WEB MINING ON INDONESIA E-COMMERCE SITE : LAZADA AND RAKUTEN

Humasak Simanjuntak, Novitasari Sibarani, Bambang Sinaga and Novalina Hutabarat

Department of Informatics Management, Institut Teknologi Del, Indonesia

ABSTRACT

E-commerce site grows rapidly since it allows someone to shop online quickly and easily without having to meet seller directly. This saves time, effort, and cost in transaction although it doesn't always provide what the customers need. They must visit several e-commerce sites to get appropriate product. Google shopping has already accumulated some foreign e-commerce sites, but it is not available for Indonesia. Therefore, it is necessary to have an Indonesia website or site summary that can show product summary from several Indonesia e-commerce site. Site summary built with applying web content mining by using web data extraction technique. Some processes in web data extraction are e-commerce site file crawling, parsing html file, saving data into database and then displaying the data into a site summary. By applying these processes, data from several e-commerce sites can now be displayed on a site called site summary.

KEYWORDS

Web content mining, web data extraction, Indonesia e-commerce site

1. INTRODUCTION

Today, technology has already changed human behavior on society globally. This changes lead the world to borderless, especially in social, economic and cultural. One example is e-commerce site. E-commerce is the use of the Internet and the Web to transact business. More formally, we focus on digitally enabled commercial transactions between and among organizations and individuals [1]. Many e-commerce sites allow customer to shop online quickly and easily without having to meet with the seller directly. Customer just adds product to cart, pay electronically, and receive their product soon. Each e-commerce provides many kinds of product, such as clothes, food, furniture, electronics and accessories.

In China and Japan, the e-commerce sites market is rapidly growing. It supported by the increasing of internet user and many research about e-commerce. Many companies compete in the e-commerce market. In fact, there are some researches was did to win the competition in e-commerce market. The research is not only about technology but also prediction when a customer will purchase and web recommendation for customer. Sato and Asahi [2] developed a model which predicts the day a customer will purchase considering customer's individual-level heterogeneity. This research also compared with some prior studies. Zhang and Yang [3] proposed a recommendation system model based on dissimilarity clustering and association rules, then Xu and Zhang [4] proposed recommendation system model based on Ontology that assists customer to choose the appropriate e-commerce site. Google also support it where on 2010 google shopping already available in Japan.

Google surveyed 1.300 Indonesians and suggests strong future growth for Indonesia's ecommerce industry. Survey shows half of the Indonesian people who currently don't shop online are

DOI : 10.5121/ijdms.2015.7101

interested in making online purchases and will probably do so in the next 12 months [5]. But, it is not supported with researches that help customer to do transaction or purchasing. There are many e-commerce sites appear, but customers difficult to choose the right site. Many e-commerce sites offer products with different prices, types, descriptions, discount, brand, and different quality so that the customer must visit several e-commerce sites to get the right product. They need to compare the products so that they get the approriate ones.This process gives difficulties to the customer because it requires longer time for them to choose the right product. Moreover, google shopping is not available in Indonesia store. Customer needs one site that provides a summary of several e-commerce sites called site summary. Site summary assists customers make right decision when they are doing online transaction.

Google Shopping (*www.google.com/shopping*) is a site that provides summary of some products from several online stores, including Shop, ArtFire eBay and Newegg [6] see figure 1. Each product offered on Google shopping also sells its products in its online store (e-commerce site). Google shopping shows name, price, review, description, and e-commerce site of product clearly in one page. In Indonesia, there are many popular e-commerce sites, such as Lazada, Rakuten, Bhinneka, Zalora, Blibli, etc. Unfortunately, these sites do not include in google shopping. Furthermore, there is no local site like Google Shopping in Indonesia, therefore it needs more effort for customer to search their needs in Indonesia e-commerce site. It is important that an Indonesia site summary give brief description about some products from many e-commerce sites. So, the question for this research is how to mine Indonesia e-commerce sites to display in one site summary.

