Extraction System Web Content Sports New Based On Web Crawler Multi Thread

Extraction System Web Content Sports New Based On Web Crawler Multi Thread

Y D Pramudita, D R Anamisa, S S Putro, M A Rahmawanto

[1,2,3,4] Departement of Informatic Engineering, University of Trunojoyo Madura, Bangkalan – Madura, Indonesia

[1]yoga@trunojoyo.ac.id

Abstract. Web crawlers are programs that are used by search engines to collect necessary information from the internet automatically according to the rules set by the user. With so much information about sports news on the internet, it takes web crawlers with incredible speed in the process of crawling. There are several previous studies that discussed the process of extracting information in a web document that needs to be considered both in terms of both aspects, including in terms of the structure of the web page and the length of time needed. Therefore, in this research the web crawler application was developed by applying a multi-thread approach. This multi-thread approach to research is used to produce web crawlers that are faster in the process of crawling sports news by involving news sources more than one address at a time. In addition to the multi-thread approach, adjusting the structure of the website pages is also done to ensure the information to be extracted by web crawling. From the results of the multi-thread implementation test on the crawling process, this study has been able to increase speed compared to the single-thread method of 122.95 seconds. But the results of web update detection, have resulted in a speed that decreased by 6.27 seconds in the crawling process with unequal data and the speed on the crawling process has also decreased by 24.76 seconds on server 1 and by 23.92 seconds on server 2. **Keywords:** W*eb Crawlers, Multi-Threads, Detection of Updated Web Page Content, Distributed Systems.*

1. Introduction

On online news sites there are many desired news categories. Visitors just choose the desired news and take it. The existence of online news sites, nowadays makes it very easy to get news, especially news about sports. Sports news is one of the headlines and is loved by various groups. Various national and local media use sports news to look for a profit. The development of print media in Indonesia which discusses sports is increasingly rapid and interesting to study. Therefore, in this research build a sports news search engine by implementing a web crawler. Web crawlers are machines that browse websites to collect documents or data found on the website visited[1]. Several previous studies built a news search engine by implementing a web crawler on online news sites to get the desired news data. While crawling is the process behind a search engine that is tasked with tracking the World Wide Web in a structured manner with certain ethics[2]. The website structure is grouped into four, such as: linear, non-linear, hierarchical, and hybrid[3]. Of the 4 groups of structures have differences, including: (1) Linear structure, is a website structure that only has a chain of links sorted and no branching. This structure is suitable for displaying information that is not too interactive,

(2) Non-linear structure is a linear structure that allows for branching. Then in this structure there is no master page or slave page, (3) Structure of Hierarchy, this structure uses branching to display data based on certain criteria, in the main view it is called the master page while for branch display is called slave page, (4) Hybrid Structure , a combined structure where this structure combines all existing structures. This structure can provide high interaction to users.

The process that is carried out on the web crawler is extracting the main URL to generate a list of pages that are accessed, then visiting all the pages listed, and identifying all the hyperlinks found on the page and adding the URL to the list of links to the pages that have been visited[4]. In previous studies the application of information retrieval with the GVSM method has been able to improve performance by more than 50%[5]. Changes from a web page can be detected by calculating the value of the checklist. The use of this method gives results that are very fast and require a short time[6]. However, a number of previous studies have used search engines by applying web crawlers by using one thread for information gathering processes, so that they are unable to fulfill user requests for information both in terms of quantity and quality[7]. But the one thread approach is only to retrieve data with certain specifications, for example the topic of 'sports', then web crawlers do web pages that only relate to the topic of sports. Therefore, this research has developed a sports news web content extraction system by applying multi-thread programming techniques. Multi-threaded approach is one approach in computer programming that aims to make the process run simultaneously[8]. The development of a publicly distributed multi-threaded web crawler proxy that is available is easier than distributing many web crawlers to many computers and controlled by one computer[9]. With this technique, it is possible to make web crawlers able to handle multiple requests for information from users at one time. The advantage of web crawlers by implementing multi-threads can provide results more relevant to keyword search and faster with the application of multithreading and distributed computing techniques[10]. But the multi-thread technique in the process of extracting the contents of a website page is very dependent on the structure of each website that will be taken so that the weight used for the extraction process must adjust the structure of the website. The website structure is grouped into 4, namely linear, non-linear, hierarchical, and hybrid structures[11]. Data stored from crawling results is large enough so that it will make the server work harder and the risk of data access failure by search engine users is more likely.

