IT FOR SOCIETY, Vol. 08, No. 01
ISSN 2503-2224

Development of Marketplace Product Data Mining and Analysis Application as a Reference in Running a Business

Joseph TN Wibawa[1], Randy Gunawan[2], Marchelleo Suhandi[3], Ricky Setiawan[4]
Faculty of Computing
President University
Bekasi Indonesia
randy.gunawan@student.president.ac.id

***Abstract* - Marketplaces has become the first choice as a third party that provides online selling and payment services that bridge sellers and buyers. The increasing need to shop in an efficient and convenient way has driven the popularity of using the marketplace. Moreover, with a variety of products and a fast transaction process, and more competitive prices, the use of the marketplace has shown significant growth in activity by the wider community. With the increasing use of marketplaces, data analysis from marketplaces is becoming a key element to understand market trends and make informed transaction decisions. Data analysis systems based on bot technology and Web-Scrapping libraries, such as Puppeteer, allow users to perform data analysis more efficiently and quickly, minimize the need for a data analysis team, and reduce time and costs. Given these factors, this research focuses on developing a web-based application that can analyze data coming from marketplaces and present it to users quickly and accurately.**

Keywords - Marketplace, Scrapping, Web-Scrapping

I. INTRODUCTION

Nowadays, there are many marketplaces for various types of products targeting various sectors of society. The marketplace itself is a third-party site that is packaged online with a role to bridge the buying and selling transaction process between buyers and sellers. The marketplace is useful as a more efficient alternative to going shopping. Marketplaces are becoming a more popular choice due to the high likelihood of users finding the right item quickly and at a lower price. This is because users do not buy through shopping malls that have to consider building management costs such as electricity procurement, building rental fees, and building maintenance. The current trend shows users' desire to shop in an easy, fast, and efficient way, causing users of marketplaces to grow exponentially. The probability of an increase in product buying and selling activities at marketplaces is increasing along with the increasing provision of information about products and the suitability of product prices to market prices.

Because the increase in users in the marketplace is accelerating, it is necessary to convert the existing system in the conventional market to the market. Locapasar data analysis is a future innovation that will shift the data analysis system in conventional markets. This is because the data analysis system in conventional markets requires a research team to come to market locations, making the system inefficient and taking a long time to analyze the data obtained, while the market data analysis system will utilize the use of bots to open various market pages where it will make it easier for users to get accurate and fast market product analysis results. Some of the properties of the marketplace product analysis described above make this system an alternative solution for market analysis that can minimize the need for an analysis team and the costs incurred to analyze data.

The use of Web-Scrapping libraries such as Puppeteer is potentially the key to modernizing future market locator analysis systems as it minimizes the need to design a mining scheme from scratch which is usually quite time-consuming. There is a need to quickly present market analysis data that matches the products that each user wants to research. The main objective of this research is to develop an application that can facilitate and present market location analysis data to users quickly and accurately. So, the focus of this research is to provide an application where users can analyze data derived from the marketplace easily and quickly using only a web-based application.

Some of the objectives of this research can be described in several points, among others:

1. Define and design a marketplace product data analysis application that uses libraries to shorten development time;
2. Developing a web application to analyze local market product data to improve scalability and

efficiency using the Laravel framework and PHP programming language; and

3. Evaluate the data generated by the web application for analyzing market location product data in terms of data quality and services provided to users.

II. LITERATURE REVIEW

Current research related to web mining can be divided into three, namely: web content mining, web structure mining, and web usage mining[1].

A. Web Content Mining

Web content mining is the process of extracting useful information from data sets derived from web documents consisting of various data types such as text data, images, audio or video data, records such as lists or tables, and hyperlinks [2]. Two approaches used in web content mining are an agent-based approach consisting of 3 agents namely an Intelligent search agent, an Information filtering/Categorizing agent, a Personalized web agent [3], and a database approach consisting of a database. The database is well organized and contains schemas and attributes with specified domains [4].

