpages of suppliers and competitors, patent databases, research results, internal databases etc.) should be screened? On the one hand a sufficient amount of information sources must be accessible; on the other hand not every source may be suitable for the current technology intelligence search process. The *cost* of the process will depend on this decision.
3. How detailed should the search be done? The result of information search cannot be benchmarked by the cost/value ratio only. In addition the *coverage* (as the ratio of identified relevant information to the assumed amount of total relevant information) of the search has to be observed.

By determining these parameters not only knowledge workers will be guided through the process but also the fundamental requirements will be defined: limit cost, measure value and optimize coverage. Even if the information search is prepared and processed well, a huge amount of results is still expectable. For this reason a following information assessment is necessary. In [1] three assessment stages are mentioned:

1. *Selection:* Due to a rating of relevance and priority the amount of information will be reduced.
2. *Analysis:* The selected information will be preprocessed regarding the current scope of work.
3. *Forecast:* Evaluation of the rated and preprocessed information to predict future developments.

With increasing stage the procedure of assessment becomes more complex. While the selection and basic analytics can be done by statistical methods, advanced analysis and especially forecasting require explicit knowledge on the explored technology topic – which causes a higher demand for manpower. So information search and assessment engage each other in reducing the effort which is necessary to fulfill the defined information needs and to get manageable results for the closing communication.

IV. CONSOLIDATION WITH FOCUSED CRAWLING

By comparing the description of the information search and assessment activities with focused crawler architecture described in the introduction, the similarities can easily be distinguished. In fact the activities merely describe a manual execution of the process which is also executed by the focused crawlers.

The input for the crawler is the result of the determination of information needs and of the preconsiderations regarding the information search. Also, the *cost* which may occur and the *coverage* that is required have to be defined for the crawling process. In addition, it has to be defined how the information *value* should be measured, so the algorithm can be provided with criteria for relevance rating. Finally the information sources and with it the seed pages have to be determined as the crawling's starting point, i.e., quasi to represent the first frontier entries.

If the input is available, the actual crawling process can start. The networking layer initially runs the source requests, whereas parsing and extraction already does a first preprocessing and preselection: URLs are picked out and their contexts prepared for rating. The representation layer reduces the extracted information to the necessary amount, so the intelligence layer can execute the analysis. On the basis of this analysis, the same layer can proceed with the relevance rating and the selection of relevant information. The consolidated process is shown in Fig. 5.

The benefits of automatic execution are the lower additional expenses to update earlier analyzed websites considering recent findings respectively to execute the process continuously even if the primary needs are met.

V. INTRODUCING AN INITIAL REQUIREMENTS MODEL

Although the consolidation of the technology intelligence processes and the focused crawling seems to be obvious at the first sight, there still remain diverse open issues. The single layers of the focused crawler work on a comprehensive task which can be accomplished in different ways. The initial requirements model that is developed in this paper should describe the decisions that have to be made to configure a crawler, and how these have an impact on the expectable results:

1. *Seed pages:* The selection of seed pages aligns the whole crawling processes. A broad collection of pages may spread the result set far but also cause an ambiguous finding. In contrast a very specialized collection may limit the whole process and hide a lot interesting results.
2. *Value estimation:* The value estimation depends on the information needs and whose needs they are. Automated value estimation extends the focused crawler topic by the field of natural language processing (NLP) as discussed in [17], otherwise a manual evaluation by the human knowledge worker is still required. Considering the complexity and still existing development need regarding NLP, the latter actually seems to be the indispensable but cost-expensive solution.
3. *Cost limit:* This requirement combines two types of cost, the access fee for sources as well as time and resources to execute the underlying algorithms. By limiting the amount of cost, the coverage and the amount of high valuable results is accordingly constrained.
4. *Coverage:* By defining a low coverage as sufficient, less effort has to be put in the process. However only a high coverage can ensure a high discovery of valuable results.
5. *Relevance algorithm:* While the value estimation handles the current analyzed website's content and evaluates based on the information needs, the relevance algorithm has to judge about the relevance of a site and especially their links to other sites regarding the examined topic. E.g., a topic-related site may not contain the required information (i.e., its value is low), but it can lead the crawler to high-value sites. An often used weight is "Term-Frequency – Inverse Document Frequency" (TF-IDF) as explained in [10, 18]. It is a mathematical statistic which indicates the importance of a keyword to a page in the whole result set.
6. *URL and context extraction algorithm:* The relevance algorithm calculates an index depending on the analyzed site's contents. For that purpose the right part of the site has to be chosen to judge a specific link. E.g. it does not make sense to rate a link low just because of a non-relevant advertisement at the same page. The approaches range from link anchor examination as evaluated in [19] to more complex partitioning algorithms like "Content-Block-Partition – Selective Link Context" (CBP-SLC) as presented in [20].