While the database implementation in this study uses the method or technique of transactional replication. Database replication can allow sharing of load access to the server, so as to minimize the risk of access failure by search engine users[12]. Database replication is a technique that can be used to optimize data access and provide fault tolerance by storing copies of data in several locations. Database replication can use 3 ways, namely snapshot replication, merging replication, and transactional replication. Snapshot replication is used to copy data on the database server to another database on the same or different server. Merging replication is used to merge data from two or more databases that are combined into one database. Transactional replication is used for complete copying of the master database, then gets periodic updates on changes that occur in the master database[13]. The purpose of using the transactional replication technique is by utilizing the master to master component rather than the master to slave provided by the database to be used. Because if you use master to slave only on the master database that can provide data updates and replicate to the slave database, the slave database can only receive updates from the master database. A web crawler needs to update the contents of the database regularly to increase the update and quality of information in the database. Updates to a website can be seen from the changes on the main page of the website. Some ways to find out changes from a website or website that have been updated, namely changes in page structure, changes in text content and image changes (hyperlinks) contained on the webpage[14].

The purpose of this research is to increase the speed of the web crawler process by applying multi-thread techniques in collecting sports news and then crawling data is distributed with master to master database replication techniques. Then in gathering sports news, only collecting the latest sports news uses a technique of comparing url or news links on 10 predetermined sports domains.

2. Research Methods

In this research using maximum up-to-date sports news data and then crawling regularly every 1 hour. The information that will be collected in this research is information about sports news with 10 predetermined domains, including:

1. www.sports.okezone.com/
2. www.bola.com/
3. www.cnnindonesia.com/olahraga
4. www.mediaindonesia.com/news/list/olahraga
5. www.merdeka.com/olahraga/
6. www.republika.co.id/kanal/olahraga
7. www.bbc.com/indonesia/olahraga
8. www.m.metrotvnews.com/olahraga
9. www.kompasiana.com/olahraga
10. www.sport.detik.com/

From the selection of 10 domains based on the top of ranking site in the scope of Indonesia from www.alexa.com on June 30, 2017, then sites - sites related to sports news were selected. And the purpose is to find out the changes found on web pages by comparing changes to the content or content. Fill or content is 'url' or the news link. Web crawlers will create a list of indexes to facilitate the search process [15]. Content changes are changes from a web page caused by the addition or reduction of the contents of the web. For example, a news site, on a news site, will always change its contents when there is more recent news[16].

In this study using a multi-threaded technique on web crawlers to compare the latest news and old news. Then the data from crawling stored in the database can be distributed by replicating from the master database to other masters, where computer servers and other master computers can communicate with each other in the database. So that any changes to the master computer can always be replicated on another master computer or vice versa. Then this system can add a website update detection feature by comparing upload dates and hours on sports news so that the news obtained is all the latest sports news from web pages with 10 predetermined domains.

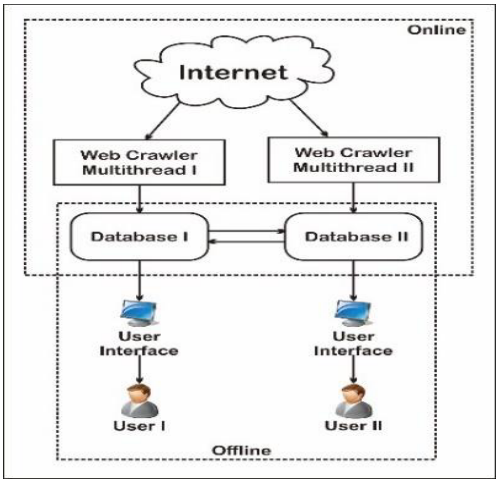

Figure 1. Distributed Multi-threaded Crawler Web Architecture

The system developed in this study is a multithreaded web crawler with distributed on two servers to collect sports news. The general description of the system to be developed can be seen in Figure 1, where the crawlerMulti-thread web architecture is distributed in this study in the form of a multi-threaded web crawler program distributed into two servers, then the results of collecting each web crawler are saved to the database. And continued with the 2nd synchronization of the database using the master to master database replication technique. Data that has been stored in the database can be seen by the user through the user interface. Distributed system is a process of distributing information

on two or more computers. There is a general architecture, namely client-server and peer to peer. In this study will use peer to peer architecture. Peer to peer architecture is part of a distributed system model where the system can simultaneously function as a client or server. Peer-to-peer is the most general and flexible model [17].

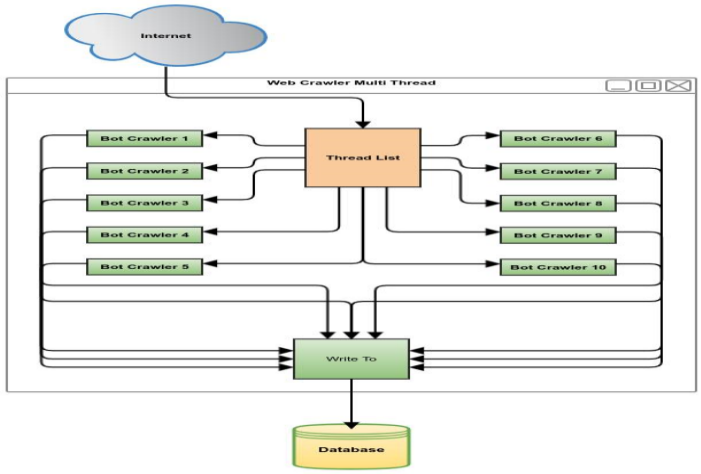

Figure 2. Multi-thread Web Crawler Architecture

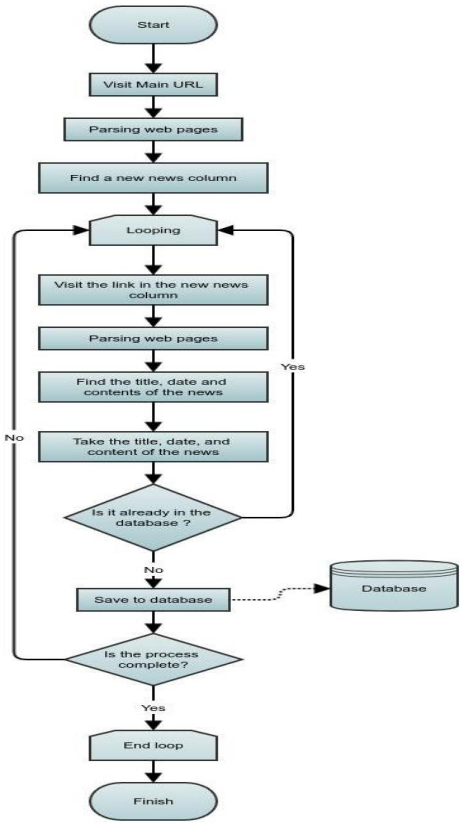

Figure 3. Flowchart Bot of Crawler

The multi-thread web crawler runs a number of 10 bot of crawlers to visit and extract content from 10 predefined web pages, shown in Figure 2. The use of 10 bot of crawlers in this extraction process is due to data from 10 websites that will be extracted features has a different programming structure and format. Then the working process is that the bot of crawler is run simultaneously without having to wait for one of the crawler bots to finish on the multi-thread architecture. The results of extracting 10

web pages will produce title elements, news dates, and news content. Furthermore, the storage of elements (data) to the database and users can see through the interface of this system.

In addition to the system architecture, in this research there is a workflow for 10 bot of crawlers used in the extraction process, shown in Figure 2, where a multi-thread bot of crawler visits the main url of the web to be crawled. After that parsing the html page. In this research, html parsing uses libraries in the python programming language BeautifulSoup version 4. After parsing the page, then look for the news column whose contents are the latest news from the webpage. After the latest news column is found, visit the url and find the elements of the title, date, and content of the news. Before saving the data to the database, check first, whether the url of news that will be stored is already in the database or not. And if it's not there, then save it first for the data to the database, and if it's already there then continue visiting the next url. This web update detection feature aims to prevent web crawlers from storing the same data. Repeat the process until no url is found in the latest news column.

3. Experiment Result and Discussion

After testing, the following is a comparison of the crawling process time needed to gather news from 10 websites that have been determined from each trial, can be seen in Table 1.

Table 1. Comparison of Test Results

Concept	Total News	Total of Crawling Time
Single-thread	213	212,03714
Multi-thread	216	89,0847
Multi-thread Update	216	95,35275
Multi-thread Update Distributed	214	134,6023
Multi-thread Update Distributed of Server 1	213	120,11111
Multi-thread Update Distributed of Server 2	216	124,20546

Based on table 1 there is a total crawling time of 10 websites needed with a single-threaded web crawler longer than a multi-thread web crawler that has a difference of 123 seconds. Whereas the comparison between multi-thread web crawlers and multi-thread web crawlers has been added to the web update detection system, while the time needed for crawling is getting longer and this has a difference of 6 seconds compared to multi-thread web crawlers without an update feature the web.

The results of the application of the concept that the effect of distribution also requires a considerable amount of time for the crawling process, shown in the 4th experiment, where it has divided the number of urls visited to each server and takes 39 seconds longer than the multi crawler web -thread update. Whereas shown in experiment 5, running the system alternately on each server takes 25 seconds longer than the multi-thread update web crawler on the first and 29 seconds on the second server.

4. Conclusion

Based on the test results of the system that has been developed, conclusions can be taken such as: (1) The application of the multi-thread programming concept can increase the crawling process speed by 123 seconds compared to single-thread web crawlers. (2) Application of web update detection on multi-thread web crawlers has decreased the crawling process speed by 6 seconds, compared to multi-thread web crawlers without detection of website updates. (3) The implementation of distribution does

not increase the speed of multi-thread web crawlers with update detection, it actually decreases the speed of the crawling process which is equal to 25 seconds on servers 1 and 24 seconds on server 2.

References

[1] H. C. Wang, S. H. Ruan, and Q. J. Tang, "The implementation of a web crawler URL filter algorithm based on caching," *2nd Int. Work. Comput. Sci. Eng. WCSE 2009*, vol. 2, pp. 453–456, 2009.

[2] N. Kumar and M. Singh, "Framework for Distributed Semantic Web Crawler," *Proc. - 2015 Int. Conf. Comput. Intell. Commun. Networks, CICN 2015*, pp. 1403–1407, 2016.

[3] T. Evi, "Analysis of Web Application Development," vol. 2009, no. semnasIF, pp. 122–127, 2009.

[4] F. Ahmadi-Abkenari and A. Selamat, "A clickstream-based web page significance ranking metric for web crawlers," *2011 5th Malaysian Conf. Softw. Eng. MySEC 2011*, no. 1, pp. 223–228, 2011.

[5] J Pardede, Multithreading Implementation to Improve Information Retrieval Performance with the GVSM Method,Journal of Computer System, Vol.4, No.1, pp. 1-6, Mei 2014.

[6] S. Shekhar, R. Agrawal, and K. V. Arya, "An architectural framework of a crawler for retrieving highly relevant web documents by filtering replicated web collections," *ACE 2010 - 2010 Int. Conf. Adv. Comput. Eng.*, pp. 29–33, 2010.

[7] S. Gupta and K. K. Bhatia, "CrawlPart: Creating crawl partitions in parallel crawlers," *Proc. - 2013 Int. Symp. Comput. Bus. Intell. ISCBI 2013*, pp. 137–142, 2013.

[8] W. Guo, Y. Zhong, and J. Xie, "A web crawler detection algorithm based on web page member list," *Proc. 2012 4th Int. Conf. Intell. Human-Machine Syst. Cybern. IHMSC 2012*, vol. 1, pp. 189–192, 2012.

[9] G. H. Agre and N. V. Mahajan, "Keyword focused web crawler," *2nd Int. Conf. Electron. Commun. Syst. ICECS 2015*, pp. 1089–1092, 2015.

[10] Z. Shi, M. Shi, and W. Lin, "The Implementation of Crawling News Page Based on Incremental Web Crawler," *Proc. - 4th Int. Conf. Appl. Comput. Inf. Technol. 3rd Int. Conf. Comput. Sci. Appl. Informatics, 1st Int. Conf. Big Data, Cloud Comput. Data Sci. Eng. ACIT-CSII-BCD 2016*, pp. 348–351, 2017.

[11] A. A. Wardekar and P. Gupta, "Smartcrawler : a Personalized Web," *2018 9th Int. Conf. Comput. Commun. Netw. Technol.*, pp. 1–4, 2018.

[12] H. R. Wang, C. F. Li, L. F. Zhang, and M. Y. Shi, "Anti-Crawler strategy and distributed crawler based on Hadoop," *2018 IEEE 3rd Int. Conf. Big Data Anal. ICBDA 2018*, pp. 227–231, 2018.

[13] S. Sharma and P. Gupta, "The anatomy of web crawlers," *Int. Conf. Comput. Commun. Autom. ICCCA 2015*, pp. 849–853, 2015.

[14] K. S. Shetty, S. Bhat, and S. Singh, "Symbolic verification of web crawler functionality and its properties," *2012 Int. Conf. Comput. Commun. Informatics, ICCCI 2012*, pp. 1–6, 2012.

[15] F. Ye, Z. Jing, Q. Huang, and Y. Chen, "The Research of a Lightweight Distributed Crawling System," *Proc. - 2018 IEEE/ACIS 16th Int. Conf. Softw. Eng. Res. Manag. Appl. SERA 2018*, pp. 200–204, 2018.

[16] E. Karlsson, "Improving landfill monitoring programs with the aid of geoelectrical - imaging techniques Distributed Web-Crawler and geographical information systems Hans Bjerkander," 2005.

[17] S. Hosen, M. A. Islam, M. M. Arshad, A. M. Khan, and M. K. Alam, "Talent Management: An Escalating Strategic Focus in Bangladeshi Banking Industry," *Int. J. Acad. Res. Bus. Soc. Sci.*, vol. 8, no. 1, 2018.