Web Content mining has the following techniques for mining data: unstructured mining, structured mining, semi-structured mining, and multimedia mining. [5]

1. Unstructured Data Mining

   This technique can be applied to text documents because text documents are a form of unstructured data. Some of the techniques used in unstructured data mining are Information Extraction, Topic Tracking, Summarization, Categorization, Clustering, and Information Visualization. [5]

2. Structured Data Mining

   Structured data mining techniques are used to extract structured data from web pages [6]. The data can be in the form of lists, tables, and trees. The data is structured data which is easier to extract than unstructured data [7]. The techniques used to mine structured data are Web Crawler, Wrapper Generation, and Page Content Mining. [5]

3. Semi-structured Data Mining

   Semi-structured data arises when the source does not impose a strict structure on the data. If we want to extract data from a web page and put the data into a database [8]. The techniques used for semi-structured data mining are Object Exchange Model (OEM), Top Down Extraction, and Web Data Extraction language [5].

4. Multimedia Data Mining

   Multimedia Data Mining is the process of discovering interesting patterns from media data such as video, audio, text, and images that cannot be accessed using queries [8]. Some of the Multimedia Data Mining Techniques are SKICAT, color Histogram Matching, Multimedia Miner, and Shot Boundary Detection [5]

B. Web Structure Mining

Referring to the use of the hyperlink structure of the web as an information source [9], it can be concluded that if the data collection process only uses traditional data collection methods and assumes that the data is independent, it can lead to wrong conclusions [10]. There are two algorithms to lead these correlations, namely HITS (Hyperlink-Induced Topic Search) developed by Jon Kleinberg [11] and Page Rank [12], [13]. These two algorithms are used to find the importance of a web page [14] which can be described as follows:

1. HITS Algorithm

   HITS ranks web pages by analyzing the inner links and outer links of the page [13]. In this algorithm, web pages are pointed by many hyperlinks called authorities whereas web pages pointing to many hyperlinks are called hubs [11], [14], [15]. Authorities and hubs are illustrated in Figure 2.1.

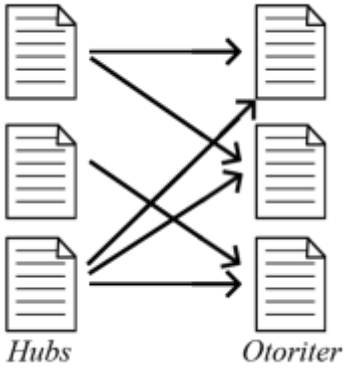


Figure 2.1 Hubs and Authorities

2. Page Rank Algorithm

   The Page Rank algorithm states that if a page has important links pointing to it, then

the links from that page are also important [13]. Therefore, Page Rank takes backward links into account and propagates ranking through links: a page ranks high if the number of links from its backward links is high [12]. Figure 2.2 shows an example of backward linking: Page A is a backward link from page B and page C while page B and page C are backward links from page D.

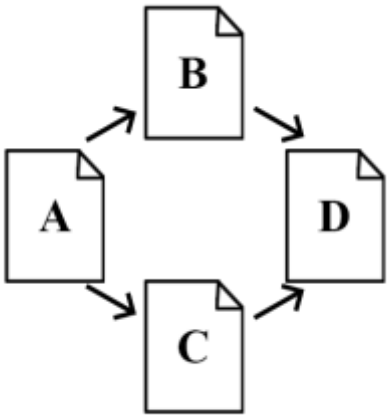


Figure 2.2 Backlink Examples

The challenge of web structure mining is having to deal with the structure of hyperlinks on the web and the age of the sector also referred to as Link Analysis is very old [10]. However, with the growing interest in web mining, research from the Web Structure Mining sector continued to grow and resulted in a research sector known as Link Mining [16]. This sector lies at the intersection of link analysis, hypertext and web mining, relation science, and inductive programming logic [10].

C. Web Usage Mining

Adapted from several sources, Web usage mining is the application of data mining techniques to find patterns of data usage originating from a particular web to understand and serve the needs of web-based applications to increase the efficiency of these applications. Web usage mining itself consists of 3 phases, namely preprocessing, pattern discovery, and pattern analysis [17].

