INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

A new approach to Web Crawling — DHEKTS Crawler in comparison with various Crawlers

K ThirugnanaSambanthan[1]*

1 Assistant Professor, Sankara College of Science and Commerce, Coimbatore, Tamil Nadu, India. Tel.: +91 9486191376

OPEN ACCESS

Received: 08.04.2021

Accepted: 13.04.2021

Published: 01.06.2021

Citation: ThirugnanaSambanthan K (2021) A new approach to Web Crawling — DHEKTS Crawler in comparison with various Crawlers. Indian Journal of Science and Technology 14(19): 1580-1586. https://doi.org/10.17485/IJST/v14i19.599

* **Corresponding author**.

Tel: +91 9486191376
dr.kts76@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To propose a crawler to visit websites for collecting information and create a search engine index for reference; To compare various crawler License, language used for creation, effectiveness with proposed DHEKTS crawler; To compare various characteristics, tasks and functions with proposed DHEKTS crawler; To identify the merits of the DHEKTS Crawler. **Methods:** A new Crawler called DHEKTS is developed to filter and synchronize documents like Images, Link, and HTML code from a given website. This Crawler is unique in nature since it returns all the details of a particular website having Images, Links, html code and contents. It can crawl through links in a specified website and crawl further to other links on the website. The DHEKTS Crawler is designed for Depth and Relevance crawling. The entire DHEKTS crawler has a few crawling mechanism supporting variety of information. The requirements are Operating System: Win 7 and higher, Front End: PHP, BackEnd: MySQL, RAM: Minimum 4GB and SERVER: High Speed Server with good storage Capacity. **Findings:** The DHEKTS Crawler has brought web related Links, Images, HTML Code, Information about to fifth level of crawling and Relevance Search giving relevant information. Multiple crawlers fulfill the major functions of crawling but DHEKTS CRAWLER is built to execute all functions in one crawler. **Applications:** This is applied in Crawling of various Websites and to retrieve valuable data.

Keywords: Crawler; DHEKTS Crawler; License; tasks; functions; effectiveness; Comparison

1 Introduction

A web crawler systematically browses WWW for the purpose of indexing. Using crawler, the web search engines updates web content, index other sites. Usually, crawler begins with a popular site and index words of pages following links within sites. Since WWW provides a great amount of useful information electronically in the form of hypertext, dynamically changing unstructured information, makes it difficult for requisite information. It is studied a web Crawler automatically traverses web by downloading documents page by page[(1)]. Further, crawling is made difficult since WWW has large volume dynamic pages[(2)].

According to crawler Junghoo Cho et al.,[3] an Internet bot systematically browses WWW[4] for web indexing. Web search engines use crawling to index sites and modernize web contents. It is noted that crawler copies pages for process by search engine. Roughly speaking, a crawler[5] starts off by placing an initial set of URLs in a queue and all URLs to be retrieved are kept and prioritized. From this queue, crawler gets an URL (in some order), downloads the page, extracts URLs from the downloaded page and puts them in URLs queue. Collected pages are later used for other applications such as a Web search engine or a Web cache.

In this study various Crawlers[6] like JSpider, Google Bot (Google), Httrack, Methabot, WebSphinix, Gnu Wget, WIRE, Pavuk, Scrapy, Bing Bot (Microsoft), Heritrix, Slurp 3.0 (Yahoo), WebHTTrack, MSN Bot (Microsoft), Web2disk are compared with DHEKTS Crawler. The functionality[7], effectiveness[8], tasks[9] performed by various crawlers are studied in detail. It is understood that the features of one crawler is not in other crawler and implementing all features in one crawler is not done. This problem is identified for this study to build a unique crawler to systematically browse WWW for indexing information, supporting multiple features of crawling like bringing links, images, HTML Source, Depth Crawling[10]. Thus, the desired work is to develop a crawler with all features of diversified crawlers[11], reducing time of referring multiple crawlers[12] to fulfill the task. Further the proposed crawler can be useful to consolidate the outcomes of crawling very easily.

DHEKTS Crawler

A new Crawler called DHEKTS is developed to filter and synchronize documents like Images, Link, and HTML code from a given website. This Crawler is unique in nature since it returns all the details of a particular website having Images, Links, Files and details of any website. It can crawl through links in a specified website and crawl further to other links in the website. The DHEKTS Crawler is designed for Depth and Relevance crawling. The entire DHEKTS crawler has a few crawling mechanism supporting variety of information.