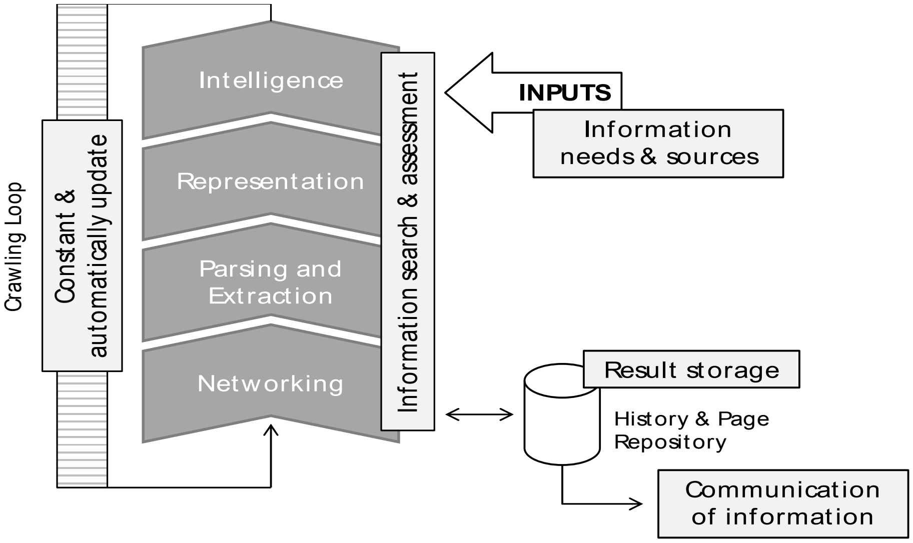
Crawling Loop
Constant & automatically update
Intelligence
Representation
Parsing and Extraction
Networking
Information search & assessment
INPUTS
Information needs & sources
Result storage
History & Page Repository
Communication of information

Fig. 5. The mentioned focused crawler architecture consolidated with the technology intelligence process.

7. *Repository processing:* Depending on the time pages in the history and page repository are stored, the crawler has to revisit already processed pages. The revisiting will raise cost, but also ensures to keep track of late-breaking news. A method for computing a revising reasonable frequency is described in [21].

If the decisions regarding all of these requirements are made, the crawler can start the actual search process. The results of the crawling pass can be evaluated by *precision* and *recall* as mentioned in [9, 10, 11]: The precision describes the number of positive relevant results at the ratio of all received results, while recall corresponds to the fraction of positive relevant results to all relevant results existing in the examination field.

Therefore the compliance of the described requirements should aim to optimize the precision and recall values. In summary the requirements and the valuation parameters can be represented as a triangle of *value*, *cost* and *coverage*, as Fig. 6 shows.

VI. CONCLUSION AND OUTLOOK

Focused crawlers are an interesting tool for the observation tasks of technology intelligence. In fact even a basic crawling architecture is a complex entity with a lot of adjustable screws. Initially the usage of the preliminary studies seems to be a good starting point: On the basis of the framework defined in [11], different algorithms were already compared with each other and at the same time the foundations are laid to continue with further tests.

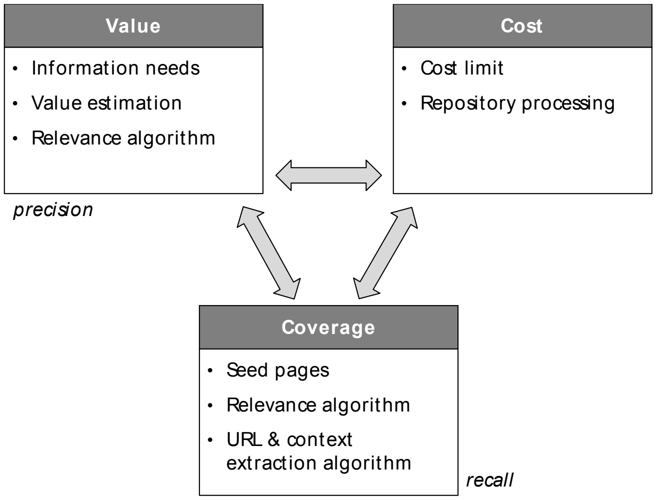

Fig. 6. The different requirements for focused crawlers categorized in the interactive triangle of value, cost and coverage.

The initial requirements model developed in this paper actually contains only the very basic requirements in the triangle of *value*, *cost* and *coverage*. But thereby it provides the basis to design different focused crawlers for the different tasks of the technology intelligence. The next step has to be an evaluation of explicit algorithms and configurations on which the requirements model can be improved, so that a general configuration model for focused crawlers in technology intelligence can be developed.

Furthermore, the value estimation using NLP has to be an important research field to reduce the required effort to get assured useful results.