Preprocessing consists of processing the usage, content, and structure of the information contained in various available data sources into the data abstractions required to find suitable patterns [17].

The description of the stages contained in this first stage is as follows:

1. Usage Pre-processing

   This stage can be said to be the most difficult process in the series of Web Usage Mining (WUM) processes because in this process there can be incomplete data available. However, this can be avoided if a client-side tracking mechanism is used, only IP addresses, agents, and server-side clickstreams are available to identify users.

2. Content Pre-processing

   Content pre-processing consists of several activities, namely the conversion of text, images, scripts, and other files such as multimedia into a data form useful for the WUM process. It often consists of content mining activities such as classification or clustering.

3. Pre-Processing Information structure

   The structure of a site can be created by the links between page views. This structure can be obtained and processed in the same way as the previous stage of content pre-processing. But keep in mind that dynamic content can cause more problems than static content, so to minimize the problems that will arise, you can build a site structure for each server session.

Pattern Discovery refers to methods and algorithms developed in several fields such as statistics, data mining, pattern recognition, and machine learning [17]. However, this section will not explain all the algorithms and techniques available from these fields so this section only describes the types of mining activities that have been applied to the web domain including the following:

1. Statistical Analysis

   Statistical techniques are the most commonly used methods to extract information regarding Web site visitors. By analyzing session files, one can perform various types of descriptive statistical analysis (frequency, mean, median, etc.) on variables such as page views, view time, and navigation path length. Many Web traffic analysis tools can provide periodic reports on statistical information such as the most frequently accessed pages, the average viewing time of a page, or the average length of the path taken by a site. These reports may include limited low-level error analysis such as detecting unauthorized entry points or finding invalid URLs in general.

Although less in-depth in its analysis, this type of information is potentially useful for improving system performance, enhancing system security, facilitating site modification tasks, and providing support for marketing decisions.

2. Association Rules

Association rule generation can be used to link the most frequently referenced pages together in a single server session. In the context of Web Usage Mining (WUM), an association rule refers to a collection of pages accessed together with a support value that exceeds a certain threshold. These pages may not be directly connected to each other through hyperlinks. For example, the discovery of association rules using the Apriori algorithm [18] (or one of its variants) can uncover correlations between users who visit pages containing electronic products and those who access pages about sports equipment. Besides being usable for business and marketing applications, the existence of such rules can help Web designers to restructure their Web sites. Association rules can also serve as heuristics for prefetching documents to reduce the latency users feel when loading pages from remote sites.

3. Clustering

Clustering is a technique to group a set of items that share similar characteristics [17]. In the Web Usage domain, there are two interesting types of clusters to be found: usage clusters and page clusters. User clustering tends to form groups of users who exhibit similar browsing patterns. Such information is very useful for inferring user demographics to perform market segmentation in E-commerce applications or provide personalized Web content to users. On the other hand, page clustering will find groups of pages that have related content. This information is useful for Internet search engines and Web help providers. In both applications, permanent or dynamic HTML pages can be created that suggest related hyperlinks to the user according to the user's request or previous history of information needs [17].

4. Classification

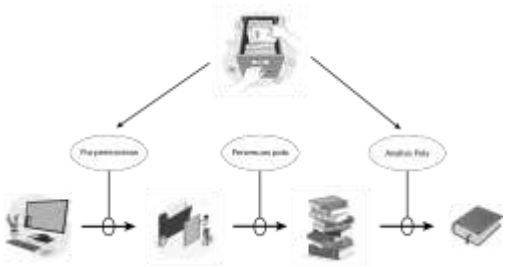
Classification is a method to map data items into one of several predefined classes [19]. In the Web domain, one is interested in developing a profile of users belonging to a particular class or category. This requires the extraction and selection of features that best describe the properties of a particular class or category. Classification can be done using supervised inductive learning algorithms such as decision tree classifiers, naive Bayesian classifiers, nearest neighbor classifiers, Support Vector Machines, etc.

