iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** IV **Month of publication:** April 2017

DOI: http://doi.org/10.22214/ijraset.2017.4140

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Brief Review of Max-Min Ant System based Focussed Crawler

Komal Upadhyay[1], Er. Suveg Moudgil(Associate Professor)[2]
[1]Department of Computer Science (M.TECH 4th sem), [2]Department of Computer Science
Haryana Engineering College Jagadhri, Kurushetra University, Haryana,India

Abstract: ***A focused crawler is Web crawler that traverses the Web to explore information that is related to a particular topic of interest only. This study, aims to find the Indian academicians webpages from foreign universities websites by selecting the features of webpage and determine its relevance on an unknown dataset. Therefore, a feature selection algorithm based on Max-Min Ant System (MMAS) is presented to improve the accuracy of focused crawler and classification process. The weights to features are assigned using cosine similarity to determine the relevancy of webpages. MMAS finds the best solution and select best fitted URLs from a large pool of URLs. The performance of the proposed methodology classification result is compared with manual classification result for Lancaster University, Stanford University and Harvard University dataset. The performance of the proposed methodology is measured using recall parameter.***
Keywords: Max Min Ant System (MMAS), Ant System, Term Frequency-Inverted Document Frequency(TF-IDF).

I. INTRODUCTION(INTRODUCTION TO MMAS)

The World Wide Web is a very big warehouse that contains and connects millions of webpages and Web resources. These interlinked resources using hypertext links, are identified using the Uniform Resource Locator (URL) and are accessed via the Internet. This information system is spread on servers all over the world in the form of webpages. It is not possible to handle all these webpages by humans. Search engines are software systems, designed to find all these webpages in a fast way. Web search engines get their information through the Web, crawl from site to site by following the hyperlinks. Web crawlers are the drivers of search engines, which update the database whenever any webpage gets added or removed from the Web. Web crawlers also known as Web robots, Web spiders, and Web ants are software that browse the Web and download the webpages in an automated manner (Manning, Raghavan, & Schütze, 2008). Web crawler algorithm is quite simple as it starts with a prepared list of links also known as seed URLs, adds them in a queue and traverses each of these webpages. It identifies new links on that particular webpage and further adds them to the queue. It traverses all the links found according to a set of policies recursively. In short, a Web crawler downloads the webpages, creates their index of the webpages and maintains them in a repository or database (Micarelli & Gasparetti, 2007).

Types of Web Crawlers

A. General Purpose Web Crawler

General purpose crawlers are also known as *universal crawlers* download all the webpages without regard to any specific subject or topic. The goal is to cover as much Web as possible within given time. These are large scale crawlers and incur high cost in terms of network bandwidth usage, but this cost is amortised over many number of queries by users (Manning, Raghavan, & Schütze, 2008). Also, the repository needs to be updated more often in general purpose crawlers.

B. Incremental Crawler

Incremental crawler crawl the Web continuously and revisit webpages periodically. During its continuous crawl, it may also oust some webpages that are stored locally to make space for newly crawled webpages. The incremental crawler has two goals: to keep the fresh collection stored locally and to improve quality of the collection. Depending upon the alteration frequency of webpages, it visits the webpages with high alteration rate more frequently and the other webpages less frequently. The advantage of incremental crawler is that it saves network bandwidth, since only webpages with high change rate are downloaded instead of all the webpages being downloaded.

C. Parallel Crawler

Volume 5 Issue IV, April 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Parallel crawlers are multiple crawlers that run in parallel. Parallel crawler consists of multiple crawling processes each of which performs the basic functionality of a single general purpose crawler. The parallel crawler can be geographically distributed or can be on a local network. All crawler installed on different machines have access to common memory. The webpages are downloaded from the Web and are stored locally. The URLs are extracted and their links are traversed. The goal is to reduce the total crawling time significantly (Cho & Garcia-Molina, 32002). Thus, bulk of webpages can be downloaded in reasonable amount of time. Parallel Crawler has many advantages like network-load dispersion, network-load reduction.

D. Distributed Crawler

Distributed crawlers are multiple instances of a crawler that download pages from the Web and are executed at geographically distant locations. It based upon the idea of distributed computing. In order to achieve wide coverage of the Web, many crawlers are geographically distributed on the Internet and a central server is used for management of communication and synchronization of the distributed nodes. Each crawler does the crawling of the part of the Web assigned to it in a distributed manner. Its advantage is that it is robust against the system crash and balance network load (Boldi, Codenotti, Santini, & Vigna, 2004).