Image Crawler

The DHEKTS Image Crawler is used to browse all images[13] (jpg, gif, png etc.) of a website recursively and collects multitude of images from the website. The images are viewed as thumbnail with respective URL links. These crawlers crawls all images of a website and display them with URL. Without storing resultant images in database, the DHEKTS Image Crawler directly display the results on the screen.

Link Crawler

The function of DHEKTS Link Crawler crawls all links of a website. The crawler crawl websites and gathers all internal and external links and produces Page heading, URL, hyperlink of the website. The Crawler acts like a site map provider for any website. It is also displaying the results without taking them to storage.

HTML Crawler

This Crawler crawl a website and lists all html links, html coding[14] of the entire website. It is useful to analyze coding techniques, structure of website. Though download or right click option is restricted, this crawler gets html code[15].

Depth Crawler

The DHEKTS Depth Crawler[16] crawl the entire website and continue crawling other websites based on the links of the website. Crawl depth is the degree to which a web search engine goes interior to a website. Majority of the sites contain multiple pages, subpages. The pages and subpages grow deeper in a manner similar to the way folders and subfolders (or directories and subdirectories) grow deeper in computer storage. By default a home page has a crawl depth 0. Pages linked within home page have a crawl depth value 1; pages linked directly within crawl-depth-1 page have a crawl depth value 2 and so on. The DHEKTS Depth Crawler is developed to have crawl depth value 5.

Relevance Crawler

Finally, the DHEKTS Crawler bringing relevant information from WWW is called Relevance Crawler. This Crawler works based on search keywords, no. of keywords present in a particular website, user relevance rating is given to the website.

Objectives:

1. To develop a crawler to visit websites for collecting information and create a search engine index for reference.
2. To compare various crawler License, language used for creation, effectiveness with proposed DHEKTS crawler.
3. To compare various characteristics, tasks and functions with proposed DHEKTS crawler.
4. To identify the merits of the DHEKTS Crawler.

Today, extracting images, links and html source are difficult on web. The existing crawlers are not sufficient for responding certain queries. Every crawler has its own specialization functions for crawling entire website and displaying the result, supporting multithreads, supporting HTTP proxies and cookies, partial local file system support etc. This paper is about a new approach in web crawling using DHEKTS Crawler which is quite different from prominent crawlers.

2 Materials and Methods

2.1 Architecture of Dhekts Crawler

The DHEKTS, a proposed crawler, is designed to overcome the difficulties of referring multiple crawlers[17] for searching. It is a unique system that can cater all information of a website. The whole system is divided into components like Image crawlering, Link crawlering, HTML crawlering, Depth crawlering and DHEKTS Search engine. The system has a component for initializing URL, loading DOM component for existence of URL, database for storage of information. The DHEKTS system plays a major role to search data on WWW for filtering intended objects[18]. It has multiple crawling functions applicable on various objects. This Crawler finds details of any website. It crawls links of a particular website. It is developed for deep crawling and implementing relevance[19] in results. The components of the proposed crawler are named after subjective crawling. The user crawl WWW for multiple objectives and the outcome can be stored in the database. It is required to choose appropriate component[20] in the system to go with purpose of crawling. The depth crawler is designed to crawl up to level 5 for interior crawling[21]. Since the work is crawling[22] WWW, the results obtained will be the existed information of website. The multiple features of the intended DHEKTS crawler is crawling images, links, HTML, depth and relevance crawling in single software.

