TEXT MINING ON NEWS AGGREGATOR FOR RECOMMENDATION SYSTEMS

Kaharisman Ramdhani[1], Galih Hermawan[2]

[1.2] *Indonesian Computer University*
Jalan Dipatiukur No. 112-116, Coblong, Lebakgede, Bandung, West Java 40132
E-mail: khrsman@gmail.com[1], galih.hermawan@email.unikom.ac.id[2]

Abstract - News aggregator is a software or application that combines content from various webs. news aggregator contains a lot of information that can be read by readers. In the news aggregator, information overload can occur due to a lot of news data that is not read by readers because readers only read interesting news. Text mining research in the news aggregator has been carried out by several researchers to group news based on the same topic using different methods, except that in this study cluster data is always initialized at the beginning. This method is less effective in the news aggregator that has a lot of news data types.The use of clustering methods that are not sensitive to initialization can be done to improve the accuracy of each cluster. This research will implement the method of cosine similarity and single pass clustering.The test was conducted with 157 news data and produced an average accuracy of 88.8%. Determination of the threshold is very influential on the accuracy of the data on the cluster produced, the test results indicate that the maximum accuracy obtained is 100% and the smallest accuracy is 63%. The results of this study concluded that the cosine similarity and single pass clustering method can be applied to the news aggregator.

Keywords: text mining, news information, news aggregator, cosine similarity, single pass clustering

I. Introduction

Text mining is the extraction of large amounts of data in the form of text or documents to find useful, hidden and previously unknown information [1]. The goal of text mining is to obtain useful information from a collection of documents. Data source used in text mining is a collection of texts that have a minimal format of unstructured or semi-structured, According to research from Lokesh Kumar [2], text mining and information extraction has become a popular area of research to extract useful information. It is very important to develop techniques and better algorithms to extract useful information.One of the areas that do text mining is on the news aggregator.

In a news aggregator system, news grouping is important because every group has is own topics containing articles from multiple sources. The quality of the news group is very important because it helps reader to select the desired news topics. News grouping of Indonesian language news has been done by several researchers with different techniques and objectives. The method of classifying news by using partitional clustering with cluster initialized is the simplest technique and commonly used for Indonesian news [3,4] because the method is easily implemented [4]. This method performs clustering by initialization of each cluster, each cluster is initialized randomly so that the group data can vary. If the random value for the initialization is not good, then the generated group is not optimal. The use of partitional clustering method with cluster initialization on the news aggregator still produce outliers or documents that should not be in one cluster [5]. Based on the research that implement one of the partitional clustering methods with cluster initialization [5], by applying a partitional clustering method on news aggregator sometimes produce over cluster if the number of documents is too much. The results of these studies suggest to use other types of clustering methods that are not sensitive to initialize or method with dynamic initialization.

II. THEORETICAL BASIS

a. web scrapping

web scrapping is a technique to gather data or information from a website or blog using the software with a specific method. Such software usually stimulate human activity on a website or blog by using low - level HTTP or using a web browser [9].Web scraping has many uses and very helpful for the retrieval of documents, one of which is for news content where the content taken directly from the targeted site. There is generally stages in implementing web scrapping techniques:

1. *Calling / Request url targets.*
   The system will call the target in this case is http url address of the web that are targeted, for example www.kompas.com
2. *Process on the target server,*
   The target server will make the process for the request that has been done, then it will present data based on what is in the request.
3. *Extraction of data.*
   The system will extract the html data that are provided by the server. Data were deemed important to be taken, then the result of this extraction will then be stored into the database

b. *News aggregator*

News aggregatoris a software or application that combines content from around the web [10]. for example, such as online newspapers, blogs or video blogs in one place for easy reading.

This technology enables users by combining data from multiple websites into one page and can show information that has been updated from the site.

c. *TF-IDF*

Term Weighting TF-IDF is a scheme that is widely used in the weighting of the word. Some things to note in the search for the document information is weighted term. Can be words, phase or other index results in a document, each word given indicator called terms of weight.

