Automatic Representative News Generation using Automatic Clustering

Diptia Zandra Eka Puspitasari , Ali Ridho Barakbah, Idris Winarno
Electronic Engineering Polytechnic Institute of Surabaya
Institut Teknologi Sepuluh Nopember (ITS) Surabaya
Email: diptia@student.eepis-its.edu, ridho@eepis-its.edu, idris@eepis-its.edu

Abstract

More than 2000 news presented by 32 online news sites in Indonesia in one day, it can make user who don't have enough time to access it being difficult to choose which news that worth enough to read for them because there are news which have same topic and content among of those news. Cluster the news automatically which can provide news representative from all similar news is the best solution to cover news redundancy problem. This paper presents a new approach of automatic representative news generation using automatic clustering. This approach involves 5 steps which are (1) Data Acquisition, (2) Keyword Extraction, (3) Metadata Aggregation, (4) Automatic Clustering, and (5) Representation News Generation. Data Acquisition is used to generate the news from RSS and present the news description that tokenized and filtered in Keyword Extraction Process. Token values, token links, and tokens are the result of Keyword Extraction and inputted into Metadata Aggregation process to provide a matrix of token values of each links. By using Automatic Clustering method, the system can identified the match number of cluster and clustered the news automatically to provide the news representative to the users. The news representation can be found by finding the news which has shortest distance with centroid in each cluster. The results of news representative depend on the token value of each links, if the difference value of cluster is too small, it means that the news are much-separated news, but if the difference value of cluster is too big, that means the news are less-separated news. The longer time that taken as a refresh-time, the automatic clustering results will be more accurate, because the more data that can be formed as a cluster.

Keywords : Data Acquisition, Keyword Extraction, Metadata Aggregation, Automatic Clustering, Representation News Generation

1. Introduction

Along with the increasing number of internet users in Indonesia, the more sites that provide information for Internet users, including information such as news. Lots of news media (both TV and newspapers) shifted to online news sites due to the increase of Internet users. Based on paper starting on 20 up to June 21, 2012, the number of news on the internet reach 2000 more each day. Sometimesthe news that displayed by a site is taken from other sites. It triggers news redundancy on the internet which led the increasing number of presented news. This phenomenon does not allow users to read even choose the news and from thousands the amount. Therefore, it urgently needs solutions to resolve news redundancy problem. The purpose of this paper was to choose a representation of news that has been taken from 32 online news sites in Indonesia. By using the Automatic Clustering algorithm, it produced a news as a representation of each different news clusters. In addition to search the news representation, this paper also can classify the news based on the description of the news content automatically.

Paper related to the news representation performed by George Adam and Vassilis Poulopoulos who collaborated with Christos Bouras[1]. In their paper, they describe a mechanism that fetches web pages that include news articles from major news portals and blogs. Constructed in order to support tools that are used to acquire news articles from all over the world, process them and present them back to the end users in a personalized manner. Another paper about document clustering process performed by Michael Steinbach who collaborated with George Karypis and Vipin Kumar [2]. They conduct paper on some common document clustering techniques. In particular, they compare the two main approaches to document clustering, agglomerative Hierarchical clustering and K-means. Another paper which does the paper about clustering search engine is “A Dynamic Clustering Interface to Web Search Results” [3]. The researchers of this paper are Oren Zamir and Oren Etzioni. They introduce Grouper – an interface to the results of the Husky Search meta-search engine, which dynamically groups the search results into clusters labeled by phrases extracted from the snippets.

This paper presents a new system for generating automatic representative news using automatic clustering. As automatic representative news, this paper must generate news from 32 online sites in Indonesia choose the news representative of each news theme. The news data obtained from RSS feed in those online news sites. Text mining is really used to read the news contents of each news data. By using tokenizing and filtering method, the tokens and token

values that really needed for clustering process can be produced. Aggregation process is the part of changing token values into matrix, so that clustering process can be done easily. The matrix of token values is also become the coordinate for link to map the news links. The clustering process used Automatic Clustering because the number of cluster not known. That is impossible to count the number of cluster manually. By using Automatic Clustering, more than 2000 news can be clustered automatically per day. This application can also be called as storage news from a variety of existing online news site. This paper work automatically every day (12 hours in per day). All news representatives collected automatically, so that users can access it every day.

2. System Design

Automatic Representative News Generation Using Automatic Clustering is a new generation of online news site that can present a news representative of all similar news from various sites. By collecting RSS from 32 online news sites in Indonesia, this paper can produce news representatives that can become the representation of news with the same topic. This paper did three experiments on 21st, 22nd, and 26th of July 2012. The time intervals of each experiment are one hour.

The result of text mining process is token/words with its value. Token values are the number of words in its documents. To get this value, there is a function to count the number of words in each document automatically. Token values are the value that will be clustered using automatic clustering.

Before clustering process, aggregation process has to done first. This process is to present the matrix of token value that will be clustered in the next process. For aggregating token values, the parameters are token link (link of its news) and token word. Token links can assumed as number of rows, while token words can assumed as number of columns. After redefining token links as rows and token words as columns, the next step is inputting the token values. The inputting process is depending on token links and token words. The system checked is the token values has same source of token link and token words. If the parameters are true, so the token value will be inputted. But, if the token value has null at those parameters, the token value will be replaced with 0.

Automatic Clustering worked after the matrix of token value has been inputted. This process found the precision number of cluster automatically. Automatic Clustering is really needed because it is impossible to find the number of cluster manually among the hundred numbers of data.

To find the news representative of each cluster, the method that used is Euclidian Distance. Centroid of each cluster has to be found first before distance counting. The news representative of each cluster is the news which has shortest distance from centroid.

Figure 1 shows the system architecture of our proposed system.

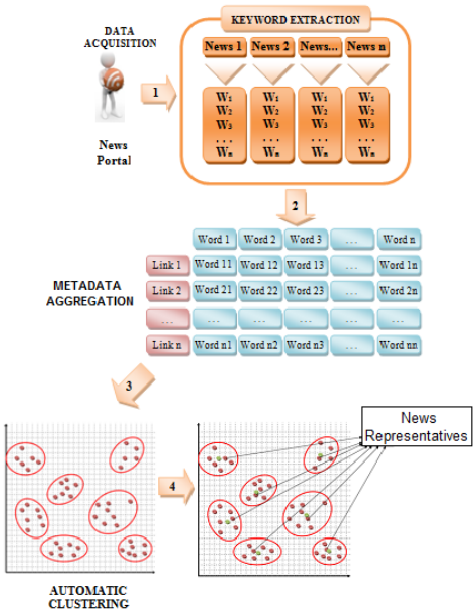


Figure 1. System architecture of our proposed system

2.1. Data Acquisition

The first step of this paper is getting RSS feeds from several online sites in Indonesia. The online site will be updated every day (24 hours fully) to get more news that may be presented to the user. The RSS which contains news links, headlines, news and news descriptions is collected in the table "news" in the MySQL database. The next step is doing text mining on the news descriptions. The process of text mining aims to get the token and its value ("tf": the number of words in the news) has been explained bellow.

1. Get RSS feed from online news sites (32 online news sites in Indonesia).
2. After RSS obtained, the contain of RSS will be separated into link, head news, and news description.

Separated RSS contain inputted into table "news" in Database.

.

2.2. Keyword Extraction

In order to get the information from the news, we have to understand what is the news contains. On this step, we will use the text mining process. The output is tokens and token values.

1. Select news description and transform it into lowercase.
2. Replace all punctuation with "space".

3. Tokenize the news description. To tokenize it, replace the "space" with "enter", so that the token produced.
4. Count the number of each token to get the token values.
5. Select stoplist, compare the stoplist with tokens.
6. If the tokens are same with stoplist, the tokens deleted.

Input the tokens and the token values into database to process it in aggregating and automatic clustering process.

2.3. Metadata Aggregation

Aggregation Process is a step to change the patter of token values into matrix. Token value aggregated to get the node of each link document. Actually, the dimension of node each link is different, because node dimension is depending on the number of words. This matter can cause a problem in the clustering process. To solve it, we need to made node dimension of all link being same. That is why aggregation process is really needed.

To enter the token value into aggregation matrix, parameters like token words and token links are needed. Token value will be entered in a dimension that has same token words and token link. If the token words and token link is not same or those words are not in some link, the token value will be 0. The algorithm of metadata aggregation is:

1. Select distinct the tokens / words from database. Select distinct is used to prevent duplicate words.
2. Select distinct the links from database. Same with step number one, select distinct is used to prevent duplicate links.
3. Select the token values.
4. Count the number of tokens and redefine it to number of column.
5. Count the number of links and redefine it to number of rows.
6. Make an array to patch the token values. It is two dimensions array. The long of first dimension length is like number of column, while the second length dimension is like number of rows.
7. Input the token values into the matrix. To input the token values, we have to compare the parameters (tokens and links). If the token values [tokens] = token, so the token value inputted into matrix / array, except this condition the token value is 0. If token values [links] = links, the token value inputted into matrix/array, except this condition the token value is 0. The result can be seen in **Table 1**.

Table 1. The Example of Aggregation Result

	Word 1	Word 2	Word 3	Word 4	Word 5
Link 1	0	1	2	1	0
Link 2	2	1	0	0	1

8. The nil (0) value on matrix is the replacement of null token values. So that, the size matrix for all links is same.

The aggregation process that carried out in this paper is stemming all the data in the table will be order by the headline, so the news will not appear double.

After all data in those table ordered by head news, the data have to be accommodated in a query which ordered by token. This function is used to sort all data in stemming table and make it easy to be compared with main value.

2.4. Automatic Clustering

Clustering process is a step to classify the news to get news which can represent other news in its cluster. Due to amount of news that will be clustered is being unknown, it will be difficult to cluster the news with usual clustering techniques. Therefore, the automatic clustering process will be used for this process.

Each cluster has their own Cluster Variance (v_c). This cluster variance can be called an identity of each cluster [4]. The equation to find the variance has been present on Eq. 1.

$$v_c^2 = \frac{1}{n_c - 1} \sum_{i=1}^{n_c} (d_i - \bar{d}_i)^2 \quad (1)$$

Where:

v_c^2 : variance in cluster c
c : 1 … k, where k = number of cluster
n_c : number of data in cluster c
d_i : i data in a cluster
$\bar{d}_i$: the average data in a cluster

Variance within (v_w) and Variance between (v_b) of this process/part is really need to count the Variance (v) from all cluster. The equation to count the value of Variance within Cluster (v_w) has been shown in Formula (2).

$$v_w = \frac{1}{N - k} \sum_{i=1}^{k} (n_i - 1).v_i^2 \quad (2)$$

Where:

v_w : variance within
N : number of all data
k : number of cluster
n_i : number of data in each cluster

v_i^2 : variance of each cluster

The equation to count the value of Variance within Cluster ($_{Vw}$) has been shown in Eq. 3.

$$v_b = \frac{1}{k-1}\sum_{i=1}^{k} n_i \left(\vec{d}_i - \vec{d}\right)^2 \quad (3)$$

Where:

v_b : Variance between
k : Number of cluster
n_i : Number of data in each cluster
$\vec{d}_i$: The average data in a cluster
$\vec{d}$: The average of all data

A good cluster is a cluster that the Variance within cluster ($_{Vw}$) is greater than the Variance between clusters ($_{Vb}$), thus obtained Eq. 4.

$$V = \frac{Vw}{vb} \times 100\% \quad (4)$$

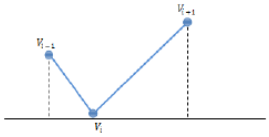
Where:

V : variance
$_{Vw}$: variance within cluster
$_{Vb}$: variance between clusters

To obtain the number of clusters, we automatically apply an automatic clustering algorithm called as Tracing Valley which proposed by Barakbah and Arai [4]. Identify of the variance moving pattern is shown in Figure 2.

Pattern	Possible?	Pattern	Possible?
	√		X
	√		X
	√		X
	X		√
	X		X
	X		X
	X		X
	X		X

(a)



$$(V_{i-1} \geq V_i) \cap (V_{i+1} > V_i)$$

(b)

Figure 1. Moving Variance Pattern consisting of (a) possibility of patterns to be a global optimum and (b) different value of altitude [4]

After the pattern of moving variance is known, then we must seek a global optimum value of each cluster.

$$\partial = (V_{i+1} - V_{i-1}) + (2 - V_i) \quad (5)$$

Where ∂ is a value difference expressing the position of the global optimum.

(3)

After the results obtained, we find the greatest value of ∂, which is the value that the global optimum value.

$$\max(\partial) = global\ optimum \quad (6)$$

To check accuracy of number of cluster, we used Eq. 7.

$$\varphi = \frac{\max(\partial)}{closer\ value\ to\ \max(\partial)} \quad (7)$$

Where:

φ : accuracy ratio
max(∂) : max value of global optimum (4)

The number of cluster that will be used for the next process is the number of cluster of process which has highest global optimum value (max ∂). The value of φ can show the distant value to get global optimum. The large φ, at least $\varphi \geq 2$, expresses possibility to construct well-separated cluster [4].

2.5. Representation News Generation

To get news representative, the system has to count the new centroids of each cluster. Based on the new centroids, the distance between centroids and all cluster members can be counted. The cluster member with shortest distance is the representative of its cluster.

1. Find the centroid of each cluster. To find the centroid. The centroid can be obtained by get the average number of all data.

$$C = \frac{x_A + x_B + \cdots + x_k}{k} \quad (8)$$

2. Count the distance between the centroid and all data in that clusters. To find the distance can use Euclidian Distance.

$$\sqrt{(xA - xB)^2 + (yA - yB)^2} \quad (9)$$

Find the nearest distance. The news with nearest distance of centroid is the representative news of this cluster.

Table 2. Experimental Result

Date	Start Time	End Time	Num. of News	Num. of Tokens	Num. of Clusters	Max Global Optimum (∂)	Accuracy Ratio (φ)	Time (second)
21-Jun	00:00	01:00	103	1045	101	0.02209	1.184991	426
22-Jun	07:00	08:00	113	959	9	0.01189	1.051636	525
23-Jun	12:00	13:00	102	843	57	0.01798	1.385382	424

3. Experimental Study

For this experiment, we generate RSS from 32 online news sites in Indonesia. There are 4 experiments for paper. The time interval of each experiment is one hour. The result of all experiment has been listed in **Tabel 2.**

Experiment 1

Tthe difference number between number of links and number of cluster is too small (103 news become 101 clusters). The number of clusters on experiment 1 cannot reached the global optimum, it has been proved by the accuracy ratio of this experiment. Based on the paper of Barakbah and Arai, the large accuracy ratio (φ), at least $\varphi \geq 2$, expresses possibility to construct well-separated cluster [8]. While, the accuracy ratio (φ) of experiment 1 is less than 2 (φ = 1.184991). As the result, the number of related news in each news cluster is small too; even it can be one news only. That condition can happen because of the news separation is too scattered. The distance of each cluster is too far, so that they cannot be clustered well. For the real description of data pattern in experiment 1 has been illustrated in **Figure 3**.

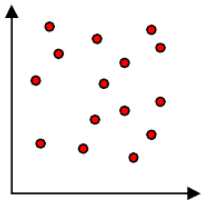


Figure 2. Scattered Data Distribution

Experiment 2

The result in experiment 2 is really different with experiment 1. The number of link is 113 news and the number of cluster that provided by automatic clustering process is 9 cluster. Based on that result, the number of related news in each news representative become more. One news representative can have more than 10 related news. The number of clusters on experiment 2 cannot reach the global optimum, it has been proved by the accuracy ratio of this experiment. Based on the paper of Barakbah and Arai, the large accuracy ratio (φ), at least $\varphi \geq 2$, expresses possibility to construct well-separated cluster [8]. While, the accuracy ratio (φ) of experiment 1 is less than 2 (φ = 1.051636). That condition can happen because of the news separation is too condensed. The distance of each cluster is too far, so that they cannot be clustered well. For the real description of data pattern in experiment 1 has been illustrated in **Figure 4**.

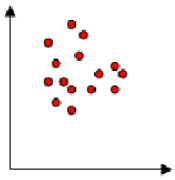


Figure 3. Condensed Data Distribution

Experiment 3

A different result has been providing by the experiment 3. Although the number of cluster is not too much or not too high, but the experiment cannot reach the global optimum. It has been proved by the number of accuracy ratio is less than 2 (φ = 1.385382).

The data on experiment 3 is a little bit above the scattered data distribution condition, but it is closer to condensed data distribution. That is the reason why the number of cluster is not good number even though the number of cluster is not too much or too less.

Time Analysis

The time interval of each experiment is one hour, and the number of news is 72 until 113 news. If the time interval of this experiment be augmented three times or more, so that the number of news will be increased too. The increasing number of news can make a possibility of well-separated news because the similar news will be increased too. The increasing number of news can be seen in **Figure 6**.

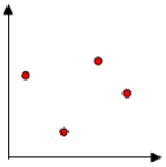


Figure 4. Scattered data distribution with time interval = 60 minutes

The red bullet is initial news, and the yellow bullet is new news. The increasing of time interval can increase the number of news too. So the automatic clustering results will be better.

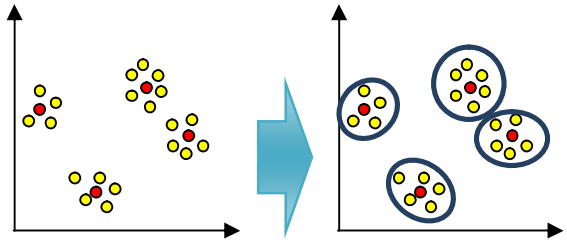


Figure 5. The Possibility of Clustering Result after Increasing Number of News

4. Conclusion

This paper presented an automatic representative news generation by applying Automatic Clustering [4]. Our proposed system generated news representatives from 32 news sites in Indonesia without determine the number of news clusters manually, but automatically. The values of link index give a significant influence for automatic clustering result. The similarity of link index can caused unrelated news become one clusters, so that the result is unexpected result. If the difference value of number of links and number of cluster is too far, it means the cluster become much-separated cluster, but if the difference value of number of links and number of cluster is too near, it means the cluster become less-separated cluster. The time interval of data acquisition also gives significant effect for automatic clustering process. The more time has taken, so the more number news that will be clustered. It can make the stretch of the data become bigger, so that the value of accuracy ratio (φ) will be increased too.

References

[1] *Efficient extraction of news articles based on RSS.* George, A., Christos, B., & Vassilis, P. (n.d.) . Computer and Informatics Engineer Department, University of Patras.

[2] *A Comparison of Document Clustering Techniques.* Steinbach, Michael, Karypis, George. and Kumar, Vipin. 2008. Twin Cities : Universty of Minnesota, 2008.

[3] *Grouper: A Dynamic Clustering Interface to Web Search Results.* Zamir, Oren and Etzioni, Oren. 2010. Seattle : Department of Computer Science and Engineering, 2010.

[4] *Determining Constraints of Moving Variance to Find Global Optimum and Make Automatic Clustering.* A.R. Barakbah; K. Arai. 2004. Surabaya : IES, Politeknik Elektronika Negeri Surabaya, 2004.

[5] Johnson, Roger A., *Advanced Euclidean Geometry*, Dover, 2007 (orig. 1929): p. 173, corollary to #272.

[6] *Clustering.* Barakbah, Ali Ridho. 2006. Surabaya : Soft Computing Research Group EEPIS-ITS, 2006.