Abstract—The development of e-commerce in Indonesia is increasing. Now this situation makes e-commerce as the land of very large data. The data is not used as a content in e-commerce only. Moreover, by the existence of Web Mining, the data are processed in a such way that will generate new information to be utilized in development system. One of usage of Web Mining is Web Usage Mining which is one technique to analyze user activity pattern in a website recorded in server log. The data that can be used for Web Usage Mining is server log data containing session information from users and session query logs containing user query session information. In this paper, the paper will utilize session query log data (external web usage mining). By using density-based Clustering method, this paper implemented a development algorithm from OPTICS (Ordering Points to Identify the Clustering Structure) algorithm, namely AD-OPTICS (Adaptive Ordering Points to Identify the Clustering Structure). The input of the system in this paper is the user's log dataset obtained from a backlink checker site that is ahrefs.com. The dataset is a list of keywords, links, and other numerical data crawling the system to the number of searches performed by users on a search engine on some e-commerce. E-commerce taken backlink dataset is 5 major e-commerce in Indonesia among others are Bukalapak, Elevenia, Lazada, Tokopedia and OLX. The result of the research is in the form of groups of items that are often sought by users along with the purity value of the formation of the item group (purity cluster). From the research, the group of items that are often searched by users are 'mobil bekas', 'hard disk eksternal', and 'ponsel' with the biggest purity value of all test is 0.00565.

Keywords: Web Mining, Web Usage Mining, Density Based Clustering, OPTICS, AD-OPTICS, Query Log

I. INTRODUCTION

One of the growing web development is the rapid e-commerce development. Based on research of Online Shopping Outlook 2015 issued by BMI research revealed that online shopping value in 2014 reached 21 trillion. There are about 4.3 million people in 2013 who are accustomed to online shopping, now the number has reached 7.4 million people and will continue to grow in the coming years [1].

The search for easy goods and services on e-commerce also helps customers find all forms of products or services as needed. From a group of items that users often looked for, it can be extracted new information. Then, this new information can be utilized for further e-commerce development such as making collection of items recommendations that will be offered in e-commerce is to increase the sales, by utilizing web mining technology.

Web mining is one data mining application form to explore useful information from the web structure of hyperlinks, web pages, and web usage data [2]. Web Usage Mining is one of the applying web mining techniques utilizing web usage data in order to know a user activity pattern on a website recorded in server data logs. However, in the development of web usage mining technology can be used by using the data query log.

The query log data is derived from a user usage logs against a site that has a search engine that is also called a search query log. Search query logs are data that contains a log of search activity performed by multiple users. The search query log can be used as reference data in implementing web usage mining. As in the web usage mining implementation, it is used modeling techniques owned of data mining called Clustering. Clustering is a technique of grouping data that has the same characteristics [3].

Some research on web mining using Clustering technique has been done, for example Clustering based on data density ie with DBSCAN (A Density Based Spatial Clustering of Applications With Noise) and OPTICS (Ordering Points to Identify the Clustering Structure). Akiladevi [4] mentions that DBSCAN is a fairly effective algorithm for Clustering web
Web mining is an excavation of data sourced from the internet that aims to find useful information or knowledge of a web, i.e., from web structure hyperlinks, web pages, and web usage data [2]. Based on the type of data used in the process of extracting information from a web, web mining consists of three types [8] namely, Web Structure Mining, Web Content Mining, and Web Usage Mining.

B. Web Usage Mining

According to Mehak, web usage mining is a type of web mining that allows a user to access the information collection in a web page by analyzing the web log files [9]. A server log file contains requests made to the web server, where these requests are recorded chronologically by the web server.

Web usage mining can be classified based on the data being reviewed:

a. Web Server Data: The user log contained on the web server. It generally consists of IP address, page reference, and access time.

b. Application Server Data: Commercial application servers, such as Web logic and Story Server, have an important feature that enables e-commerce applications built on it. Its main feature is the ability to explore various business events and logs in the application server logs.

c. Application Level Data: existing events can be defined in the application and logging can be recorded to record the event.