E. Focused Crawler

Focused crawler downloads webpages on specific topic or subject. The goal of focused crawler is to retrieve the maximal set of relevant webpages while concurrently traversing the minimal number of irrelevant webpages on the Web. A Focused Web crawler learns to identify webpages that are relevant to the interest of user (Zhang, Du, & Li, 2009). User defines the seed URLs and query. Seed URLs are the starting point of the focused crawler and such URLs are related to query. User defines query to search in seed URLs and other originating links from such URLs. The fetch module fetches the webpage from the corresponding URL using HTTPS/HTTP protocol. The parse module performs scanning of a webpage to identify different HTML tags of a webpage. The page relevancy is calculated using relevance calculation module so that only those URLs that are related to query are stored in crawler queue and rest of the URLs are ignored. The webpage repository is a database where are the parsed URLs are stored for future reference.

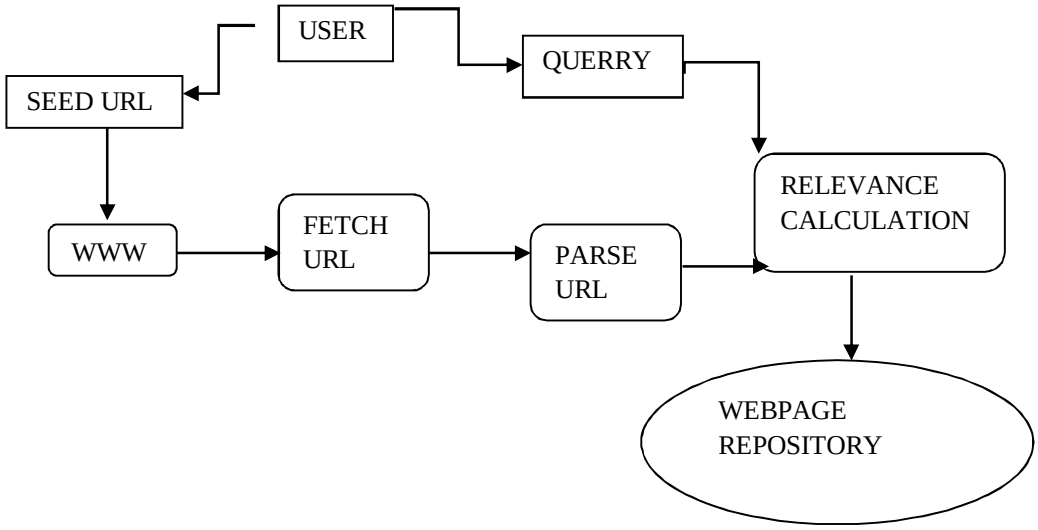


Figure 1 shows the working of basic focused crawler.

II. LITERATURE REVIEW

A. Evaluating Topic-Driven Web Crawlers (Menczer, Pant, Srinivasan, & Ruiz, 2001)

Discussed best first crawler that searches the pages ranked according to the similarity score. The similarity score is the cosine similarity between the topic terms and web page. Best first crawlers use only term frequencies vectors for calculating page's relevance. Due to its simplicity and efficiency, best first crawling is considered to be the most successful approach for web crawling.

B. Adaptive Focused Crawling (Micarelli & Gasparetti, 2007

Discussed an overview of the Focused crawling domain and discussed approaches that include some adaptive techniques. These technique makes focused crawler possible to change the system behaviour according to the particular environment and its associations with the given input parameters during the search.

C. Text Feature Selection using Ant Colony Optimization (Aghdam, Ghasem-Aghaee, & Basiri, 2009

Discussed novel feature selection algorithm that is based on ant colony optimization. First of all documents are pre-processed, then feature extraction transform the input text document into a feature set. To reduce the dimensionality of feature set feature selection is applied to the feature set. Ant Colony Optimization is used to discover the domain of all subsets of given calculated feature set. The finest feature subset found is then produced as the recommended set of features to be used in the actual design of the classification system

D. A Web Page Classification System Based on a Genetic Algorithm Using Tagged-Terms as Features (Özel, 2011

Discussed a webpage classification system using genetic algorithm (GA). They have taken both HTML tags and terms as feature set. Different weights are assigned to HTML tags and terms. The selected best features improve the run time performance of the classifier. Webpages are classified based on the selected top features obtained using genetic algorithm

E. An Ant Colony Optimization based Feature Selection for Webpage Classification (Saraç & Özel, 2014

Discussed an ant colony optimization (ACO) algorithm to select the best features. C4.5, naive Bayes, and *k* nearest neighbour classifiers were applied to assign class labels to Webpages. Their system consists of feature extraction, ACO based feature selection, and classification components. The training dataset is prepared according to binary class classification problem. From the dataset, features are then taken out, after then the finest subset of features is being selected by the ACO algorithm and these selected subsets are used to classify Webpages with Weka data mining software.

