EMITTER International Journal of Engineering Technology
ISSN: 2443-1168, Vol. 7, No. 2, December 2019, pp. 467~479
DOI : 10.24003/emitter.v7i2.378

Cluster-Based News Representative Generation with Automatic Incremental Clustering

Irsal Shabirin, Ali Ridho Barakbah, Iwan Syarif

Graduate School of Informatics and Computer Engineering
Politeknik Elektronika Negeri Surabaya
Jl. Raya ITS Sukolilo Surabaya 60111, Indonesia
Telp: 6231 5947280 Fax : 6231 5946114
E-mail: irsal@pasca.student.pens.ac.id, {ridho, iwanarif}@pens.ac.id

Received March 26, 2019; Revised May 8, 2019; Accepted June 21, 2019

Abstract

Nowadays we are facing much abundant information, especially news, and makes us confused in sorting out the information, so that it wastes our time in filtering that information. Though the news often contains similar contents that should save our time for reading. In this paper, we propose a new approach to provide aggregation mechanisms from cluster-based news and produce representative news, using our proposed Automatic Incremental Clustering. This approach presents a mechanism for clustering incremental news data and dynamically providing an automatic creation of new clusters. This approach consists of six main functions, which are (1) Data acquisition with incremental news sources from several news service providers, (2) Keyword extraction for term representation of news data, (3) Metadata aggregation for creating vector space of terms, (4) Automatic clustering for initiating news cluster generation, (5) Automatic incremental clustering for clustering incoming news data to pre-determined clusters or creating a new cluster of news data, and (6) News representation for selecting the most representative news of data clusters. For experimental study, we involved 95 news data service providers with 751 news data for for creating initial clusters with automatic clustering and 110 news data for incremental automatic clustering. Our approach performed 85.14% accuracy for incremental automatic clustering, and is able to dynamically create new clusters for incremental news data.

Keywords: Clustering, Metadata Aggregation, Automatic Incremental Clustering, Representative News

1. INTRODUCTION

The number of internet users in Indonesia is increasing rapidly. This is proven by statistics saying that as of June 30, 2012, internet users in Indonesia have reached 55 million compared to the year 2000 with only 2 million users [1]. The statistics of internet users throughout 2014 increased by 6% compared to the previous year. According to data released by APJII (*Asosiasi Penyelenggara Jasa Internet Indonesia*/Association of Indonesian Internet Service Providers), the number of internet users in 2014 was 88.1 million. This number has increased from 71.2 million in the previous year [2][3].

The growing number of internet users also has an impact on the current media. The mass media saw a shift or resolution to a more sophisticated direction, i.e. online media like news portals, as one of the primary needs to deliver information to present the latest news for internet users [1]. The number of news until June 12, 2012 reached more than 2000 news every day. There are some news with the same title or content loaded in different portals, leading to an increased number of news with the same content [4], which positions the readers in a trapped situation among non-representative news (news with the same content).

In this study, we try to provide a new approach using clustering science. This new approach aims to provide representative news using the Automatic Incremental Clustering algorithm. This research began with cluster formation initialization using the Automatic Clustering algorithm, then new news entered the system in real-time so as they could be combined with previously defined clusters or placed in a new cluster. Furthermore, this research would provide an exploration of facts and knowledge about the proposed method.

2. RELATED WORKS

Diptia et al [4] built a system that was able to group news automatically. The system was also able to provide news representatives from each group with the aim of overcoming news redundancy through the Automatic Clustering algorithm. The news used in this study was static or offline news, the collection of news that was gathered before. The results of this study were that the system could identify the number of groups and news or members of each group automatically and provide news representatives to users. Link index value had a significant influence on the use of Automatic Clustering algorithm. If the difference of links and clusters/groups was large, then the distance among clusters would be farther away. On the contrary, if the difference of links and clusters was too close, then the distance of each cluster was getting closer or less separated. The time interval from data acquisition also had a significant impact on the grouping process. The smaller the time interval used, the more news would be obtained and grouped.

Marlisa et al [5] built a system that was capable of automatically generating news representatives using online clustering. This system allowed grouping to be dynamic with time update features and creating new clusters/groups. At the experimental stage, researchers applied a system using

460 news in Indonesian. This experiment yielded precision ratio of 70.9%. This was caused by incorrect results of keyword extraction resulting in only one or two keywords for an article. Centroid keyword distribution (the center point of each group) also affected the grouping results.