C. Clustering

Clusters are the set of several objects that have similarities. Clustering is a process of grouping a set of data naturally into a meaningful subset called cluster [10]. The clustering method separates data groups into subset data groups that have either the same or different characteristics. It will create different clusters. Clustering is unsupervised learning in the data mining, i.e., training data is not known so that will be known the existence of data in a cluster determined by the measurement parameters [10].

Good clustering will have a high quality in terms of similarities where the similarity between objects in intra-clusters have a high similarity and inter-objects in the inter-class have a low similarity. The quality of a Clustering result depends on the size of the similarity used, the way it is implemented and measured by its ability to form a hidden pattern. However, in Clustering the problem is an objective evaluation that is usually done by human inspection [10].

There are several approaches that can be used in developing Clustering methods. Some Clustering approaches are grouped into several types of approaches with several algorithms namely, Partition Approach, Density Based Approach, Network Based Approach, Model Based Approach, and Hierarchical Approach.

D. Performance Efficiency

OPTICS (Ordering Points to Identify the Clustering Structure) is a Clustering algorithm based on data density (density based clusters). OPTICS is an algorithm that has the same principles as the DBSCAN algorithm, where the goal is to find clusters of data and noise. In DBSCAN, there are two main components that must be known: the Eps parameter (\(\epsilon\)) and the exact MinPts parameter values of each cluster and at least one point (object) of each cluster [5]. OPTICS will generate a sequence of objects considered Has the closest distance, be its neighbor in sequence. OPTICS will store the order of objects to be processed and two value information for each object i.e, core distance and reachability distance. The following definition of core distance and reachability distance [5].

\[
\text{core distance}_{\text{OPTICS}}(p) = \begin{cases} 
\text{undefined} & \text{if } \text{Card}(\text{Ne}(p)) < \text{MinPts} \\
\text{MinPts} - \text{distance}(p), \text{ lainnya} & \text{otherwise}
\end{cases}
\]  

\[
\text{Reachability distance}_{\text{OPTICS}}(p,q) = \begin{cases} 
\text{undefined} & \text{if } |\text{Ne}(q)| < \text{MinPts} \\
\max(\text{core distance}(q), \text{distance}(p)) & \text{otherwise}
\end{cases}
\]
The development of the OPTICS algorithm, one of which is the AD-OPTICS (An Adaptive Dynamical Clustering) which has the main features of the development of OPTICS algorithm. The advantages of AD-OPTICS is multi-view clustering, tuning data using data quintiles, and step in the dynamic clustering process. The existence of the development of the OPTICS algorithm into AD-OPTICS is adjusted to some problems in OPTICS [5] namely:

a. In the application of an algorithm, it is desirable that effective time for large databases. In AD-OPTICS, values of $\varepsilon_{max}$ parameters are selected in an adaptive way.

b. The way of cluster extraction in OPTICS is where the cluster cut is determined in a new way $\varepsilon$ with the quantity of reachability distance.

c. The existence of time constraints used to determine the value of $\varepsilon_{max}$, part of the clustering structure may be missing. Then the object will be used in the new round of OPTICS algorithm ie AD-OPTICS.

d. Grouping of complex objects, objects with multiple views (multiple views) where multiple metrics can be applied. AD-OPTICS uses a simple strategy that is for each iteration, a metric is selected.

E. **Web Query Log**

Web Query Log is a file that records search requests submitted by users along with results returned by search engines and results that have been clicked by the user. The query log contains a history record of user activity in search engines that exist within a web containing new knowledge [2]. This data log query is considered to be the usage data of a Web.

Query Logs if properly processed and can generate new knowledge can provide several benefits such as to improve the effectiveness and efficiency of a search engine on a web, improve filtering on spam items, and streamline ambiguous query handling. The primary data source used in Query Log Mining is the query log of a search engine on a web that can be obtained from the server side, the client side or from the Internet Service Provider (ISP) side through the proxy log. Query data can be collected through the existing access log on a web directly if the web has its own search engine or can through the log of conventional access on the web that is connected on external search engines.

F. **TF IDF (Term Frequency Inverse Document Frequency)**

Term Frequency Inverse Document Frequency is one method that can be used in assessing and measuring the attachment of a document, not just a word. With TF-IDF, words in all documents in the dataset are weighted, as measured by TF-IDF is the relevance rather than frequency. TF-IDF is a combination consisting of TF and IDF. TF specifies how much the existence of a term (word) in a document. The greater the number of occurrences of a term in the document, the greater the weight or the value of the suitability of large documents [12].