5. Sequential Patterns

The sequential pattern discovery method is one of the methods used to try to find inter-session patterns that are formed in such a way that the presence of one set of items is followed by another item in a set of sessions or episodes sorted by time. Using this approach, Web marketers can predict future visitation patterns which will help place ads aimed at specific user groups.

6. Dependency Modeling

Dependency modeling is a useful pattern discovery method in Web Mining. The goal of this method is to develop models capable of representing significant dependencies among various variables in the Web domain. Several probabilistic learning techniques can be used to model user browsing behavior. They include Hidden Markov Models and Bayesian Belief Networks. This modeling will not only provide a theoretical framework for analyzing user behavior but can also potentially predict future Web resource consumption.



Gambar 2.3 Proses WUM

Dependency modeling is a useful pattern discovery method in Web Mining. The goal of this method is to develop models capable of representing significant dependencies among various variables in the Web domain. Several probabilistic learning techniques can be used to model user browsing behavior. They include Hidden Markov Models and Bayesian Belief Networks. This modeling will not only provide a theoretical framework for analyzing user behavior but can also potentially predict future Web resource consumption.

III. RESEARCH METHODOLOGY

The method chosen for this research aims to increase efficiency and make it easier for the general public to analyze data for various purposes. The methodology design starts with a literature review of

the most efficient data scraping techniques by considering various factors that can affect the time taken in collecting data from various marketplaces available in Indonesia. This then led to the formation of hypotheses using certain factors to measure their feasibility. The hypothesized most efficient data scraping technique was validated through simulation and evaluation. The data collected from the simulation is the time required to collect and process data between humans and the web-scraping application that has been created. The obtained results were then compared and verified with existing solutions to determine the improvement in performance efficiency.

An important consideration in the evaluation of future web-scraping applications lies in the data-scraping technique itself and the features provided by the application. A key component to designing an optimal web-scraping application is a comprehensive knowledge and understanding of the influencing factors and different parameters that the application is intended for.

Known critical parameters for creating an efficient web-scraping application quantitatively expressed in values (such as the speed of data collection, the amount of data that can be calculated, and the accuracy of the results) are investigated. The purpose of the evaluation is to improve the efficiency in processing scalable market data to reduce the time required to process such data by utilizing various scraping and data analysis techniques.

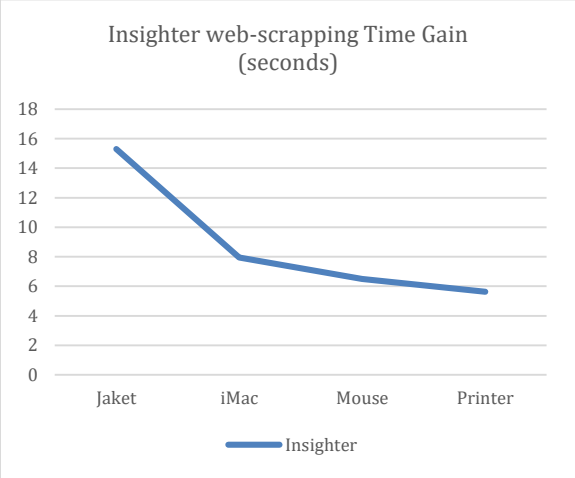
IV. RESULTS AND DISCUSSIONS

The expected outcome of this study is a web-scrapping application that can capture data, process it, and produce results quickly and consistently. This contribution will result in the implementation of web-scrapping which has an impact on improving the performance of market analysis in various types of products and types of information. Realizing future web-scrapping in market analysis is expected to improve conventional market analysis methods that operate using human labor. With a significant increase in efficiency and effectiveness in the field of market analysis, making market analysis more accessible to all levels of society without knowledge of data analysis.

The purpose of the experiment is to investigate the performance of the web-scrapping application in (i) time utilization. (ii) tracking of different types of products. (iii) provision of accurate analysis results compared to human analysis results with various samples of analyzers. Some of the sample analyzers are individuals who have been accustomed to performing data analysis using tools owned by the individual and individuals who have not been accustomed to performing data analysis using unfamiliar tools.