Francesco Cambi et al. [6] analyzed two news agency websites and four of the most popular Italian newspapers. The researcher developed a web-based application that collects articles / news every hour by implementing a modified (Two-Layer) Incremental Clustering algorithm. It is used to find the most relevant news, how much the average time of the news stays in the news portal, and how much of the same news is posted on several news portals. The researcher uses the TF-IDF formula to form the features of each news. And to measure the distance between news using cosine similarity. The algorithm used by researcher turned out to be very efficient in terms of computational time and quite good in terms of precision and recall.

Siti Rofiqoh et al. [7] have done experimented in the field of topic detection on online news using K-Means with a mini batch algorithm. The simulation in this research uses nine news portals via RSS. The data obtained are the date of publication, the author, the title, and the first few sentences. Then each article through a tokenization process, the filter uses a stop word list, and finally uses TF-IDF for word weighting. On topic detection research, researcher interpreted the midpoint or centroid as the topic. The use of K-Means for large datasets will make the clustering process run slowly. However, with the approach given by the researcher, the computation of clustering process can be reduced by 30-40% with reduced accuracy around 2-4%.

Arun Kumar Sangaiah et al [8] have done experimented in the field of clustering related to Arabic documents / news. The Clustering process aims to improve understanding or summarizing a group of news. It starts with collecting data from EL-Watan News and EL-Jazeera News. The news is given one label according to human judgment which is divided into 5 categories. The next process is the term extractor, which consists of tokenizing, filtering, and weighting each word using TF-IDF. Then the authors grouped the news with four algorithms, convinced by K-Means, Incremental K-Means, Threshold + K-Means, and K-Means with Dimensionality Reduction. Calculation of distance between news uses cosine similarity. The proposed method shows better accuracy and fewer errors for the new classification test case. The dimensionality reduction process is very sensitive because it can eliminate important information from each news itself.

3. ORIGINALITY

The ease and breadth of Internet access usage on online media through news spread in Indonesia has made news presentation on a topic to be redundant or unrepresentative. The purpose of this study is to utilize the science of clustering to group news and present representative news in each group. This research proposes a new approach through combining Data Acquisition techniques, Keyword Extraction, Metadata Aggregation,

Automatic Incremental Clustering, and Representative News. Automatic Incremental Clustering includes two phases: first, forming the initial cluster by collecting some news and clustering them using the Automatic Clustering algorithm; second, the latest news was received in real-time, therefore the news could be combined with a cluster that had been previously defined or placed in a new cluster. Each cluster would be selected as a representative news through the clustering approach, i.e. the closest distance to the centroid.

4. SYSTEM DESIGN

There were 5 stages in this research: Data Acquisition, Keyword Extraction, Metadata Aggregation, Automatic Incremental Clustering, and News Representatives. The overall system design is shown in Figure 1.