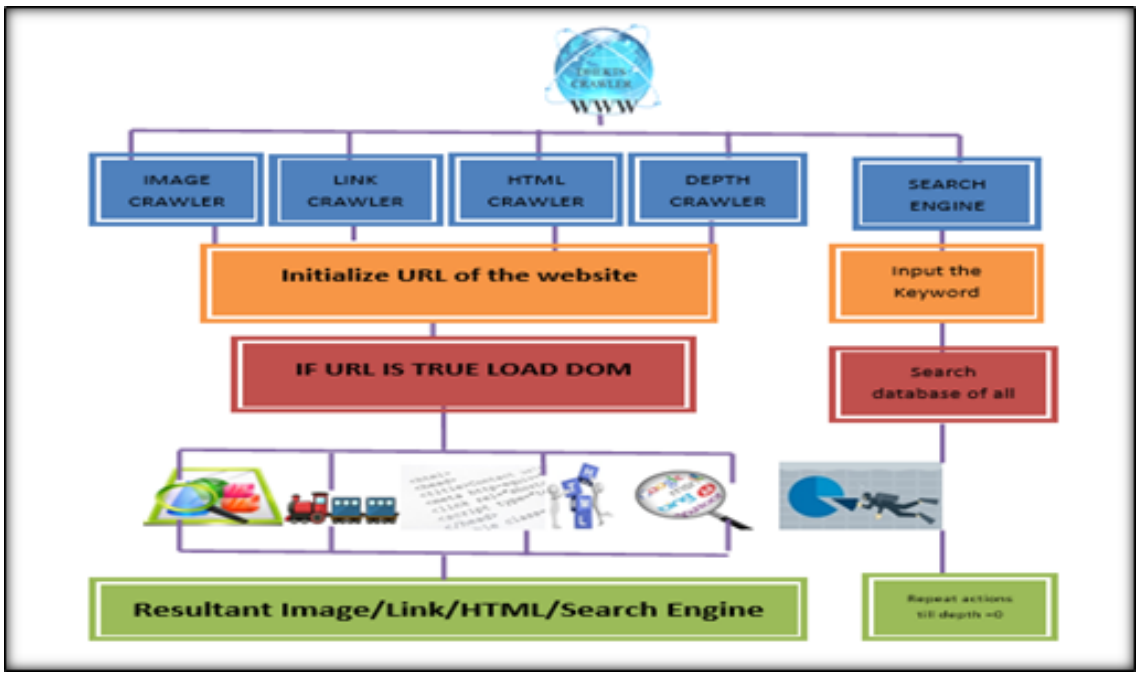

Fig 1. Architecture of DHEKTS WEB Crawler

2.2 Working of Dhekts Web Crawler

1. Initiate crawling

2. Input seed URL and determine an IP address for the target server using DomainNameServer.
3. Extract Robot.txt file from the server and verify permission.
4. Verify protocol of underlying host like http, ftp, gopher etc.
5. Based on protocol host, download document.
6. Identify the document format like doc, xls, ppt, html, or pdf etc.
7. Extract links or references of the websites
8. Store the document and URLs in search engine buffer
9. Repeat steps 1 to 9 till the queue is empty.

Starting with the seed URL, the DHEKTS crawler crawls all links found in HTML page till the URL in a designated queue is empty.

3 Results and Discussion

3.1 Functions of Various Crawlers

The tasks of various crawlers are organized below.

Table 1. Functions of various Crawlers

S.No	Crawler Name	Characteristics
1	JSpider	It checks the site for errors, Outgoing and/or internal link checking. It can download complete website.
2	Google Bot (Google)	It retrieves the content of web pages (i.e.) the words, code and resources that build up. It then sends the collected information to Google Search Engine [23] .
3	Httrack	It sequences download sites. It updates download sites and open page by browser.
4	Methabot	It has complete, new and very capable configuration system. It has Full HTTP crawling support. It has partial local file system support
5	WebSphinix	It supports retrieval of Multithreaded web pages. It supports reusable page content classifiers. It has the Support to robot exclusion standard.
6	Gnu Wget	It supports HTTP proxies and HTTP cookies. It can resume downloads which are aborted. It can also use filename wild cards.
7	WIRE	It is highly scalable. It has all the parameters for crawling and indexing. It has high Performance as it is executed in a number of machines.
8	Pavuk	It provides full information about transfers. It has optional multithreading support. It has JavaScript bindings to allow scripting of meticulous tasks.
9	Scrapy	It is easy to setup and use. It is faster and it is excellent developer documentation. It is an excellent choice for focused crawls. It has a established framework with redirection handling, full unicode, odd encodings, integrated http cache etc.
10	Bing Bot (Microsoft)	It fetches page from your website and sends to the mobile friendliness classifier for a real-time verdict. It takes some time to fetch and analyze the page and show the judgment.
11	Heritrix	Web crawler for archiving. It crawl contents of Internet archives.
12	Slurp 3.0 (Yahoo)	It crawl websites, meta tags, and traverse through links for search engine indexing and then gives back to Yahoo searchable database.
13	WebHTTrack	It has an integrated help system. It sorts the original site's relative links. It can update a mirrored site, resume interrupted downloads.
14	MSN Boot (Microsoft)	Documents of web build a searchable index for MSN Search engine
15	Web2disk	It automatically saves snapshots of the website. It monitors websites for update, downloads dynamic pages. It has powerful filtering capability.
16	DHEKTS	Image, Link, HTML, depth and Relevance Crawling are implemented in this crawler. Depth Crawling to the level 5 is implemented. Size of the software is less compared to other software and effective in crawl the website. It is a good choice to crawl websites.