Query	Insighter	Human
Jacket	00.15.29	06.51.96
iMac	00.07.94	13.39.56
Mouse	00.06.50	15.49.20
Printer	00.05.62	06.43.59

Tabel 2.1 Time acquisition of web-scrapping Insighter and analysers



Graph 3.1. Time performance of Insighter web-scrapping in analyzing data of various products

The time required to record data can measure the efficiency of each method and the results of data recording can assess the ability of each method to analyze data based on the conclusions drawn by the perpetrators of the analysis of previously obtained data. Figure 3.1 shows how the time acquisition of the web-scrapping application differs. Figure 3.2 shows the acquisition of data analyzers. Figure 3.3 shows the achievement of statistical elements obtained by the web-scrapping application compared to humans.

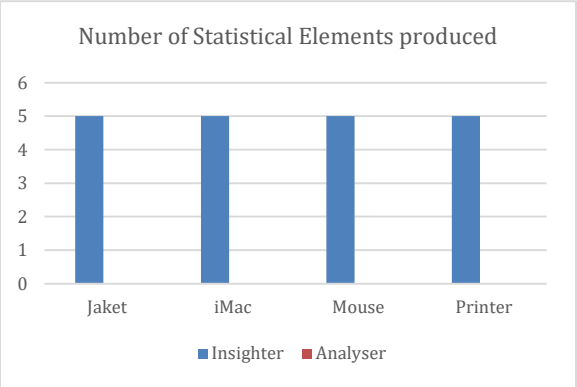
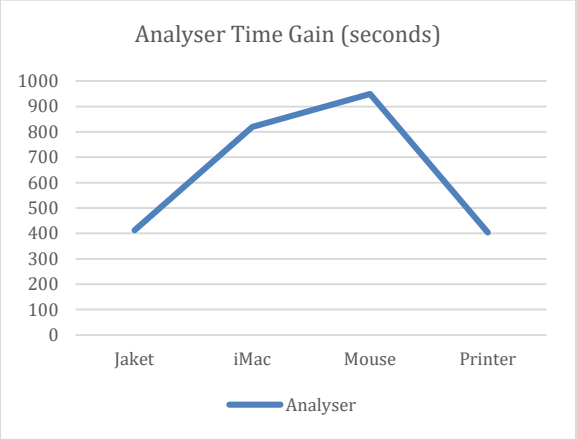


Figure 3.3. The number of Statistical Elements generated at the end of data analysis using web-scrapping and Conventional Analysis by Analysts

The time performance using the web-scrapping method shows a decrease as more analysis is performed. In contrast, the time performance using



Graph 3.2. Time performance of Analyst (Human) in analyzing data of various products

the conventional method fluctuates without following any pattern as it depends on the analyzer. In addition to getting superior time gains, web-scrapping also consistently provided an above-average number of statistical elements across all experiments with analyzers.

V. CONCLUSIONS AND FUTURE WORKS

The use of web-scrapping applications in market data analysis has the potential to improve efficiency and accessibility for the wider community. Through literature review, simulation, and evaluation, the web-scrapping method is proven to be more efficient in collecting and processing market data compared to conventional methods that rely on human labor. The web-scrapping application can generate data quickly, consistently, and accurately, and provides a greater number of statistical elements than conventional methods. With proper implementation, web-scrapping applications can significantly improve the performance of market analysis and enable easier accessibility for all levels of society without specialized knowledge of data analysis.

ACKNOWLEDGEMENT

The authors are very grateful to the Faculty of Computing of President University, President University, and Mr. Hasanul Fahmi of President University for being willing to direct the writing of articles in the Indonesian language course.

REFERENCES

[1] J. Jackson, “Data Mining; A Conceptual Overview,” *Communications of the Association for Information Systems*, vol. 8, 2002, doi: 10.17705/1cais.00819.