1. *Term Frequency (TF)*

TF is the frequency of the appearance of a word or term in a document. The greater the number of appearance greater the weight or value greater compliance. Based on the book written by Jiawei Han [12] the formula to get the value of this TF is

$$TF(d,t) = \begin{cases} 0 & \mathit{if\,freq}(d,t) = 0 \\ 1 + \log(1 + \log(\mathit{freq}(d,t))) & \mathit{otherwise}. \end{cases} \qquad \textbf{(1)}$$

Information:
TF (d, t) : Term value t d in the document.
freq (d, t) : The number of terms in the document d t.

2. *Inverse Document Frequency (IDF)*

IDF shows the relationship of a term in the entire document. The fewer the number of documents that contain the term in question, the greater the value of the IDF. Inverse Document Frequency (IDF) can be calculated by using a calculation

$$IDF(t) = \log \frac{1+|d|}{|d_t|}, \qquad \textbf{(2)}$$

Information:
IDF (t) : Value of IDF.
d : Total number of documents
dt : number of documents that contain the term t

Once the value of its TF and IDF has been known, next is to determine the value of its ITF-IDF using:

$$TF\text{-}IDF(d,t) = TF(d,t) \times IDF(t). \qquad \textbf{(3)}$$

Information:
TF-IDF (d, t) : Tf-idf value
TF (d, t) : Value of tf
IDF (t) : Value of idf

d. *cosine Smilarity*

cosine similarity is the method used to calculate the level of similarity / similarity between two objects. In this case the comparison is two documents. This is the cosine similarity value of 0 toward 1, the greater the value, the greater the similarity between the two documents.

$$sim(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1||v_2|}, \qquad \textbf{(4)}$$

$$sim(_{A,B}) = \frac{\sum_{i=1}^{n} \boldsymbol{A_i B_i}}{\sqrt{\sum_{i=1}^{n} \boldsymbol{A_i^2}} * \sqrt{\sum_{i=1}^{n} \boldsymbol{B_i^2}}} \qquad (5)$$

Information:
$\sum_{i=1}^{n} \boldsymbol{A_i B_i}$ *: The number of words contained in the document A and contained in document B.*

$\sqrt{\sum_{i=1}^{n} \boldsymbol{A_i^2}}$ *: The number of words contained in the document A.*

: The number of words contained in the document D2. $\sqrt{\sum_{i=1}^{n} \boldsymbol{B_i^2}}$

e. *Single Pass Clustering*

single pass clustering is a method used to combine multiple documents into one group by grouping the data one by one to form a group with the evaluation of any data input into the cluster. Evaluation of similarity between data and the cluster *can be done in various ways including using distance functions, vector similarity and others.*

In using this algorithm, two things that needs attention is the determination of the objective function and the determination of the threshold value. Specified objective function should be as far as possible reflect the state of the data that is modeled and can provide the level of similarity or difference values contained in the data. Determination of the threshold value is also a subjective thing, the greater the threshold value, the harder the data to merge into a cluster, and vice versa.

The algorithm that often used in a Single Pass Clustering is as follows: [13]

1. For each iteration documents

a. Find new clusters by comparing the document D and the other documents to look for the value of its cosine similarity.

b. if the value of the cosine similarity is greater than the threshold value, then the comparison document is included in cluster C. The value of cluster C thus formed is the middle value between the document D and the documents being compared.

c. If the cosine similarity value is less than the threshold value, then create a new cluster consisting of documents D alone

2. The loop ends

III. RESEARCH METHODS

The method used in constructing this sistem is descriptive research method [7]. This method is used because in this study data sources are from the internet and no manipulation of variables in the data. The data used is data obtained as it is. Therefore, the descriptive method deemed suitable for use in this study. Stages of research to be conducted are:

a. Identification of problems

b. Data collection

c. analysis

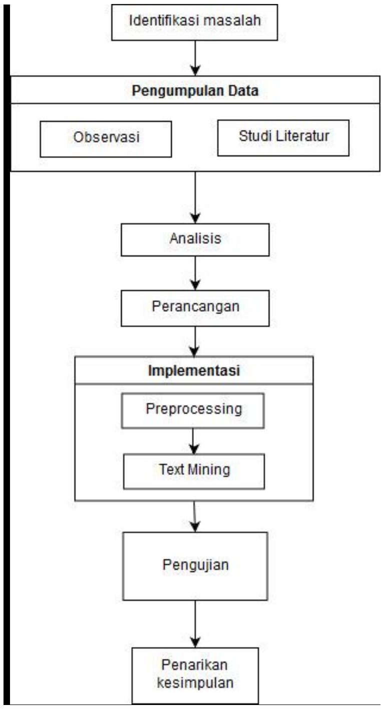
d. Design

e. Implementation

f. Examination

g. Conclusion

Overview of the steps being taken can be seen in Picture 1



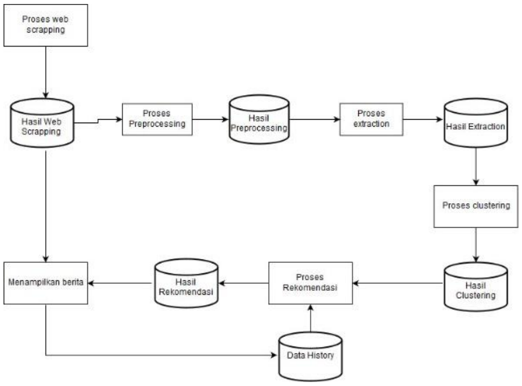
Picture 1 Research methods

IV. ANALYSIS

This stage is to identify and evaluate all components of the system to be built.

a. System Overview

General overview system to be built can be seen in picture 2



Picture 2 System Overview

b. Input Data Analysis

The data that will be processed for this system is textual data which obtained from news content. News content will be used as the input data that were collected from several news portals using the web scrapper. The data is then stored into the database.

Other input data that used are news history data read by the reader. This data will be used as a reference for the recommendations given by the news aggregator.

c. *preprocessing*

In the preprocessing stage, there are some steps being taken before news content data processed by cosine similarity method. Preprocessing process is divided into several stages: Cleaning, Case Folding, tokenizing, and Stopword Removal.

1. ***cleaning***

   This is the stage for removal process of markup tags and special format of news content that will be processed. News content on the database that resulted from the scrapping process is still contain HTML tags. Therefore, at this stage, news data will be cleared from the html tag. At this stage all the html tags such as <p>, <strong> <a href="#"> will be removed from the data.

2. ***Case folding***

   This is the process of converting uppercase to lowercase or also called case folding. Case folding phase change all the letters from "a" through "z" in the document to lowercase. Not all words in the document are consistent use of capitalization, and therefore the case of folding convert the entire text in the document to lowercase.

4. ***tokenizing***

   Tokenizing is decomposition process in the form of sentences containing words and separators between words as a period (.), comma (,), space () and another separator sign into tokens or single word pieces.

5. ***stopword removal***

   Stopword removal is the process of taking important word from results of tokenizing. This step is using stopword removal algorithm (removing the less important). This method is done by eliminating the unimportant words (stopword) documents through words checking whether the results of tokenizing included in the list of words is not important or not. If included in the stopword then these words will be removed from the document so that the remaining words in the document is considered as the important words.

d. *clustering*

The clustering process that used in this research is cosine similarity and single pass clustering. This method is done in three stages, namely the TF-IDF weighting, cosine similarity values, and clustering. Here are the steps done to make clustering documents.

1 ***TF-IDF stages***

- *The first step is finding the value of TF and DF between the compared documents.*
- *The next step is to find TF value of each document.*

$$TF(d,t) = \begin{cases} 0 & \text{if } freq(d,t) = 0 \\ 1 + \log(1 + \log(freq(d,t))) & \text{otherwise.} \end{cases}$$

TF (d, t)is a value of term d in document t. freq (d, t) is the number of terms d in the document t.

- *After the value of TF and DF has been known, find the value of IDF.*

$$IDF(t) = \log \frac{1 + |d|}{|d_t|},$$

- *The final step is to find the value of TF-IDF,*

$$TF\text{-}IDF(d,t) = TF(d,t) \times IDF(t).$$

2. ***Cosine Similarity stages***

- *The first step is calculating the value of $v_1 v_2$ by summing the value of the values in the document D1 (v_1) and D2 (v_2),*

$$sim(_{v1,v2}) = \frac{v_1 v_2}{|v_1|\,|v_2|}$$

$v_1 v_2$ is the number of words contained in the documents D1 and D2.
$|v_1|$ is the number of words contained in the document D1.
$|v_2|$is the number of words contained in the document D2

- *The next step is to determine the value of $|v_1|$ by calculating the root of the product of the value of every word in D1.*
- *The next step is to determine the value of $|v_2|$ by calculating square root of the result of every word on the D2.*
- *The final step is to determine the cosine similarity value from values which have been obtained.*

3. ***Clustering Stages***

- *This stages is comparing cosine similarity values with the specified threshold. If the cosine similarity value exceeds the threshold then the document compared will become one cluster, but* *if it is less than the threshold then it becomes new cluster.*

That three stages is repeated until all documents are successfully compared and all documents inserted into clusters.

e. Recommendations Process

News aggregator will display all of the data that is stored on a database. By the time the user selects one of the news that is shown, the system will save selected news. The system will then display the news in the cluster that is similar to user selected news.

V. TESTING

Cluster accuracy testing was conducted to determine the accuracy of the clusters. The test is performed by comparing the cluster formed by the clustering system with manually checked data. The test is performed with 157 news data that were extracted using web scrapping methods from www.kompas.com and www.tribunnews.com.

The similarity between documents in clusters is evaluated using the standard method that can be seen in table 1

Table 1 Category Classification Results

	in Event	*Not In event*
in cluster	*a*	*b*
Not In Cluster	*c*	*d*

The above table ***Error! Reference source not found.****shows that the classification results may be the correct data (a) or the wrong data (b). While the document that is not included in the classification results sometimes is the wrong data (d) and sometimes the correct data but not in clusters (c).*

To test the accuracy or precision level of the classification results of all classification documents, the formula p = a / (a + b) if a + b> 0 is used. The test is performed with a different threshold value.

1. *The first test with a threshold of 0.3*

   The first test is done by using threshold value 0.3. In this test, the number of formed clusters is 118. The test results with threshold 0.3 can be seen in Table 2

Table 2 Testing Results With Threshold 0.3

Total Cluster Generated	***Total Data In Clusters***	***Total Data***	***Correct Data***	***Wrong Data***
104	*1*	*104*	*104*	*0*
18	*2*	*36*	*36*	*0*
3	*3*	*9*	*9*	*0*
2	*4*	*8*	*8*	*0*
118		*157*	*157*	*0*

The test result by threshold 0.3 on 157 data generates 157 correct data, so the accuracy is 100%.

2. *The second test with a threshold of 0.2*

   The test results with threshold 0.2 can be seen in Table 3

Table 3 Testing Results With Threshold 0.2

Total Cluster Generated	***Total Data In Clusters***	***Total Data***	***Correct Data***	***Wrong Data***
85	*1*	*85*	*85*	*0*
14	*2*	*28*	*28*	*0*
8	*3*	*24*	*25*	*0*
2	*4*	*8*	*8*	*0*
1	*5*	*5*	*5*	*0*
1	*7*	*7*	*7*	*0*
111		*157*	*157*	*0*

The test result by threshold 0.2 on 157 data generates 157 correct data, so the accuracy is 100%.

3. *The third test with a threshold of 0.1*

   The test results with threshold 0.1 can be seen in Table 4

Table 4 Testing Results With Threshold 0.1

Total Cluster Generated	*Total Data In Clusters*	*Total Data*	*Correct Data*	*Wrong Data*
68	*1 -5*	*100*	*100*	*0*
3	*6 -10*	*24*	*22*	*5*
1	*11-15*	*13*	*13*	*0*
1	*16-20*	*20*	*15*	*5*
73		*157*	*147*	*10*

The test result by threshold 0.1 on 157 data generates 147 correct data, so the accuracy is 93%.

4. *The third test with a threshold of 0.07*

   The test results with threshold 0.07 can be seen in Table 5

Table 5 Testing Results With Threshold 0.07

Total Cluster Generated	*Total Data In Clusters*	*total Data*	*Correct Data*	*Wrong Data*
44	*1-5*	*67*	*67*	*0*
5	*6-10*	*41*	*28*	*13*
1	*11-15*	*13*	*13*	*0*
2	*16-20*	*36*	*29*	*7*
52		*157*	*137*	*20*

The test result by threshold 0.07 on 157 data generates 137 correct data, so the accuracy is 87%.

5. *The fifth test with a threshold of 0.05*

   The test results with threshold 0.05 can be seen in Table 6

Table 6 Testing Results With Threshold 0.05

Total Cluster Generated	*Total Data In Clusters*	*total Data*	*Correct Data*	*Wrong Data*
21	*1-5*	*36*	*34*	*2*
5	*6-10*	*35*	*29*	*6*
2	*11-15*	*23*	*15*	*8*
1	*16-20*	*16*	*6*	*10*
1	*> 20*	*47*	*17*	*30*
30		*157*	*101*	*56*

The test result by threshold 0.07 on 157 data generates 101 correct data, so the accuracy is 64%.

From the five test that has been done with the different threshold value, test results can be simplified as seen in Table 7

Table 7 Test result

threshold	*the number of clusters*	*accuracy*
0.3	*118*	*100%*
0.2	*111*	*100%*
0.1	*73*	*93%*
0.07	*52*	*87%*
0.05	*30*	*64%*

From the above test results can be seen that the threshold value greatly affect the number of clusters formed. The larger the specified threshold value, the accuracy can reach 100% but with few cluster members. Otherwise, the smaller threshold value resulting more cluster member but with less accuracy and may cause some data appear in wrong cluster. Average accuracy obtained from the test results was 88.8%.

VI. CONCLUSION

Based on the implementation and tests performed on the news aggregator system by applying cosine similarity and single pass clustering method, it can be concluded that the method used successfully classify the news based on the contents of the news with an average accuracy of 88.8%. The result clusters have different levels of accuracy depending on the specified threshold. Even so, the method still has drawbacks. By using this method there are still some news that should not exist in a cluster because this method is only checking the text and not its subject.

REFERENCE

[1] Daniel Waegel. 2006. The Development Of Text-Mining Tools And Algorithms. Thesis. Ursinus College.

[2] Lokesh Kumar, Parul Kalra Bhatia. 2013. Text Mining: Concepts, Process and Applications. Journal of Global Research in Computer Science. 4(3): 36-39

[3] Wibisono Y., dan Khodra, M. L. 2005. Clustering Berita Berbahasa Indonesia, Jurnal FPMIPA UPI.

[4] Husni, Y.D.P. Negara,M. Syarief. 2015. Clusterisasi Dokumen Web (Berita) Bahasa Indonesia Menggunakan Algoritma K-Means. Jurnal SimanteC. Vol. 4, No. 3.

[5] Thoriq B Achmad, Nelly Indriani. 2015. Web Content Mining Menggunakan Partitional Clustering K-Means Pada News Aggregator. Jurnal Sistem Komputer. 5(2): 42-46

[6] Ferbruariyanti Henry, Zuliarso Eri. 2012. Algoritma Single Pass Clustering Untuk Klastering Halaman Web. Prosiding Seminar Nasional Komputer dan Elektro (SENOPUTRO) 1: 1-8

[7] Sugiyono. 2010. Metode Penelitian Kuantitatif dan Kualitatif dan R & D. Bandung: CV Alfabeta.

[8] M. Romli. 2012. Jurnalistik Online: Panduan Mengelola Media Online Bandung: Nuansa.

[9] Google Inc, “Google,” 2010. [Online]. Available: http://www.google.com/webmasters/docs/search-engine-optimization-starter-guide.pdf

[10] Angela M. Lee. 2015. The Rise of Online News Aggregator : Consumption and Competition. International Journal on Media Management. 00:1–22

[11] O. Maimon dan L. Rokach. 2010. Data Mining and Knowledge Discovery Handbook. New York: Springer-Verlag New York Incorporated

[12] J Han, M Kamber, 2006. Data Mining Concepts and Techniques (2nd Edition). Massachusetts: Morgan Kaufmann

[13] Salton, G., 1989. Automatic Text Processing, The Transformation, Analysis, and Retrieval of Information by Computer. Boston: Addison – Wesly Publishing Company, Inc. All rights reserved.

[14] Anhar. 2010. PHP & MySql Secara Otodidak. Jakarta: PT TransMedia

[15] Alexander F. K. Sibero. 2011. Kitab Suci Web Programing. Yogyakarta: MediaKom.