3.2 Crawlers Platform

This table is helpful to identify the Authorization, Operating System, and Language for development of web crawlers.

Table 2. CrawlerLanguages

S.No	Crawler Name	License	Language	Operating System
1	JSpider	General Public License version 2.0	Java	Windows
2	Google Bot (Google)	-	Python	Windows
3	Httrack	GPL License	C	Cross-platform
4	Methabot	ISC License	C	Cross-platform
5	WebSphinix	Apache Software License	Java	Windows, Mac, Linux, Android, IOS
6	Gnu Wget	GNU GPL	C	Cross-platform
7	WIRE	GPL License	C/C++	Cross-platform
8	Bing Bot (Microsoft)	-	-	Windows
19	Heritrix	Apache License	Java	Linux/Unix/Windows Unsupported
10	WebHTTrack	GPL License	C/C++	Cross-Platform
11	MSN Boot (Microsoft)	-	-	Windows
12	Web2disk	-	-	Windows
13	DHEKTS	-	PHP	Cross-platform

Performance Measure

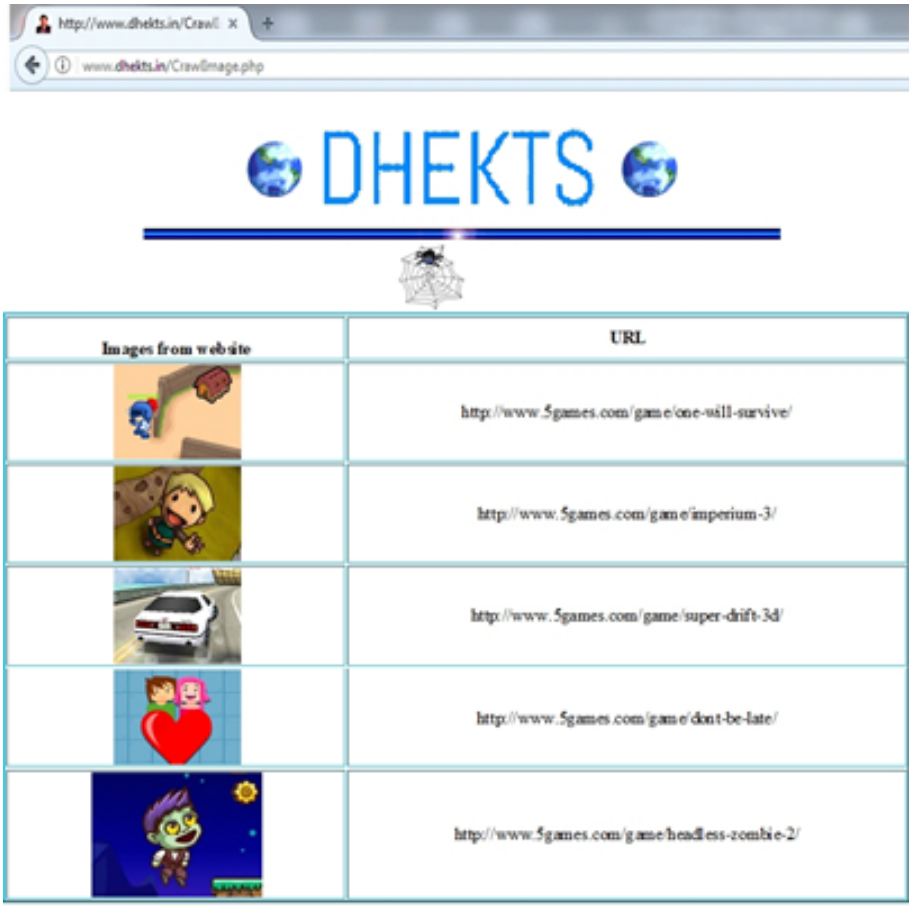

Fig 2. Output of the DHEKTS Image Crawler

The first feature of DHEKTS is crawling images of websites. It has crawled images of 5games.com, clearly the mining process of DHEKTS crawler mined WWW and has displayed all images linked with concerned websites. The results are tabulated with images and appropriate URLs.