According to Defeng (in Robertson, 2004), the type of formula to be used for calculating the term frequency (tf) is pure tf (raw tf) [13]. Thus, the formula used to calculate the TF-IDF value is by multiplying the value of TF with the IDF value, which is as follows:

$$w_{ij} = tf_{ij} \times (\log (D/df_j) + 1)$$

G. **Cosine Similarity**

Cosine similarity is a measure of the similarity between two vectors in measuring the cosine of angles between two document vectors. Each vector is represented from a word frequency value that appears on each document. Cosine similarity has a semantic meaning to detect similarities between documents. As for getting the cosine value for every two pairs of documents, is as follows:

$$similarity = \cos(\theta) = \frac{\sum_{i=1}^{n} A_i B_i}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\left(\sum_{i=1}^{n} A_i^2\right)^{1/2}\left(\sum_{i=1}^{n} B_i^2\right)^{1/2}}$$

In the case of information retrieval, cosine similarity of two documents will be worth between 0 and 1, since the weighting frequency present in each document will not be negative. Two documents are said to be identical or to have a resemblance if the cosine is one (1), and otherwise the document is said to have no resemblance at all, the cosine value is no (0) [14].

H. **Purity**

Performance measurement system is done by using purity calculation method of clusters formed, namely Purity [5]. Purity is a measure of the purity accuracy of a cluster [15]. Purity calculation is:

$$purity(\Omega, C) = \frac{1}{N} \sum_{k=1}^{N} \max |w_k \cap c_j|$$

Where, $\Omega = \{w_1, w_2, ..., w_k\}$ is a set of elements in the cluster and $C = \{c_1, c_2, ..., c_j\}$ is the set of clusters. A bad cluster will have a purity value close to 0 and a good and good cluster will have a purity value approaching 1. A high purity will be easily obtained if the number of large clusters and purity will be close to 1 if all elements enter in the cluster [15].
III. MEASUREMENT

A. System Description

This system implements a data mining approach by clustering, precisely clustering based on data density. The system will process the query log data from the e-commerce website, which will then be grouped based on the AD-OPTICS algorithm. The AD-OPTICS algorithm uses the eps and MinPts values of a dataset by adapting the new way adaptively, i.e., with the quantitative range of a data. AD-OPTICS is basically the concept of OPTICS algorithm; it's just AD-OPTICS has other advantages that overcome the weakness of OPTICS algorithm that is 1) Multi-view Clustering, 2) Can be queried with Quantile data, 3) The steps include dynamic clustering.

The system design process is illustrated in Figure 1 below.

![Figure 1: System Design Process](image)

Here's an explanation of the system overview in Figure 1.

1. The first stage, the system reads the dataset in the form of search logs on search engines that have been collected from the source.
2. The system performs the preprocessing phase of the data, i.e., performing the process of data cleaning, integration, transformation, reduction, and discretization.
3. Once the data has passed through the preprocessing stage, data that has been considered good will be incorporated into the implementation phase. In the initial implementation phase, the data is divided into two parts: the data type, namely string data and numerical data. In the implementation of the data string is done TF-IDF process first, then combined with numerical data to enter the calculation of Cosine Similarity. The result of cosine similarity value is the value of similarity or similarity of each data in the dataset. The cosine similarity value will be used as the data vector of each data in the dataset which will then be used for implementing the AD-OPTICS algorithm. The implementation of AD-OPTICS algorithm refers to the basic algorithm, OPTICS. After the system implementation is complete, the next step is the test scenario implementation.
4. Furthermore after the test scenario is done, the results of the test are evaluated and analyzed. The result of the analysis is the best evaluation value in the cluster.