Table 1 SURVEY OF OPEN SOURCE WEB CRAWLERS

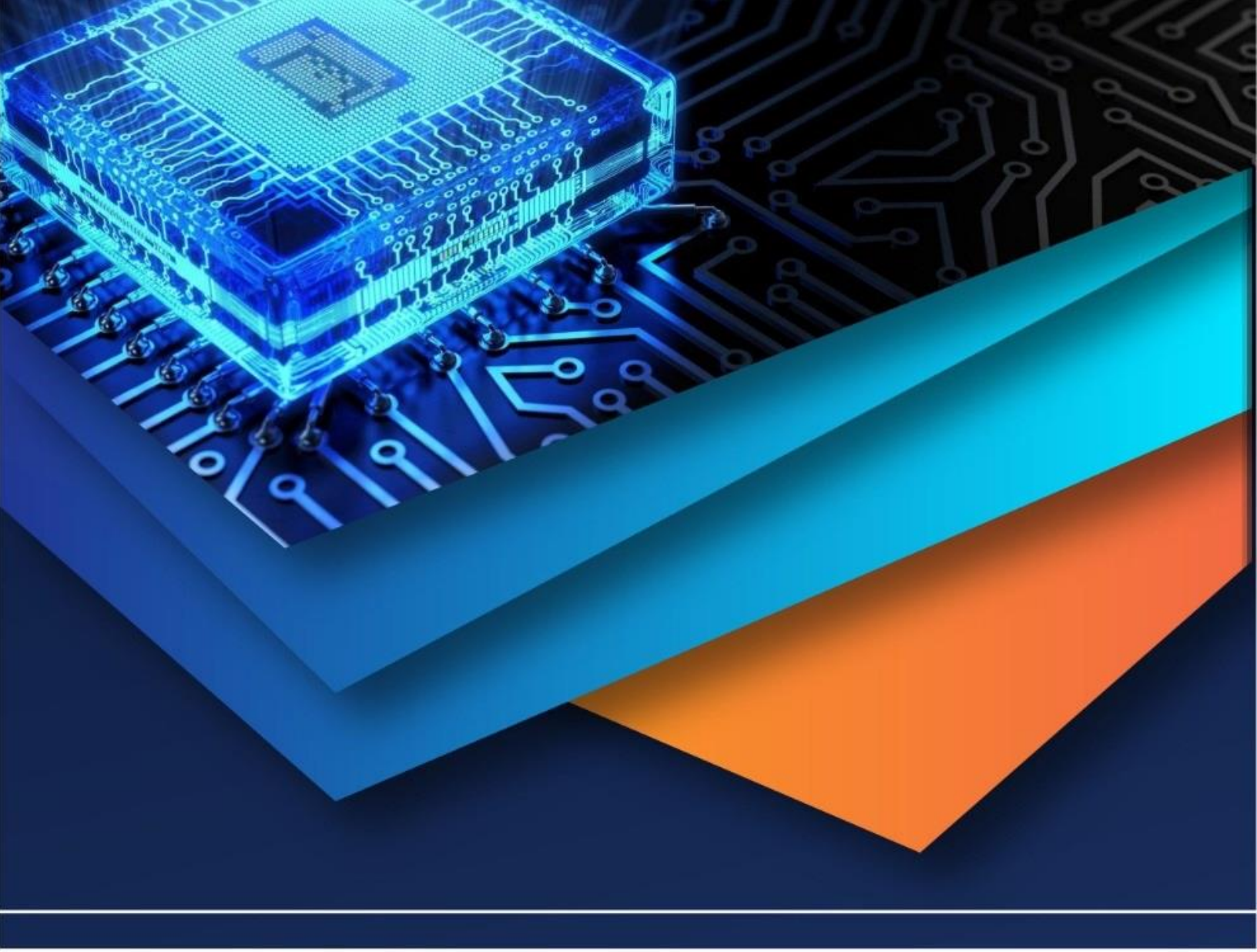
Features	Nutch	Scrapy	Heritrix	Norconnex http collector	Crawler4j	YaCy	Webs Phinx	jSpider	Xapian	Ebot
Language	Java	Python	Java	Java	Java	Java	Java	Java	C++	Erlang
Flexible	Yes	Yes	Yes	Yes				Yes		
Scalable	Yes							Yes	Very Well	Yes
Multi Threaded	Yes		Yes	Yes	Yes		Yes		Does not provide explicitly but can be used	
Distributed	Yes			Yes		Yes		Yes		Yes
Cross Platform	Yes	Yes	Yes	Yes		Yes	Yes		Highly Portable	Linux
Documentation	Bad	Developer	User	Good					Yes	No
Configurable			Yes		Yes		Yes			Yes
Focussed	Yes	Yes	Yes	Universal						
Extensible	Yes	Yes	Yes	Yes	Yes			Yes		
Continuous Crawl			Yes							
Interface	Command line	Command line	Both			GUI	GUI		Command line	