REFERENCES

[1] Wellensiek, M., G. Schuh, P. A. Hacker, J. Saxler: Technologiefrüherkennung. In: *Technologiemanagement*, G. Schuh, S. Klappert, Eds. Springer Berlin, Heidelberg: 2011.

[2] Spath, D., S. Schimpf, C. Lang-Koetz: Technologiemonitoring. Technologien identifizieren, beobachten und bewerten. Fraunhofer IAO, Stuttgart, 2010.

[3] Manyika, J., M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, A. Hung Byers: Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, 2011.

[4] Schuh, G., T. Drescher, K. Apfel: Wissensexplosion im technologischen Umfeld beherrschen und benutzen. In: *10. Aachener Management Tage. Navigation für Führungskräfte*, G. Schuh, A. Kampker, V. Stich (Eds.), Apprimus Aachen, 2013.

[5] Manning, C. D., R. Prabhakar, H. Schütze: *Introduction to information retrieval*. Cambridge University Press, Cambridge, 2008.

[6] Chakrabarti, S., M. van den Berg, B. Dom: Focused crawling: a new approach to topic-specific Web resource discovery. In: *Computer Networks* 31(11), pp. 1623-1640. Elsevier, 1999.

[7] De Bra, P. M. E., R. D. J. Post: Information Retrieval in the World Wide Web: Making client-based searching feasible. In: *Proceedings of the 1st International World Wide Web Conference*. CERN Geneva, 1994.

[8] Kwiatkowski, M., S. Höhfeld: Thematisches Aufspüren von Web-Dokumenten – Eine kritische Betrachtung von Focused Crawling-Strategien. In: *Information – Wissenschaft & Praxis* 58(2), pp. 69-82. DGI e.V. Frankfurt, 2007.

[9] Pant, G., P. Srinivasan: Learning to Crawl: Comparing Classification Schemes. In: *ACM Transactions on Information Systems* 23(4), pp. 430-462. ACM Pittsburgh, 2005.

[10] Rawat, S., D. R. Patil: Efficient Focused Crawling based on Best First Search. In: *Proceedings of the 3rd IEEE International Advance Computing Conference (IACC)*, pp. 908-911. IEEE, 2013.

[11] Srinivasan, P., F. Menczer, G. Pant: A General Evaluation Framework for Topical Crawlers. In: *Information Retrieval* 8(3), pp. 417-447. Kluwer Academic Publishers Dordrecht, 2004.

[12] Singh, S., N. Tyjagi: A Novel Architecture of Mercator: A Scalable, Extensible Web Crawler with Focused Web Crawler. In: *International Journal of Computer Science and Mobile Computing (IJCSMC)* 2(6), pp. 244-250. IJCSMC, 2013.

[13] Zhuang, Z., R. Wagle, C. Lee Giles: What's There and What's Not? Focused Crawling for Missing Documents in Digital Libraries. Digital Libraries, 2005. *Proceedings of Joint Conference of Digital Library (JCDL '05)*, pp. 301-310. IEEE, 2005.

[14] Pal, A., D. S. Tomar, S. C. Shrivastava: Effective Focused Crawling Based on Content and Link Structure Analysis. In: International Journal of Computer Science and Information Security (IJCSIS) 2(1). IROCS, 2009.

[15] Chen, H., M. Chau, D. Zeng: CI Spider: a tool for competitive intelligence on the Web. In: *Decision Support Systems* 34(1), pp. 1-17. Elsevier, 2002.

[16] Pant, G., F. Menczer: Topical Crawling for Business Intelligence. In: *Research and Advanced Technology for Digital Libraries.* Springer Berlin, Heidelberg, 2003.

[17] Bellot, P., L. Bonnefoy, V. Bouvier, F. Duvert, Y.-M. Kim: Large Scale Text Mining Approaches for Information Retrieval and Extraction. In: *Innovations in Intelligent Machines-4. Studies in Computational Intelligence* 514, pp. 3-45. Springer International, 2014.

[18] Kumar, M., R. Vig: Focused Crawling Based Upon Tf-Idf Semantics and Hub Score Learning. In: Journal of Emerging Technologies in Web Intelligence (JETWI) 5(1), pp. 70-77. Academy Publisher, 2013.

[19] Craswell, N., D. Hawking, S. Robertson: Effective Site Finding using Link Anchor Information. In: *Proceedings of SIGIR '01*, pp. 250-257. ACM, 2001.

[20] Peng, T., L. Liu: Focused crawling enhanced by CBP-SLC. In: *Knowledge-Based Systems* 51, pp. 15-26. Elsevier, 2013.

[21] Dixit, A., A. K. Sharma: A Mathematical Model for Crawler Revisit Frequency. In: *2nd International Advance Computing Conference (IACC)*, pp. 316-319. IEEE, 2010.