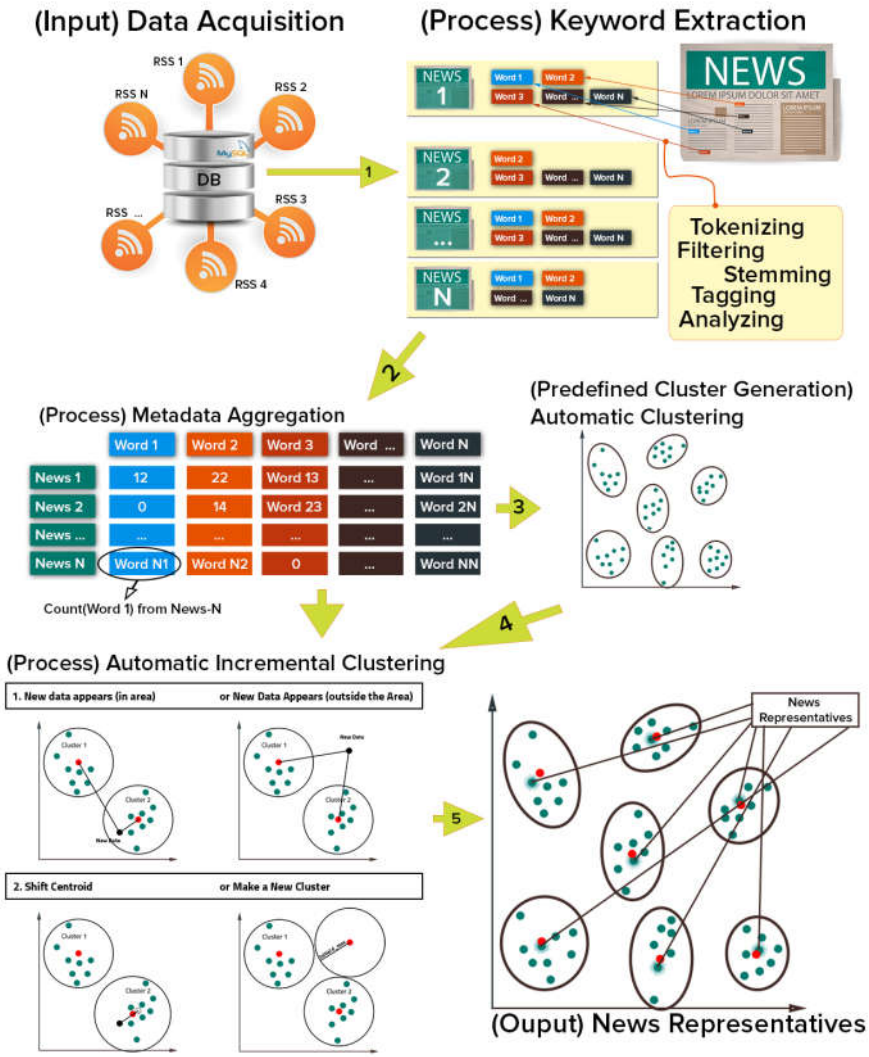

Figure 1. System Design

4.1 Data Acquisition

The proposed system began by getting an RSS feed list of open news portals. The next step was building a news crawling system from the RSS Feed to obtain 751 news from 95 news portals. Examples of RSS Feed links can be seen in Table 1.

Table 1. Example of RSS Link

No	Name	RSS Link
1	detik – Finance	http://rss.detik.com/index.php/finance
2	Suara.com – ALL	http://www.suara.com/rss
…	…	…
95	Suara Surabaya	http://www.suarasurabaya.net/rss/

The RSS Feeds showed results of news articles in the form of text but still contained html formats in them. The examples of news formats obtained can be seen in Figure 2. At this stage, the process of removing html tags was done in order to be processed further to keyword extraction.

```
<div cLass="text detail" readabiLity="86">
    <strong>Jakarta</strong> -Kondisi perekonomian global yang tengah melemah saat ini
    menjadi tantangan bagi banyak negara, termasuk Indonesia.<p>Menteri Keuangan Bambang
    Brodjonegoro mengatakan, lembaga dunia International Monetary Fund (IMF) baru saja
    merevisi kembali ke bawah pertumbuhan ekonomi global. Meskipun hanya direvisi 0,1%,
    tapi tendensi bahwa revisi ke bawah ini terjadi sudah berulang-ulang secara
    berturut-turut.</p><p>"Ini menegaskan kondisi ekonomi global ini jauh dari cerah atau
    kondisinya sedang <em>gloomy</em> atau suram. Ini terjadi di hampir semua negara di
    dunia yang ekonominya tergolong besar di dunia," ujarnya saat memberikan sambutan pada
    acara Sosialisasi Amnesti Pajak di Hotel Ritz Carlton, Jakarta, Selasa (26/7/2016).</p><
    p>Ia mengungkapkan, pola besarnya jarak antara fluktuasi/naik-turunnya harga saham pada
    saat ini berbeda dengan pola volatilitas pada masa lalu. Pola krisis keuangan global
    yang biasanya terjadi pada periode 5 hingga 10 tahun sekali saat ini tidak bisa
    diprediksi.</p><p>"Kalau kita bicara dulu mengenai krisis keuangan, apakah global atau
    regional, kita melihat in! <em>even</em> yang mungkin kejadiannya 10 tahun sekali, 5
    tahun sekali. Dan kalau pun ada ancaman sudah diketahui jauh sebelumnya. Itu pola
    sektor keuangan masa lalu. Dalam kondisi hari in!, ini menjadi faktor yang makin sukar
    untuk ditebak," tandasnya.</p><p>"Sekarang in! lebih <em>volatile</em> dibandingkan
    yang dirasakan krisis 1990 dan 2000-an. Ini terjadi setelah global <em>financial crisis
    </em> 2008. Di sini bisa kita lihat, global <em>financial crisis.</em> Itu yang membuat
    kondisi sistem perekonomian dunia menjadi <em> totally different</em>," tambahnya.</p><p
    >Pertumbuhan ekonomi global yang saat ini masih rendah, menurut Bambang, menjadi
    pertanda bahwa ekonomi dunia semakin sulit mencari sumber pertumbuhan.</p><p>"Apalagi
    kalau kita lihat negara per negara, tidak ada satu pun negara yang bisa menghindari
    volatilitas. Ini juga peringatan buat kita semua bahwa ekonomi Indonesia akan
    senantiasa berhadapan dengan volatilitas global," pungkasnya.</p><strong>(drk/drk)</
    strong>
</div><p>Redaksi: redaksi[at]detikfinance.com<br /><strong>Informasi pemasangan iklan</
strong><br /> hubungi : sales[at]detik.com </p>
```