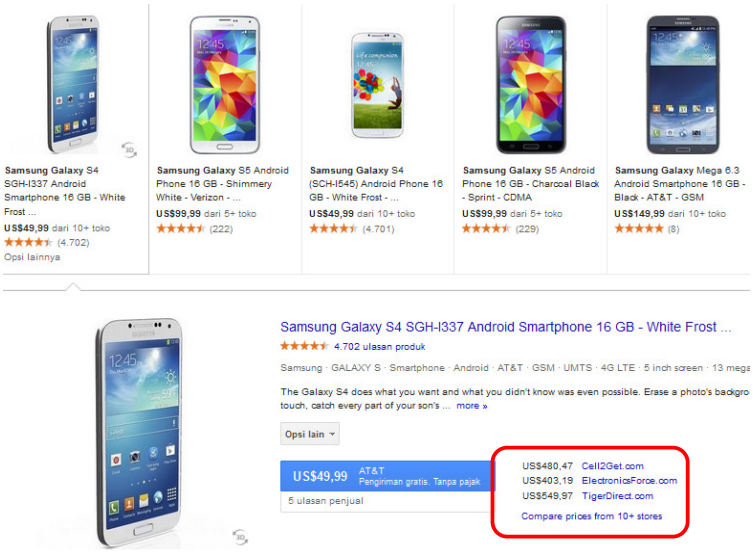

Figure 1 Samsung Galaxy S4 summary in Google Shopping [6]

There is a technique used for making a site summary (to get data from some sites and show the relevant product in one page) called web mining. There are three type of web mining, they are: Web structure mining, web content mining, and web usage mining [7] [8]. Web structure mining or web log mining is a technique used to find link structure from hyperlink and build summary of website [7]. Web structure mining is often used to decide a website pagerank. Web content mining is an extraction process to find useful information from data, document, audio, video or metadata in the web. This technique extracts the keyword from the web to build useful information [9]. The latter one, web usage mining is a technique used to discover pattern of user behavior related to one site from web data, log, click stream, cookies to improve the site service to the user [10]. This method use data mining techniques to discover interesting usage patterns

from Web data, in order to understand and better serve the needs of Web-based applications [11]. This web mining research focuses on web content mining, especially in web data extraction.

Some researches related to the web mining on Indonesia e-commerce site has not been available yet, but there are various researches about web content mining which are already published. Arvind Kumar Sharma [7] discussed in his research paper about "Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining". He gives detailed study and analysis of each web mining tools to scan the many HTML documents, images, and text provided on Web pages. The resulting information is provided to the search engines, in order of relevance giving more productive results of each search. This research also gives the taxonomy of content mining technique. One type of web content minings that will be used in this research is web data extraction language that converts web data to structured data and delivers it to the end user. The data is stored in the form of table [8]. Some purposes of web data extraction, are as the followings:

a. To get information from web that is used on other application
b. To retrieve information such as web search engine
c. To allow user to access data from a web

The recent research about web mining using tag HTML did by Khirade Rajratna Rajaram. He did text mining to the tree of tag HTML based on query result to find out the required data [12]. But, it need more time to execute it, because we need to build tree of HTML tag for each page in the website.

Web content mining becomes complicated when it has to mine unstructured, structured, semi structured and multimedia data. One tool that helps mine semi-structured or structured data mining technique is web crawler. A Web crawler (also known as Web spider or robot) is a computer program that inspects the Web in a methodical manner and retrieves targeted documents [13]. The program starts from a seed list of the URLs to be segmented in a queue. Web pages are then fetched from this queue and links are extracted from these web pages. This process ends after a specified time or when no new interesting URL can be found.

Therefore, the proposed solution in this research is implementing mining especially web data extraction on several Indonesia e-commerce sites for supporting data to the site summary that displays comparation of the same products from several sites in one site. The scope of this research is web mining carried out on two examples of Indonesia e-commerce sites which have products' similarity. Web data extraction did to the data offline in local hardisk, process the data, stored in the database and display in the summary site. The application and the site summary were built using Java Web.

This paper is organized as follows: Section 1 describes clearly the problems occurred and the background of doing the research. The general objective and proposed solution also be stated in this section. Continue with section 2 explains the proposed method for mining Indonesia e-commerce site. Continued with section 3 describes steps of the research in order to get the solution of problem stated in section 1. In section 4 there will be explanation about analysis result to present our solution to solve the problem and define the result of evaluation process. Finally, in section 5 we point out the result of the research and also raise future research work related to this field in section 6.