Page Heading	URL	Hyperlink
Mobile	www.icc-cricket.com/mobile	Mobile
Mediazone	http://www.icc-cricket.com/mediazone/	Mediazone
Facebook	www.facebook.com/icc	Facebook
Twitter	www.twitter.com/icc	Twitter
Youtube	www.youtube.com/cricketicc	Youtube
Instagram	http://instagram.com/icc	Instagram
International Cricket Council	http://www.icc-cricket.com	International Cricket Council
Home	http://www.icc-cricket.com	Home

Fig 3. Output of the DHEKTS Link Crawler

The second feature of DHEKTS crawler is crawling links of websites. It has crawled links of icc-criket.com. Clearly, DHEKTS crawler has displayed all associated links of crawling websites by mining process. The results are tabulated with page heading, URL and hyperlinks.

The performance of DHEKTS Crawler crawling different websites are compared with time

Table 3. Performance measure of DHEKTS Crawler

Website	DHEKTS(Sec)
5games.com	0.053 (Relevant)
http://jic-edu.sa/	0.035 (Relevant)
http://www.wallpaperstop.com/	0.026 (Relevant)

The Precision of DHEKTS Crawler is always one since the number of retrieved the relevant pages are same. Multiple crawlers fulfill the major functions of crawling but DHEKTS CRAWLER is built to execute all functions in one crawler.

Merits of Dhekts Crawler

1. **Robustness:** The web server has spider traps misleading crawler, responding unwanted pages, displaying countless page in a particular domain. DHEKTS Crawler is designed to sustain such traps.
2. **Extensible:** Crawlers must be extensible (i.e.) it should handle new data formats, fetch new protocols etc. DHEKTS Crawler is modular in nature and it is extensible.
3. **Scalable:** The scalability of a system is attained by adding resources. DHEKTS Crawler has a good crawl rate.
4. **Performance efficiency:** The crawl output is compared to anticipated result. DHEKTS Crawler has performed well with expected results.
5. **Politeness :** Two types of policies, implicit and explicit in crawler regulate the rate of visiting websites. DHEKTS Crawler is formulated to politely crawl on sites.
6. **Distributed:** The crawler function must work in distributed environment. The proposed approach has worked in different

requirement of the hour which can better understand the user requests. The work can be extended to build an effective content mining crawler to satisfy future trends.