[2] C. E. Dinuca and D. Ciobanu, “WEB CONTENT MINING,” *Annals of the University of Petroşani, Economics*, vol. 12, no. 1, pp. 85–92, 2012, Accessed: Jul. 20, 2023. [Online]. Available: https://econpapers.repec.org/article/petannals/v_3a12_3ay_3a2012_3ai_3a1_3ap_3a85-92.htm

[3] S. Prakasam, “An agent-based Intelligent System to enhance E-Learning through Mining Techniques,” *IJCSE) International Journal on Computer Science and Engineering*, vol. 02, no. 03, pp. 759–763, 2010, Accessed: Jul. 20, 2023. [Online]. Available: https://www.researchgate.net/publication/49618603_An_Agent-based_Intelligent_System_to_Enhance_E-learning_through_Mining_Techniques

[4] M. H. Dunham, *Data Mining*. New Jearsey: Pearson Education, 2003. Accessed: Jul. 20, 2023. [Online]. Available: https://theswissbay.ch/pdf/Gentoomen%20Library/Data%20Mining/Dunham%20-%20Data%20Mining.pdf

[5] S. Saini and H. M. Pandey, “Review on Web Content Mining Techniques General Terms Data Mining, Web Content Mining. Keywords Web content mining, structured data mining, unstructured data mining, semi-structured data mining,” *Int J Comput Appl*, vol. 118, no. 18, pp. 975–8887, 2015, doi: 10.5120/20848-3536.

[6] K. Pol, N. Patil, S. Patankar, and C. Das, “A Survey on Web Content Mining and Extraction of Structured and Semistructured Data,” in *2008 First International Conference on Emerging Trends in Engineering and Technology*, IEEE, 2008, pp. 543–546. doi: 10.1109/ICETET.2008.251.

[7] F. Johnson and S. Kumar Gupta, “Web Content Mining Techniques: A Survey,” Jun. 2012. doi: 10.5120/7236-0266.

[8] D. Sharda, "WEB CONTENT MINING TECHNIQUES : A STUDY," Chandigarh, 2014. Accessed: Jul. 20, 2023. [Online]. Available: http://ijirts.org/volume2issue3/IJIRTSV2I3050.pdf

[9] J. Fürnkranz, "Web Structure Mining Exploiting the Graph Structure of the World-Wide Web," Austria, Aug. 2002. Accessed: Jul. 20, 2023. [Online]. Available: https://www.researchgate.net/publication/294218073_Web_structure_mining_Exploiting_the_graph_structure_of_the_world-wide_web

[10] S. B. Boddu, A. V. P. Krishna, R. R. Kurra, and D. K. Mishra, "Knowledge discovery and retrieval on World Wide Web using Web structure mining," *AMS2010: Asia Modelling Symposium 2010 - 4th International Conference on Mathematical Modelling and Computer Simulation*, pp. 532–537, 2010, doi: 10.1109/AMS.2010.108.

[11] J. M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment *," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, Sep. 1997, doi: 10.1145/324133.324140.

[12] Lawrence Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," 1998. Accessed: Jul. 20, 2023. [Online]. Available: https://www.semanticscholar.org/paper/The-PageRank-Citation-Ranking-%3A-Bringing-Order-to-Page-Brin/eb82d3035849cd23578096462ba419b53198a556

[13] W. Xing and A. Ghorbani, "Weighted PageRank Algorithm," Second Annual Conference on Communication Networks and Services Research, May 2004. doi: 10.1109/DNSR.2004.1344743.

[14] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, "PageRank, HITS and a Unified Framework for Link Analysis *," Aug. 2002. doi: 10.1145/564376.564440.

[15] J. Wang, Z. Chen, L. Tao, W.-Y. Ma, and L. Wenyin, "Ranking User's Relevance to a Topic through Link Analysis on Web Logs," 2002. doi: 10.1145/584931.584942.

[16] L. Getoor, "Link Mining: A New Data Mining Challenge," Maryland, 2003. doi: 10.1145/959242.959253.

[17] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," Minneapolis, Jan. 2000. doi: 10.1145/846183.846188.

[18] R. Agrawal and R. S&ant, "Fast Algorithms for Mining Association Rules."

[19] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases) (© AAAI)," 1996. [Online]. Available: www.ffly.com/