2. PROPOSED METHOD

According to the literature review, this research proposes method to mine Indonesia e-commerce sites. The proposed web data extraction method can be seen in Figure 2.

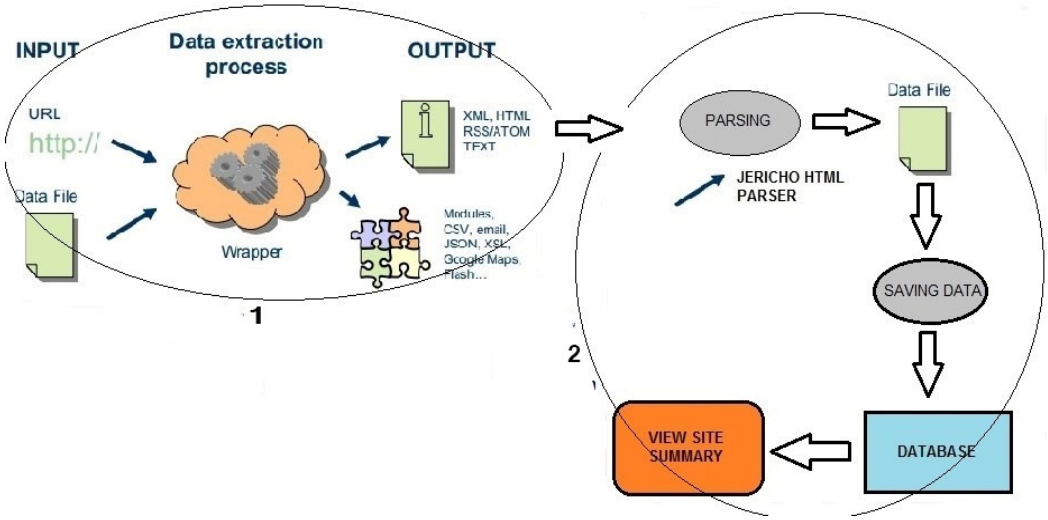

Figure 2 Proposed method for mining Indonesia E-commerce sites

Two stages of the proposed method for mining Indonesia E-Commerce sites are:

1. Web Data Extraction Process

Web data extraction is a web content mining technique used for extracting data from web into Html file.The purpose of this process is to get data from e-commerce site and copies it to the local hardisk. This process uses Web Crawler tool. User needs to define the address of the site and crawler tool extracts data from site and copies the data to the local hardisk. The result of this process is all HTML file including the directory, CSV, xml, and images.

2. HTML Data Processing

HTML Data processing is a process to extract summary data from HTML file and show the summary data to the site summary. This process consists of 4 sub processes, such as:

a. Parsing HTML file

Parsing is a process used to get piece of data which is needed to display on site summary. This research uses and improves the Jericho HTML parser to extract summary data from HTML file. Parser extracts data based on tag inputted by user. The tag that needs to be parsed will be discussed later in section 4 (analysis).

b. Storing to the relational table

The result of HTML parser is stored in the database. Database design was resulted based on data analysis in section 4. All data from two Indonesia e-commerce sites will be stored in the same tables

c. Displaying the data in site summary

Data saved in database will be displayed in site summary. Site summary displays product data from two e-commerce sites: Lazada and Rakuten. All the same products will be compared in order to provide kindy user interface.

3. RESEARCH METHOD

Research was conducted by the following steps:

1. Conduct Literature study in order to learn and understand web mining type specifically web data extraction, web crawling, and HTML parsing.

2. Explore and analyze several Indonesia e-commerce sites as a case study in this research. The Indonesia e-commerce site analyzed related to the web structure, the important data in the web, and important HTML tag related to the data. Based on this analysis, table in database will be designed.

3. Conduct questionnaire to 30 responden about the important information in e-commerce site should know by the customer before they are buying the product.

4. Explore and analyze several Web Crawler tools and HTML parser tools. Some tools are provided in the form of libraries such as Jericho, JTidy, JavaCC HTML Parser.The result was used as practical consideration in the selection of the tools that will be used in this research.