References

1) Dhenakaran SS, Sambanthan KT. Web crawler-an overview. *International Journal of Computer Science and Communication*. 2011;2(1):265–272. Available from: www.csjournals.com/IJCSC/PDF2-1/Article_49.pdf.
2) Amudha S. Web crawler for mining web data. *International Research Journal of Engineering and Technology*. 2017;4(2):128–136. Available from: https://www.irjet.net/archives/V4/i2/IRJET-V4I225.pdf.
3) Cho J, Garcia-Molina H. Synchronizing a Database to Improve Freshness. *ACM SIGMOD Record*. 2000;29(2):117–128. Available from: https://doi.org/10.1145/335191.335391.
4) Berners-Lee T, Cailliau R. WorldWideWeb: Proposal for a Hypertext Project. 1990. Available from: https://www.w3.org/Proposal.html.
5) AbuKausar M, Dhaka VS, Singh SK. Web Crawler: A Review. *International Journal of Computer Applications*. 2013;63(2):31–36. Available from: https://dx.doi.org/10.5120/10440-5125.
6) Sambanthan KT, Dhenakaran SS. Web change monitoring and tracking tools. *International journal of Computer Science & Communication*. 2011;2(2):451–454. Available from: http://www.csjournals.com/IJCSC/PDF2-2/Article_32.pdf.
7) Kumar N, Aggarwal D. LEARNING-based Focused WEB Crawler. *IETE Journal of Research*. 2021. Available from: https://doi.org/10.1080/03772063.2021.1885312.
8) Devi RS, Manjula D, Siddharth RK, Ackerman MS, Starr B, Pazzani MJ. An efficient approach for web indexing of big data through hyperlinks in web crawling. *The Scientific World Journal*. 1997;97:17–31. Available from: https://doi.org/10.1155/2015/739286.
9) Lu H, Zhan D, Zhou L, He D. An improved focused crawler: using web page classification and link priority evaluation. *Mathematical Problems in Engineering*. 2016;6406901. Available from: https://doi.org/10.1155/2016/6406901.
10) Zheng Q, Wu Z, Cheng X, Jiang L, Liu J. Learning to crawl deep web. *Information Systems*. 2013;38(6):801–820. Available from: https://doi.org/10.1016/j.is.2013.02.001.
11) Kumar M, Bhatia R. Design of a mobile Web crawler for hidden Web. In: 2016 3rd International Conference on Recent Advances in Information Technology (RAIT). 2016;p. 186–190. Available from: 10.1109/RAIT.2016.7507899.
12) Patil Y, Patil S. Review of web crawlers with specification and working. *International Journal of Advanced Research in Computer and Communication Engineering*. 2016;5(1):220–223. Available from: 10.17148/IJARCCE.2016.5152.
13) PS. An image crawler for content based image retrieval system. *International Journal of Research in Engineering and Technology*. 2013;02(11):33–37. Available from: https://dx.doi.org/10.15623/ijret.2013.0211006.
14) Xu S, Yoon HJ, Tourassi G. A user-oriented web crawler for selectively acquiring online content in e-health research. *Bioinformatics*. 2014;30(1):104–114. Available from: https://dx.doi.org/10.1093/bioinformatics/btt571.
15) Bra PD, Houben GJ, Kornatzky Y, Post R. Information Retrieval in Distributed Hypertexts. *InRIAO*. 1994;48:1–493.
16) Agrawal N, Johari S. A Survey on Content Based Crawling for Deep and Surface Web. In: and others, editor. 2019 Fifth International Conference on Image Information Processing (ICIIP). 2019;p. 491–496. Available from: 10.1109/ICIIP47207.2019.8985906.
17) Sarveshachodankar A, Michael, Walke S, Dr CH, Patil. Literature review on Web Crawling. *International Journal of Engineering Research & Technology*. 2020;8(5). Available from: https://www.ijert.org/literature-review-on-web-crawling.
18) Yu L, Li Y, Zeng Q, Sun Y, Bian Y, He W. Summary of web crawler technology research. *In Journal of Physics: Conference Series 2020*;1449(1):12036. Available from: https://doi.org/10.1088/1742-6596/1449/1/012036.
19) Daneshpajouh S, Nasiri MM, Ghodsi M. A Fast Community Based Algorithm for Generating Web Crawler Seeds Set. *WEBIST*. 2008;p. 98–105. Available from: https://www.researchgate.net/profile/Shervin_Daneshpajouh/publication/220724572_A_Fast_Community_Based_Algorithm_for_Generating_Web_Crawler_Seeds_Set/links/0046352652e14da34a000000.pdf.
20) Devi RS, Manjula D, Siddharth RK. An Efficient Approach for Web Indexing of Big Data through Hyperlinks in Web Crawling. *The Scientific World Journal*. 2015;2015:1–9. Available from: https://dx.doi.org/10.1155/2015/739286.
21) Shaker M, Ibrahim H, Mustapha A, Abdullah LN. A framework for extracting information from semi-structured web data sources. *In2008 Third International Conference on Convergence and Hybrid Information Technology*. 2008;1:27–31. Available from: https://doi.org/10.1109/ICCIT.2008.60.
22) van Steen M, Tanenbaum AS. A brief introduction to distributed systems. *Computing*. 2016;98(10):967–1009. Available from: https://dx.doi.org/10.1007/s00607-016-0508-7.
23) Bar-Yossef Z, Mashiach LT. Local approximation of pagerank and reverse pagerank. *InProceedings of the 17th ACM conference on Information and knowledge management*. 2008;p. 279–288. Available from: https://doi.org/10.1145/1458082.1458122.