B. System functionality

a) System functionality

This system is useful for grouping a number of items searched by users who entered on the searching feature in e-commerce, so that it will get a group of items that are often sought by the user. Grouping the items using a data mining approach that is clustering with AD-OPTICS algorithm. Grouping results can be used as a reference for developers offering special items that are more often searched for by users. Some of the system functionality in this thesis research include:

1. The system can perform preprocessing of the dataset.
2. System can implement TF-IDF method along with Cosine Similarity value calculation.
3. The system can implement AD-OPTICS algorithm.

b) System input

Enter the system in the form of product search queries that the customer inputs into e-commerce. In this case, the e-commerce that is analyzed are 5 big e-commerce in Indonesia those are Tokopedia, OLX, Bukalapak, Lazada and Elevenia. Data used from 27th January 2016 to 19th September 2016. Then the data is entered into the system by applying the variable values used in the AD-OPTICS algorithm that is the minClSize (eps) and minsizeToCluster (minPts) variables.

c) System output

Output from the system in the form of clustering results grouping items that users often search on e-commerce. The AD-OPTICS algorithm which refers to the OPTICS algorithm is one of the non-explicit clustering methods, but removes the points in a specific order along with the smallest reachability distance value [6]. Although the OPTICS algorithm does not explicitly describe the clusters that are formed, but based on the reachability distance value obtained can be formed a graph that if formed a valley and hill, then in each valley will be considered to be a cluster [6].

As for the result, the algorithm will generate a point in a certain order along with its reachability value but at the time of cluster extraction with quantile data [11], it will not necessarily contain multiple items whose overall items have a natural direct resemblance.

C. Stages of System Design

1. Query Log
In this process the data collection from the source backlink checker platform, namely ahrefs.com. The data collected in the form of user query log data in search engines on some e-commerce in Indonesia. The selected e-commerce shops is Bukalapak, Elevenia, Lazada, Olx and Tokopedia. The example of the search query log data is in the attachment. Table 1 contains a description of each attribute in the data.

<table>
<thead>
<tr>
<th>Atribut</th>
<th>Deskripsi Atribut</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keyword</td>
<td>User search keyword</td>
</tr>
<tr>
<td>Position</td>
<td>Keyword ranking updated by ahrefs</td>
</tr>
<tr>
<td>Position History</td>
<td>History ranking of previous keywords</td>
</tr>
<tr>
<td>Date</td>
<td>Keyword ranking ranking (last date update ahrefs)</td>
</tr>
<tr>
<td>Volume</td>
<td>Number of searches with keywords that are meant by many users</td>
</tr>
<tr>
<td>URL</td>
<td>The target URL of the keyword the user selected</td>
</tr>
<tr>
<td>Difficulty</td>
<td>The approximate difficulties for a keyword are on the search engine's first page search page</td>
</tr>
<tr>
<td>Traffic</td>
<td>Estimate the traffic that will go to a specific URL or website that contains the keyword</td>
</tr>
<tr>
<td>CPC</td>
<td>The value per dollar unit for each user who selects an article (url) containing a particular keyword, when the link in question is already advertised on Google Adwords</td>
</tr>
<tr>
<td>Results</td>
<td>The number of documents that search engines can provide for keywords</td>
</tr>
<tr>
<td>Last Update</td>
<td>The last date ahrefs checks search engines crawls data per each data</td>
</tr>
</tbody>
</table>

2. Preprocessing Data

From the data obtained from the source, the data in the form of a query log that contains several attributes and types of different attributes, then required analysis of data and attributes whichever is good used in the implementation of the system. The series of preprocessing stages of data both data cleaning up to data analysis refers to the preprocessing stages of data for data mining [16], this preprocessing stage is illustrated in Figure 2.

3. System Implementation

Figure 3 describes the sub processes in the system implementation.

![Figure 3. Implementation Steps](image)

In the implementation phase of the system, each data in the dataset is considered as a document, so the implementation of a special TF-IDF method for the data string in each document is then combined with the numerical data that has been discretized for each document. Next is calculated by Cosine Similarity method to get the value of similarity between documents. The cosine similarity value will be used as the reference vector value required for the AD-OPTICS algorithm. In the implementation stage, test scenarios are also conducted, including changes to some reference parameters including the minClSize (eps) and minSizeToCluster (minPts) parameters. The result of the implementation of the algorithm in the form of reachability distance value of each processed data, then will be described manually into a cluster graph. Once cluster found, it will be calculated the quality of the cluster with Purity method. Examples of TF-IDF implementation process are then calculated cosine similarity per each attribute until the implementation phase of the algorithm is described in the attachment.