5. Develop an application for parsing HTML file, store data in the database and visualize the data summary from two e-commerce sites in one site summary.

6. Conclude the result of the application in order to summarize the research question and to answer the problem

4. ANALYSIS

In this section, the result of analysis and result finding are explained.

4.1 Tag Structure Analysis: Lazada and Rakuten

As a study case for this research, the researchers analyzed some Indonesian e-commerce sites.The researchers tried to found Indonesia e-commerce site that sell most similar product category. Based on the analysis, Lazada and Rakuten have some similarities in the web structure and the product category, they are:

1. Web Structure:

    a. Header has a logo and e-commerce site name
    b. Product category are placed in sidebar
    c. Content with advertisement or products sold to the customer
    d. Footer has information about e-commerce site and payment method.

2. Product Category

    a. Lazada and Rakuten have many similar product categories. Based on the analysis, Rakuten and Lazada have only 10% different product category such as: bag only sold in Lazada, while flower, food & drink, books & stationery only sold in Rakuten.

As is illustrated in Figure 2, data extraction process with web crawler tool produced HTML file. HTML file stores product data in HTML tag. Analysis on tag structure of HTML was carried out in order to define all tags needed to be parsed. Those tags store common information related to the product shown in one e-commerce site. Based on the questionnaire to the 30 responden, the important information in e-commerce site should know by the customer before they are buying the product are product name, price, picture, description, product category/subcategory, and product specification. Therefore, the researchers explored this information in Lazada and Rakuten site to get the tag structure. Figure 3 shows the web structure of Lazada and Rakuten related to the position of product information.

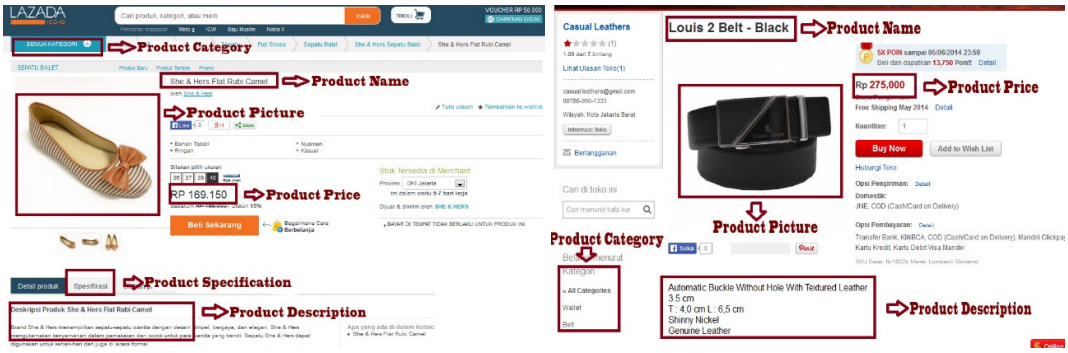

Figure 3 Web structure of Lazada and Rakuten site

All product information in figure 3 was stored in a HTML tag. The specific product information will be related to the same tag for each product. We used inspect element tool in the browser to get tag name that use in the specific information. Figure 4 and figure 5 shows the example of tag that store product information in Lazada and Rakuten site.