III. CONCLUSION

Focused Crawler aims to retrieve the maximum relevant webpages and minimize the irrelevant webpages concurrently. MAX-MIN ant system based focused crawler for webpage classification is proposed. The proposed method aims to find all Indian academicians webpages from foreign universities websites. Lancaster University and Stanford University has been crawled and screened relevant and moderately relevant URLs to avoid the problem of tunnelling. MMAS finds best fitted URLs thereby minimizing the number of URLs from large pool of moderately relevant URLs. The computed results are compared with manual classification result. This study concludes that due to no uniformity among websites, sometimes it is difficult for focused crawler to extract all the relevant information. Therefore, websites structure should be analysed and parameters should be adjusted accordingly in order to get high accuracy. However, problem of tunnelling is removed and recall obtained for the Lancaster University, Stanford University and Harvard University are 0.86, 0.87 and 0.83 respectively. For future work, the proposed model is to be implemented for other university websites. We also intend to improve recall parameter for the proposed methodology and reduce the crawling time and space.

REFRENCES

[1] Aggarwal, C., Al-Garawi, F., & Yu, P. (2001). Intelligent crawling on the World Wide Web with arbitrary predicates. Proceedings of the 10th international conference on World Wide Web (pp. 96-105). ACM.
[2] Aghdam, M. H., Ghasem-Aghaee, N., & Basiri, M. E. (2009). Text feature selection using ant colony optimization. Expert systems with applications, 36(3) , 6843-6853.
[3] Boldi, P., Codenotti, B., Santini, M., & Vigna, S. (2004). UbiCrawler: A scalable fully distributed Web Crawler. Software: Practice and Experience 34.8, 711-726.
[4] Brin, S., & Page, L. (2012). Reprint of: The anatomy of a large-scale hypertextual web search engine. Computer networks 56, no. 18, 3825-3833.
[5] Chakrabarti, S., Berg, M. V., & Dom, B. (1999). Focused crawling: a new approach to topic-specific Web resource discovery. Computer Networks, (pp. 1623-1640).
[6] Chakrabarti, S., Dom, B. E., & Berg, M. v. (9 Jul. 2002). System and method for focussed web crawling. U.S. Patent No. 6, 418-433.
[7] Cho, J., & Garcia-Molina, H. (2002). Parallel Crawlers. Proceedings of the 11th international conference on World Wide Web. ACM.
[8] De Bra, P., Houben, G.-J., Kornatzky, Y., & Post., R. (1994). Information retrieval in distributed hypertexts. Intelligent Multimedia Information Retrieval Systems and Management-Volume 1 (pp. 481-491). LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE.
[9] Dorigo, M., & Gambardella, L. M. (1997). Ant colony system: a cooperative learning approach to the travelling salesman problem. IEEE Transactions on evolutionary computation 1, no. 1, 53-66.
[10] Dorigo, M., Maniezzo, V., & Colorni, A. (1996). Ant system: optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 26, no. 1, 29-41.
[11] Edwards, J., McCurley, K., & Tomlin, J. (2001). An adaptive model for optimizing performance of an incremental web crawler. Proceedings of the 10th international conference on World Wide Web (pp. 106-113). ACM.
[12] Gasparetti, F., & Micarelli, A. (2003). Adaptive Web Search Based on a Colony of Cooperative Distributive Agents. Cooperative Information Agents, Springer , 168-183.
[13] Heydon, A., & Najork, M. (1999). Mercator: A scalable, extensible web crawler. World Wide Web 2, no. 4, 219-229.
[14] Holden, N., & Freitas, A. (2004). Web page classification with an ant colony algorithm. International Conference on Parallel Problem Solving from Nature (pp. 1092-1102). Springer Berlin Heidelberg.
[15] Kim, S., & Zhang, B.-T. (2003). Genetic mining of HTML structures for effective web-document retrieval. Applied Intelligence 18, no. 3, 243-256.
[16] Manning, C., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Vol. 1. Cambridge University Press.
[17] Menczer, F., Pant, G., Srinivasan, P., & Ruiz, M. (2001). Evaluating topic-driven web crawlers. Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 241-249). ACM.
[18] Micarelli, A., & Gasparetti, F. (2007). Adaptive focused crawling. In The Adaptive Web (pp. 231-262). Springer Berlin Heidelberg.
[19] Navrat , P., Jastrzembska, L., & Jelinek, T. (2009). Bee hive at work: Story tracking case study. Web Intelligence and Intelligent Agent Technologies (pp. 117-120). IEEE/WIC/ACM International Joint Conferences on. Vol. 3. IET.
[20] Özel, S. (2011). A web page classification system based on a genetic algorithm using tagged-terms as features. Expert Systems with Applications 38, no. 4, 3407-3415.
[21] Saraç, E., & Özel, S. A. (2014). An ant colony optimization based feature selection for web page classification. The Scientific World Journal.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY