EMITTER International Journal of Engineering Technology
Vol.1, No.1, December 2013
ISSN : 2443 - 1168

Automatic Representative News Generation using On-Line Clustering

Marlisa Sigita, Ali Ridho Barakbah, Entin Martiana Kusumaningtyas, Idris Winarno

Department of Information and Computer Engineering
Electronic Engineering Polytechnic Institute of Surabaya
Email: marlisa@student.eepis-its.edu, ridho@eepis-its.edu , entin@eepis-its.edu, idris@eepis-its.edu

Abstract

The increasing number of online news provider has produced large volume of news every day. The large volume can bring drawback in consuming information efficiently because some news contain similar contents but they have different titles that may appear. This paper presents a new system for automatically generating representative news using on-line clustering. The system allows the clustering to be dynamic with the features of centroid update and new cluster creation. Text mining is implemented to extract the news contents. The representative news is obtained from the closest distance to each centroid that calculated using Euclidean distance. For experimental study, we implement our system to 460 news in Bahasa Indonesia. The experiment performed 70.9% of precision ratio. The error is mainly caused by imprecise results from keyword extraction that generates only one or two keywords for an article. The distribution of centroid's keywords also affects the clustering results.

Keywords: News Representation, On-line Clustering, Keyword Aggregation, Text Mining.

1. INTRODUCTION

The number of Internet user is growing from 42 million in 2010 to 80 million users in 2013 according to a prediction by Indonesian Ministry of Information and Communication [1]. Similarly, the number of online news provider is more in quantity because some mass media like newspaper, magazine, and TV channel also provide their content in the Internet. The huge volume of online news can be an advantage but also can be inefficient too. By 2009, the approximate number of the number of online news sites was 45 sites [2]. It would be difficult to digest all information by visiting each link. According to our experiment, the number of gathered news for 19 hours

from 25 online news providers can reach up to 2118 different news. And a survey held by Diptia Zandra since 21st June 2012 showed that the number of news that has been presented by 32 online news sites can reach more than 2000 news every day or approximately 83 articles/news every hour [3].

Generally to keep up-to-date using Internet is by visiting news sites or browsing through any different sites. This way leads to inefficiency due to news redundancy because it is possible that we find different title yet the same news content. But existing RSS present myriad number of news and redundancy. Consequently, finding information would be inefficient. Considering the problem about how to consume news easily and efficiently behind big number news in the Internet, a system that clusters the news and generate its representative become highly demanding.

2. RELATED WORKS

There are some researches about clustering dynamic data. Diptia Zandra [3] presented an automatic representative news beneration with applying an automatic clustering. The system use automatic clustering to determine the number of centroid automatically. The system crawls and cluster retrieved news every 3 hours. Another research related to news clustering is presented by Oren Zamir and Oren Etzioni [4] in their paper A Dynamic Clustering Interface to Web Search Result. This research they present Grouper, a clustering interface for web search engine result. Grouper creates cluster by merging base clusters (a phrase and the set of document that contain it). This is often beneficial but it can be confusing, especially when the clusters fail to capture the semantic distinctions the users are expecting. A research about discovering information among news site presented by George Adam et al search describes advaRSS, a crawling mechanism, which is created in order to support “peRSSonal”, a mechanism that produce personalized RSS feeds [5]. As the mechanism intend to be based utility for system offering collections of news article in real time to the Internet users, it has to maintain a fresh collection of the latest news. The algorithmic procedure is divided into two phases; they are training phase and crawling using the posting history. The training phase is utilizing simple metrics in order to fetch articles from RSS but in parallel it construct a change rate history of each article per hour and per day. While crawling the posting history is retrieving the posting history of an RSS by utilizing the information that are recorded during the training phase.

3. ORIGINALITY

This research presents a new system to generate automatic representative news using on-line clustering. The system determines single news for each cluster as its representative. News data obtained from 32news sites in Bahasa Indonesia. Vector quantization is implemented to update winning weight position based on new member position. A new cluster is

generated when an article's distance to the winning weight exceeds the threshold.

4. SYSTEM DESIGN

This system cluster news from RSS of 32 Indonesian online news sites. The process begins with news/article acquisition shown in diagram labeled by News enters. The article is extracted to obtain its keyword. After determining the keyword, the system then builds the metadata shown by metadata aggregation. Then the articles are clustered to obtain their centroids and the outermost members. This phase called predefined-cluster creation. The cluster generates centroids' position and the outermost members of each cluster. Then they become the basis for next the clustering process.

When the automatic clustering is done, the next process is online clustering. This process begins with articles/news acquisition and then applies text mining to the articles' content and extracts their keyword shown by label a new entry article. Then the next process is metadata aggregation, which every keyword is being aggregated in actual dimension. After that, the news then being clustered using VQ (Vector Quantization) to create dynamic cluster. Using VQ enables clusters to change centroids' position shown by centroid label update or it can create new clusters, shown by cluster label creation. Figure 1 shows the diagram of our proposed system architecture.

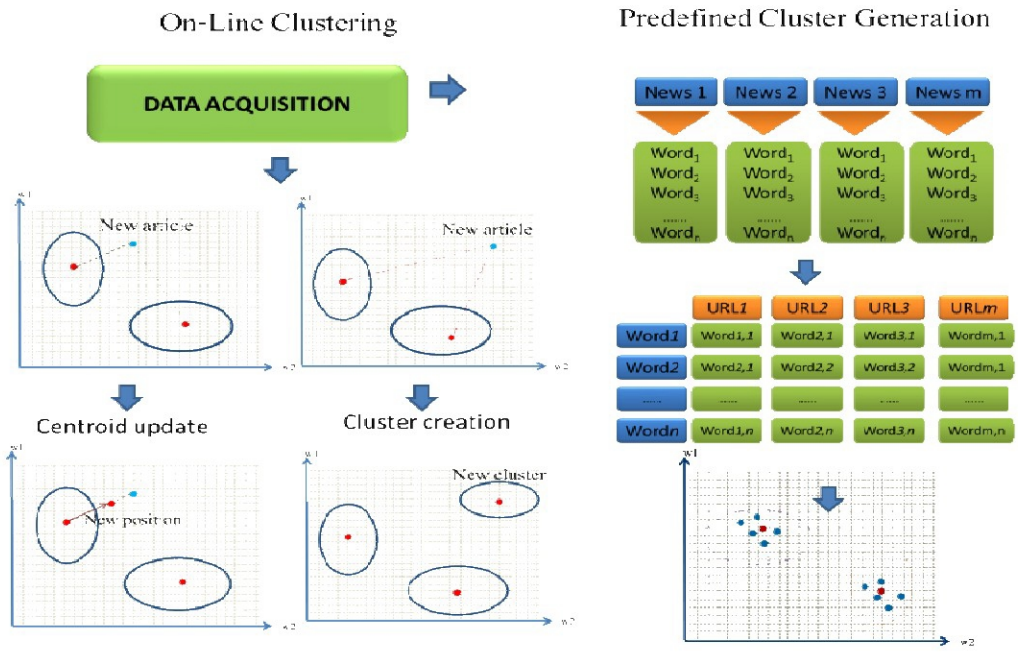

Figure 1. System architecture of our proposed system

4.1 Predefined Cluster Generation

The system workflow, shown in Figure 2, begins with news or articles acquisition. This process aims to collect online news from online news sites. News content isstemmedin order to find their keywords. Keyword extraction aims to get information from news or articles using text-mining process. We extract some of the most representative words (token words) and count the number of each token (token values). So the output of this process is tokens and the values for every article. These keywords then enter the metadata aggregation to aggregate the keywords length. The clustering in this step would be done manually. This cluster is become the predefined cluster for the next process (on-line clustering). The process then goes to metadata aggregation. Metadata aggregation works to generate the matrix of token value that will be clustered. This metadata contains URL and the number of token values of news.

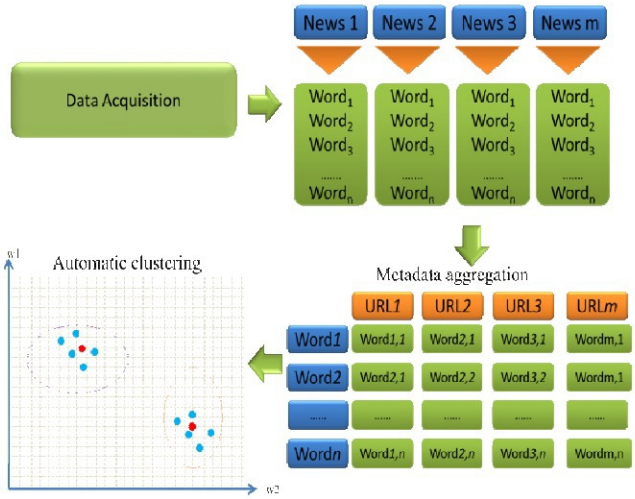

Figure 2. Predefined Cluster Generation

Then the URL is represented by token link. Token link represented by row while token word is number of columns. After defining the row and column, we input the token value depends on the token link and token words. If the token value has same source of token link and token word, the token value will be inputted. After that, it continues clustering them manually to obtain the winning weight of each cluster.

4.2 On-Line Clustering

After building predefined cluster, the next process continues to online clustering. Data acquisition process in in this phase would run continuously so when a news is obtained it would be calculated using VQ (Vector Quantization) to define the cluster it enters.

Figure 3 shows when there an article enters, it has two possibilities, enters a cluster or creates new cluster. When its distance to the closest centroid is less than the threshold, it enters the cluster. In It happens when

the article distance's to the centroids more than the treshold defined and consequently, the winning weight is updated using Vector Quantization. But when the distance to the closest centroid exceed threshold defined, it creates new cluster and itself is being the centroid.

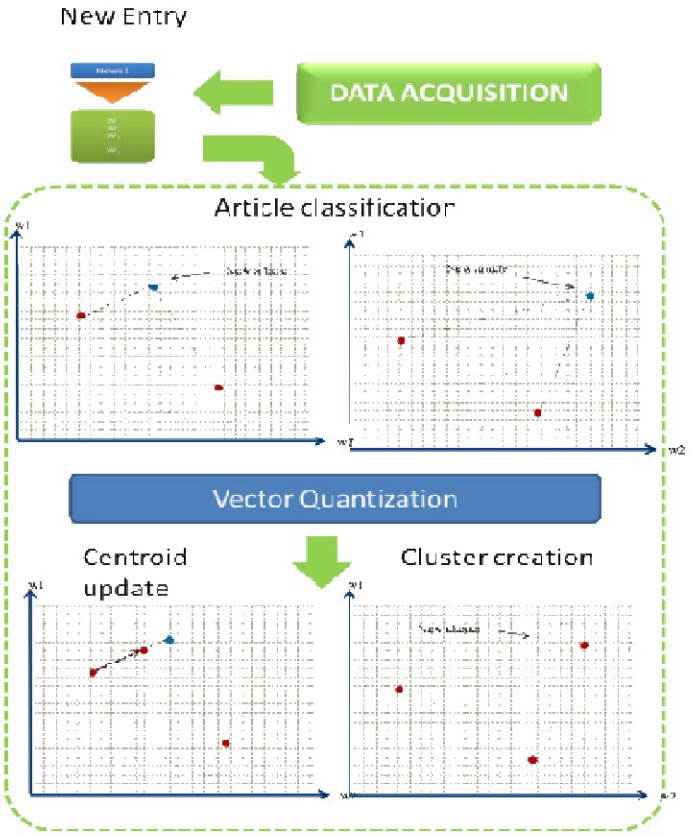

Figure 3. On-Line Clustering Workflow

4.3. Vector Quantization

Vector quantization is one of competitive learning methods. This method aims to discover the structure in the data by finding how the data is clustered [6]. The algorithm can be described as below.

1. Choose the number of cluster *M*
2. Initialize the prototypes $w_1^* \ldots w_m^*$ (one simple method is to randomly choose M vector from the input data)
3. Repeat until stopping criterion is satisfied
   - Randomly pick an input *x*
   - Determine the "winning" node k by finding the prototype vector that satisfies Eq. 1.

$$w^*_k - x \mid <= \mid w^*_i - x \mid \textit{(for all i)} \quad (1)$$

   - Update only the winning centroid weights according to Eq. 2.

$$w^*_k(new) = w^*_k(old) + m(x - w^*_k(old)) \quad (2)$$

Where w^*_k(new) is winning centroid weights after being updated from its old value w^*_k(old) and m is learning rate; having the value between 0 and 1. In the experiment we use $m = 0.5$.

5. EXPERIMENTS AND ANALYSIS

The news clustering uses news data from various online news sites in Bahasa Indonesia. Predefined cluster consist 500 articles/news those are divided into 50 clusters. Each cluster has maximum distance (Mi) that represent the distance from the centroid to the outermost member. It is used as threshold to determine new cluster creation. The clustering run incrementally uses offline data consists of 460 news in Bahasa Indonesia. The example of result is shown in Table 1 and Table 2.

Table 1. Example Result Obtained from On-Line Clustering

Id	Enter	New	Title	Cluster	Correct Cluster
14	√		Rekapitulasi Suara Bali tegang	2	Yes
15	√		Hasil Pilgub Bali Imbang	2	Yes
16	√		PDI-P Segera Gugat Hasil Pilkada Bali ke MK	2	Yes
11		√	Puspayoga Diminta Lengkapi Alat Bukti Gugatan Hasil Pilkada Bali	52	Yes
12		√	PDIP Pertanyakan Netralitas Polri dalam Pengamanan Pilkada Bali	54	Yes
13		√	MK Tolak Gugatan PDIP dalam Pilkada Bali	55	Yes
41	√		Justin Bieber Berniat Jalan-Jalan ke Luar Angkasa	34	No
64	√		Lacak Otak Penipuan Investasi Bodong via CCTV	13	No

The cluster label shows cluster id of each cluster. Experiment in the system performance that use 50 centroids from predefined cluster and 500 news test data shows that there are 42 new clusters creation or it can be said that 80% of data enter existing cluster. For example in figure 2, article with the id 14 enter centroid 2 and the status is b indicating that it enters the right cluster. Another example, article with the id 11 have a check (v) in New column; it means that the article creates new cluster with the centroid id 52.

The experiment shows the result that the precision is 70.9%. The examination on the experiment shows that the errors caused mainly by inadequate keyword of article. It means that input data has too little keyword such as only one keyword exist. When the system acquisites news from the Internet, it stems news content and determine news' keywords. The

keywords is terms from stemmed news content that appears more than 4 times according to an examination to define best keywords. But it failed in adapt to the length of news content. When the news is very short, the threshold can be inadequate.

Representative news is article that has the closest distance to its centroid. Based on our experiment of 50 news clusters, 46 of them have the right representative news while 4 of them have wrong representative news. In Table 2, 'B' shows the right ones in the column 'Status'. This result is strongly affected by clustering precision.

Table 2. The Example Result Obtained from Representative News Generation

No	Cluster Id	Representative News	Correct Cluster
1	2	Terkait Miss World 2014, Bali Belum Lakukan Persiapkan	Yes
2	3	BlackBerry Investigasi Penyebab Gangguan Jaringan BBM	Yes
3	4	Waktu Kenaikan Harga BBM Dinilai Tidak Tepat	Yes
4	5	Justin Bieber Akan Pergi ke LuarAngkasa	Yes
5	6	Pelaku bom Boston dipindahkan kerutan	Yes
6	7	BI Tunggu Respons Singapura Soal Syarat Akuisisi Bank	Yes
7	8	Adi Bing Slamet: Subur Itu Sudah Bau Tanah, Harusnya Tobat!	Yes
8	9	Peminat Palapa Ring Melonjak	No

6. CONCLUSION

We made a series of experimental study to see the performance of our system. The number of gathered news for 19 hours from 25 online news providers can reach up to 2118 different news. Based on this problem, we propose a system to cluster news using on-line clustering. The uniqueness of this research is the feature to update centroid position and create new cluster. The Vector Quantization is implemented to incrementally cluster news. Using 460 data as experimental purpose, our system performed 29.1% of the error rate

We examined the precision of clustering including the precision of new cluster creation. An article enters the right cluster if its title has close similarity to the members of the cluster. The article creates new cluster if there's no single article similar to the previous articles. The result shows that 269 of 383 articles enter right cluster while 124 of them enter wrong cluster. The errors are caused by 2 conditions:

1. When the keywords are too short, only one or two keywords exist in an article. For example article with the id 269 having the title 'Juara,

Marquez Puncaki Klasemen'. This keyword is not sufficient enough to represent news content and also leads to condition stated in point 2.

2. Some centroids has lots of features while the others have little features. This case happened for example in article with the id 64 that has the title 'Lacak Otak Penipuan Investasi Bodong via CCTV'. This article has the closest distance to the centroid 13 (C13) that has representative news 'Belum Diketahui Nasibnya 27 Karyawan Freeport' (the cluster talks about Freeport). The distance to the C13 is 19.28 while its distance to C10 that has the representative 'OJK Harus Menindak Kasus Investasi Bodong' is 20.91. The distances are calculated using Euclidean distance; it makes the number of features affect the result. C13 has 54 features (keywords) while C10 has only 21 features (keywords).

The result of creating new cluster shows 72 of 460 articles creates new clusters. There are 50 new clusters created. Some clusters have member one until four members. These members enter the right cluster except two articles with the id 11, 82, and 460. Overall experiment shows that keyword extraction that fails in generating representative keyword lead to the imprecise clustering result. Keyword distribution in each centroid is also determining the clustering result. Some cluster has big number of keywords while others just have a little amount of keywords. When the distance is calculated, centroid that has big number of keywords tends to result in higher distance value.

The system needs a better approach in keyword extraction, which gives a content-representative keyword. Some bad results might be obtained from keyword extraction affect the clustering results. To reduce the error of keyword distribution, the cosine distance metric is better to used for distance measurement of articles and centroids

REFERENCES

[1] Kominfo Pekalongan, **Pengguna Internet Indonesia BisaTembus 82 Juta**, *http://kominfo.pekalongankota.go.id*, Retrieved June 19, 2013.

[2] I. Moggi, **Daftar Situs Berita Online yang ada di Indonesia**, *http://www.speechmagazine.blogspot.com*, Retrieved May 13, 2011.

[3] Diptia Zandra Eka Puspitasari, Ali Ridho Barakbah, Idris Winarno, **Automatic Representative News Generation using Automatic Clustering**, *Industrial Electronics Seminar (IES) 2011*, Surabaya, 2012.

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

[5] A. C. George, **Efficient Extraction of News Articles based on RSS**. Computer and Informatics Engineering Department, University of Patras.

[6] Ali Ridho Barakbah, **Pursuit Reinforcement Competitive Learning: An approach for on-line clustering**, *The 2nd Information and Communication Technology Seminar (ICTS),* Surabaya, 2006.