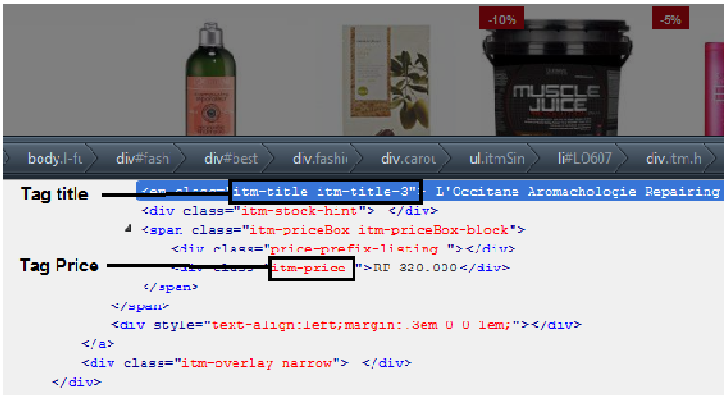

Figure 4 Example tag in Lazada page to store title and price of product

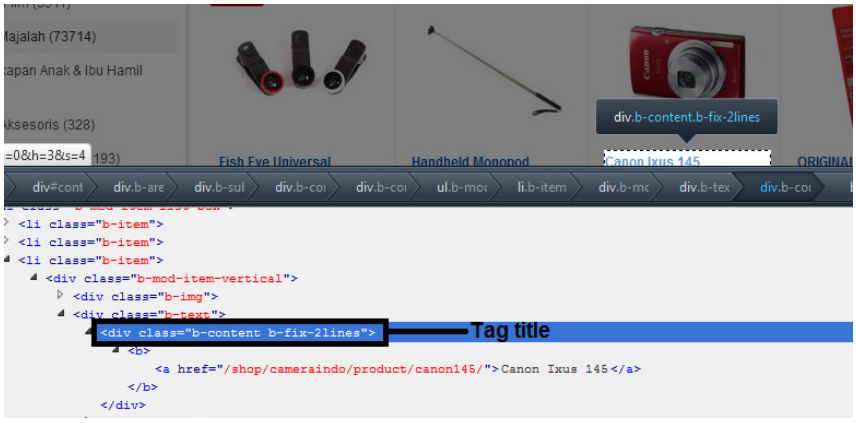

Figure 5 Example tag in Rakuten page to store title of product

The result of tag structure analysis in Lazada and Rakute site can be seen in table 1. Table 1 shows Lazada and Rakuten have different tag to store information about product. It depends on how the information placed in the site and what information will be displayed to the customer. Every e-commerce site is free to define their tag name itself according to their developer. If we

already know the tag name then it is easier to parse all the information from the site based on the tag that already defined.

Table 1 Tag structure for product information in Lazada and Rakuten

No	*Tag Name*	Description	Lazada	Rakuten
1	*Tag Box*	Space for information about product, price and picture		
2	Name	Store the name of product		<h1 class="b-content b-fix-2lines" itemprop="name"></h1>
3	*Price*	Store information of product		
4	*Picture*	Store picture of product	<img class="b-thumb-128px "></img>	<img alt="Tops C048 0" src="itm-img "></img>
5	*Breadcumbs*	Store the information about category, subcategory of product	<ul class="b-breadcrumb">	<ul class="bcr box breadcrumbs">
6	*Description*	Store the description of product		
7	*Spesification*	Store the specification of product		-

The tag structure analysis also resulted database design based on product information in the e-commerce site. The data were recorded are Product Name, Price, URL, Image, description, specification, and category. Commonly, each category has many products and it is possible to have sub category. Example: Fashion category has sub category man and woman. Sub category has sub sub category dress, bag and soon. So, it is possible one category has many sub category. The database design should supported this case so the semantic of data displayed on e-commerce site same with data were stored in database. Figure 6 shows Physical data model to store e-commerce site data.

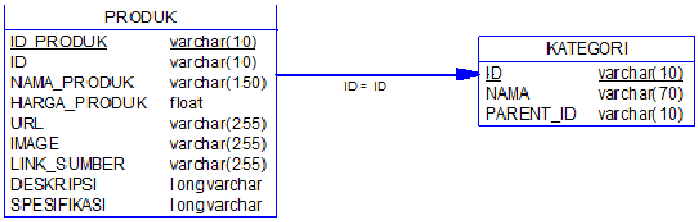

Figure 6 Physical data model design for Lazada and Rakuten

4.2 Analysis Web Crawling

In this research, data processing did in the local hardisk where the data will be available offline. It means e-commerce data were not directly retrieve from e-commerce site, but data taken from e-commerces site and copying some or all content of website to the local disk. This process called web crawling. There are several tools or web crawler that can be used, such as Visual Web Spider, Visual Web Ripper, Win Web Crawler and HTTrack. We tried these four tools and the result of comparison can be seen in table 2.

Table 2 Web Crawler comparison

No	Description	HTTrack	Visual Web Ripper	Visual Web Spider	Win Web Crawler
1	Input	URL	URL	URL	URL
2	Output	Folder dan file	File .xls or mysql or oracle	File .xls or .HTML file or mysql	Link
3	License	Open source	proprietary	proprietary	proprietary
4	Type	Free	Trial	Trial	Trial

Based on the exploration on web crawler tools, HTTrack used as a web crawler for some following reasons:

a. HTTrack is an open sorce tool, while other tools are not open source. We do not have limitation to use this tool.

b. HTTrack generates output files and folders that can be saved in local directory become an offline data.

c. HTTrack can be used continuously depends on requirement of this research. Other tools must be bought if we want to use it more than trial period.

The exploration result using web crawling tools HTTrack are:

a. Produces folder contains image file, .txt, css file, and an index.html, hts-log.txt. cookies.xt, backblue.gif and fade.gif

b. The root directory is a folder contains information about product, several Html files named similar with the product name.

c. There are image folders save icon and shortcut for each Html file in the root directory.

d. Generates index.html file in the website main folder.

e. Each folder which has similar name with product name has one index.html file and others folders.

The structure directory of crawling result by using HTTrack can be seen in Figure 7.

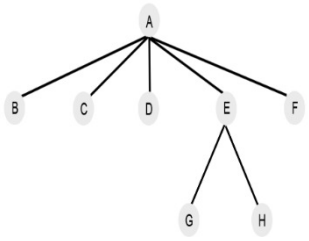

Figure 7 Crawling Result using HTTrack

Description:

A. Root folder (ex: www.lazada.co.id)

B. File (index.html, backblue, cookie, fade, hts-log)

C. Image folder

D. Configuration folder (hts-cache)

E. Folder (ajax.googleapis.com, apis.google.com, assets.pinterestcom)

F. Category and subcategory folder

G. File (index and product details)

Therefore, the researchers conclude that crawling result of HTTrack would be parsed and stored to the database. Parsing process should considered the structure of directory especially content of index.html that exist in product directory.

4.3 Analysis HTML Parser: Jericho Html Parser

Jericho HTML Parser is one example of HTML Parser which consist of Java library used to analyze and manipulate html tag. Jericho parsed tag that store product information. It validated the product information tag so the researchers did not need to take all data on e-commerce page. Jericho HTML parser already has some libraries such as:

a. FindSpecificTags.java, this class search tag with specific name in certain namespace or special tag like document declaration, XML declaration, PHP tag, Mason tag, and HTML comment.

b. ExtractText.java, class to extract all text in document, title, description, keyword, and link.

c. RenderToText.java, class to render simple text from HTML file.

In addition to the java library in Jericho, the researchers need to improve or add other classes for parsing lazada and rakuten site. Jericho parsed product information tag that already defined and stored data to the database. Some methods which are needed to be add to the Jericho HTML parser are:

a. getItem, method used to get product data such as product name, price, picture, breadcumbs, specification, description.

b. recursiveTraversal, method used to read all index file with .html extension in the all directory.

c. substituteKey, method used to maintaining category id and product id in the database.

d. generateId, method used to create category id with format “KATG000001” and product id “PROD000001”.

e. placeFileImage, method used to copy image from source file into destination folder.

f. readCharFromFile, method used to read content of html file from recursiveTraversal method.

The example algorithm for method that must be added to the Jericho can be seen in algorithm 1 below:

```
arguments: file: String value, prod: reference to a DataBefParser
BEGIN
       IF files ditemukan
              content <- readCharFromFile(files)
              source <- new Source(content)
              if getFirstElementByClass(parameter: getBreadcumbs prod) source is
       not nil
                     net.htmlparser.jericho.Source    sou    <-    allocate    new
       Source(parameter:  getFirstElementByClass(parameter:  getBreadcumbs  prod)
       source)
                     Bread : List
                     Bread <- getAllElements(“A”) sou
                     Clear lscat
                     for i <- 0 to size Bread do
                            add elementString in Bread into lscat
                     endfor
                     if size lscat > 0
                            remove value of last index from lscat
              endif
              if getAllElementsByClass(parameter: getBox prod) source is not nil
       then
              for Element element : list1

                     IF AllElementClass is not nil
                            IF elementTitle is not nil
                                   namaProduk <- element list
                            IF elementPrice is not nil
                                   price <- element list
                            IF elementImage is not nil
                                   image <- element list
                            IF elementDesripction is not nil
                                   desc <- element list
                            IF elementSpesification is not nil
                                   spec <- element list
                            sourceSite <- getAttributeValueFromTag(“A”)
                            saveToDatabase(parameter:  namaProduk,  price,  image,
               desc, spec)
                     ENDIF
              Endfor
              Endif
       ENDIF
       END
```