4. Evaluation

At the evaluation stage, the evaluation is conducted against all results of the test scenario. Evaluation is done by doing combination process of reference variable that used in algorithm. After each of the combinations of variables entered on the system has had results, then performed the analysis process. The analysis is performed on several things such as numerical attribute analysis that influenced the clustering result, the best eps attribute analysis obtained from each iteration of OPTICS algorithm, analysis of iteration time of algorithm implementation, and value index analysis of clustering quality result.

5. Test Result

Based on test scenario, running program is done with one dataset and some combination of variables. The system is built
into two general stages of the OPTICS algorithm and the AD-OPTICS algorithm. In the first iteration of the test, the random variables obtained for the first variable in the running program are minCISize (eps) of 0.6 and minSizeToCluster (minPts) of 5000. The determination of the random value range for the minCISize variable (eps) refers to the value of cosine similarity per data set. The determination of the range of random values for the minSizeToCluster variable (minPts) refers to the number of datasets. The results of running the program in the form of a list of data along with the smallest reachability distance value. Table 2 describes the example of the running program, ie the clustering result of the AD-OPTICS algorithm.

### Table 2. Example Results of the Program

<table>
<thead>
<tr>
<th>Data</th>
<th>Reachability distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>batu bacan batu cincin batu akik batu bacan</td>
<td>0.264194262</td>
</tr>
<tr>
<td>kamera kamera</td>
<td>0.166331421</td>
</tr>
<tr>
<td>batu bacan doko batu cincin batu akik batu bacan</td>
<td>0.169654104</td>
</tr>
<tr>
<td>semprot onderil mobil aksesoris mobil aksesoris eksterior robest carlas rubber paint cat rubber coating cat spray semprot hasil karet</td>
<td>0.169654104</td>
</tr>
<tr>
<td>mobil bekas mobil bekas</td>
<td>0.167594986</td>
</tr>
<tr>
<td>sungai dareh batu cincin batu akik batu sungai dareh</td>
<td>0.169654104</td>
</tr>
<tr>
<td>adidas wanita fashion wanita sepatu sneaker sepatu slip on wanita nike adidas</td>
<td>0.169654104</td>
</tr>
<tr>
<td>rokok elektrik lobi koleksi vapor rokok elektrik</td>
<td>0.169654104</td>
</tr>
<tr>
<td>baterai elektronik baterai baterai charger ultra safire baterai baterai baterai baterai baterai</td>
<td>0.169654104</td>
</tr>
<tr>
<td>batu sapir batu cincin batu akik batu sapir</td>
<td>0.169654104</td>
</tr>
</tbody>
</table>

### IV. CONCLUSION

Conclusions that can be obtained from this research are:

1. The clustering results of some tests show frequently searched items are "used cars", "external hard disk" and "handphone". In the end, the keywords obtained from the results of this study can be used as an e-commerce developer reference opportunity to develop new e-commerce.

2. In this study, the clustering results obtained have the highest Purity value of 0.00565. The purity value is still relatively low. A good cluster is one that has a purity value close to 1. So it can be said that in the clustering process in this study, it produces a less good cluster. This may be influenced by several factors, such as clustering data used has poor distribution of data, the participation of numerical attributes in the cluster affects the results obtained, the cluster extraction is not maximized, and the sensitivity of selecting variables minCISize (eps) and minSizeToCluster (minPts).

3. In this study, the dataset used was not good enough, measured through the distribution of keyword keywords and user-selected urls, not to 10% having a tendency of similarity. This affects the cosine similarity method used to measure the distance between the data, because the basis of cosine similarity calculation for a data, the data has a tendency of good similarity if you want the clustering results obtained a good cluster.

4. The effect of using numeric attributes in calculating data that has no resemblance to keywords and urls affects the result of reachability distance value. The AD-OPTICS algorithm requires appropriate accuracy in selecting the minCISize (eps) and minSizeToCluster (minPts) variable values, since giving them too low or too high will affect the density of the cluster results.

5. The results of this study with respect to the results of previous studies in the paper [5] have different results, especially from the purity value obtained is between 0.946 and 0.984. The results are much different with the results of this study that shows purity of 0.00565. It is also affected by the difference in data sets used, and preprocessing the data used.

### REFERENCES