5. RESULTS AND DISCUSSION

The proposed method in section 2 was implemented. Implementation was started with crawling lazada and rakuten site using HTTrack. HTTrack crawled 12503 files, 22013 directory of Lazada site and 7368 files, 10747 directory of Rakuten site. Crawling process ran in 2 hours and did not crawled all data in Lazada or Rakuten site.

The crawling result was parsed by the Java application that implementing Jericho HTML parser that already improved. User must entry base path (directory location), css tag of product name, price, image, breadcumbs, specification, and description to the text that provided by application. Then, application read and parsed index.html file in root directory (root folder with name www.lazada.co.id for Lazada and www.rakuten.co.id for Rakuten) in accordance with tag that inputted by user. Content of css tag were fetched and stored in the database. Screenshot of java application to parse e-commerce site can be seen in figure 8.

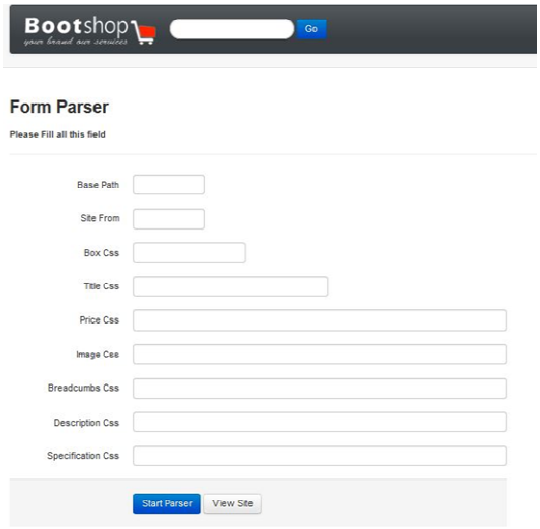

Figure 8 Java Application that parsed Lazada and Rakuten index.html file

The application parsed 5197 files with size 1.02 GB from Lazada and 3158 files with size 332 MB successfully. Parsing result for one product have data structure as shown in figure 6 and stored in database. Finally, table has 2892 records lazada product and 15517 records Rakuten product. All data can be displayed in site summary to help customer for buying same product from Lazada or Rakuten. The example of site summary can be seen in figure 9.

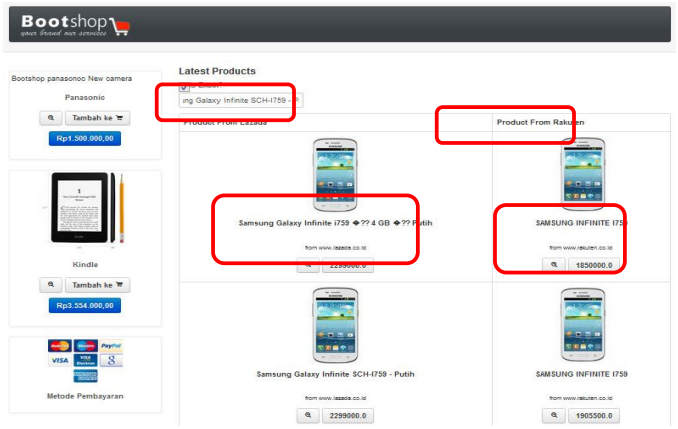

Figure 9 Prototype of site summary for Lazada and Rakuten

Figure 9 show proposed method for mining Lazada and Rakuten site with web data extraction process can be implemented successfully. Customer searched and compared products that he/she need by visiting site summary without visit lazada and rakuten site.

Based on result of implementation, there are some evaluations related to the result displayed in site summary. The evaluations are about number of file compared with number of record in database and naming or category of each product in lazada and rakute site. The explanation related to the evaluation are:

1. Number of files that successfully parsed and number record in database were not equal with number of files that already crawled by HTTrack. Based on analysis it was occured because:

a. Lazada and Rakuten has some different tags image, design, and naming convention for list of products. If the design different, then css attribute must be different. The application only defined one tag for list of products, so content of other different tags were not able to be parsed by the application. This resulted number of product in database must be less than number of product in lazada site.

b. There are 5 index.html files were not contain with list products so defined tag css in the parsing form couldn't be found in the file.

2. There are several different name of product categories for same product between Lazada and Rakuten. Example: Product category in Lazada site is Handphone & Gadget. This product category was not found in Rakuten site, but the category is Smartphone & Gadget. The application was not able to compare this product as a same product category.

6. CONCLUSIONS

The created application was able to do web content mining for Lazada and Rakuten site. The user enables to use the tool and ensure it will display the product based on user needs. User can compare every product data in Lazada and Rakuten only in one site summary. But, there are some limitations of application related to the different css tag and product category. Therefore, the expected future development is to mine all Indonesia e-commerce site with define all the related css tag that exists in e-commerce site. The proposed method also can improve with implementing natural language processing or ontology to identify product data in e-commerce site. Data mining method also can be used to improve the accuracy.

The result of this research acts as initial research for future large application development which is able to mine all Indonesia e-commerce site like google shopping.

REFERENCES

[1] Kenneth C.L, Carol G.T. E-Commerce: Business, Technology, Society. Third Edition. New York. Prentice Hall. 2007: 10.
[2] Sato S, Asahi Y. A Daily-level Purchasing Model at an E-commerce Site. International Journal of Electrical and Computer Engineering (IJECE). 2012; 2(6): 831-839.
[3] Zhang F.L, Yang W.S, Zhang W.M, E-commerce Website Recommender System Based on Dissimilarity and Association Rule. TELKOMNIKA Indonesian Journal of Electrical Engineering. 2014; 12(1): 353-360.
[4] Xu H, Zhang R, Lin C, Gan W. Construction of E-commerce Recommendation System Based on Semantic Annotation of Ontology and User Preference. TELKOMNIKA Indonesian Journal of Electrical Engineering. 2014; 12(3): 2028-2035.
[5] Lukman E. Google survey finds that half of Indonesia's ecommerce non-adopters will shop online soon. https://www.techinasia.com/google-survey-finds-indonesias-ecommerce-nonadopters-shop-online/. 2 April 2014.
[6] www.google.com/shopping, google shopping, August 2014.
[7] Arvind KS, Gupta PC. Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining. IJARCET Journal. 2012; 1(8): 287-293.
[8] Achmad S. Web Usage Mining dengan Google Analytics: Studi Kasus Situs Achmatim.Net. Universitas Indonesia. 2008.
[9] Nimgaonkar S, Duppala S. A Survey on Web Content Mining and extraction of Structured and Semi structured data. IJCA Journal. 2012; 47(11): 44-50.
[10] Jaideep S, Robert C, Mukund D, Pang-Ning T. Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations. 2000; 1(2): 12-23.
[11] Addanki R, Konda S, P RK. Preprocessing and Unsupervised Approach for Web Usage Mining. International Journal of Social Networking and Virtual Communities. 2012; 1(2)

[12] Khirade RR, Balaji S. A survey: Web mining via Tag and Value. International Journal of Computer Science and Information Technologies. 2014; 5(2): 5310-5314.
[13] Muhammad F. Intelligent Crawling of Web Applications for Web Archiving. Proceedings of WWW '12 Companion International conference companion on World Wide Web. Lyon. 2012; 127-131.

AUTHORS

Humasak Tommy Argo Simanjuntak got a Master of Information Systems Development in HAN University of Applied Sciences in The Netherlands. At the moment he is head of and lecturer at the Information System Study Program at Del Institute of Technology in Indonesia. His research interests are mainly about information systems development, data management, data mining, and business intelligence.

Novitasari Sibarani, Novalina Hutabarat, and Bambang Sinaga was graduated from Del Institute of Technology on 2014. They was involved in this project to do initial step for mining Indonesia e-commerce site.