Figure 2. Example of News obtained from RSS

4.2 Keyword Extraction

Keyword extraction was used to get information in the form of words from every available news. At this stage, the process was, in order: case folding, tokenizing, filtering, stemming, tagging, and analyzing. In this study, the analyzing process was carried out using Term-Frequency (TF), in which the repetition of words appeared in a news was calculated. Results from this stage are shown in Table 2.

Table 2. Example of keyword extraction result

Example of news article:	
...Tahun 2014 naik 8% menjadi Rp 205,123 miliar. Selanjutnya tahun 2015 total transaksi mencapai Rp 253,056 miliar atau naik 23%. Dan di 2016 transaksi KUPVA Bukan Bank mencapai Rp 257,479 miliar atau naik 2%. ...	
***Analyzing* result:**	
Words	**Total**
Miliar	3
Naik	3
Transaksi	2
tahun (stop word)	-
Atau (stop word)	-
…	…

4.2 Metadata Aggregation

Metadata Aggregation was the process of creating a two-dimensional matrix. The columns in the matrix were words appeared from all the news involved and each line was news. Thus, every word contained in a news would have a value based on the previous process (Keyword Extraction). If a news did not have a certain word, it would be given a value of 0 as shown in Table 3.

Table 3. Example of metadata aggregation result

	Word-1	Word-2	Word -3	Word -4	Word -N
News-1	0	1	2	1	...
News-2	2	1	0	0	...
News-N	...	...	...	...	...

In this study, not all keywords were entered into the metadata matrix. The sorting process was by calculating the number of keywords that must be greater than or equal to the determined threshold value. There were two types of threshold value used in this study; the first (*t1*) was worth two, and the second (*t2*) was obtained by calculating the highest number of words (*max_kata*) on each news divided by two (1).

$$t_1 = 2 \;\&\; t_2 = \frac{\max\_kata}{2} \tag{1}$$

4.3 Automatic Incremental Clustering

The next step was news grouping based on the obtained metadata matrix. This grouping process used Automatic Incremental Clustering as the method proposed in the study, which aimed to respond new data on existing clusters. The response was formation of a new cluster or becoming a member of a cluster. Therefore, this method required initial cluster formation of a metadata using Automatic Clustering.

Automatic Clustering was a clustering method used to form the initial cluster without defining the number of clusters to form. The number of clusters would be automatically identified/formed based on the data used [4].

In this study, the used data was metadata matrix, which was the result of news acquisition (offline) previously processed using metadata aggregation. This process would produce several news clusters and had a high dimension midpoint or centroid (multi-dimension) on each cluster.

After obtaining the initial cluster, the next step was the Automatic Incremental Clustering process. This process consisted of two stages, namely initializing the distance among d_{max} clusters and determining whether new data would be put in the initial cluster or in a new cluster. This determination process used the Vector Quantization approach. Overall, this process is illustrated in Figure 3.

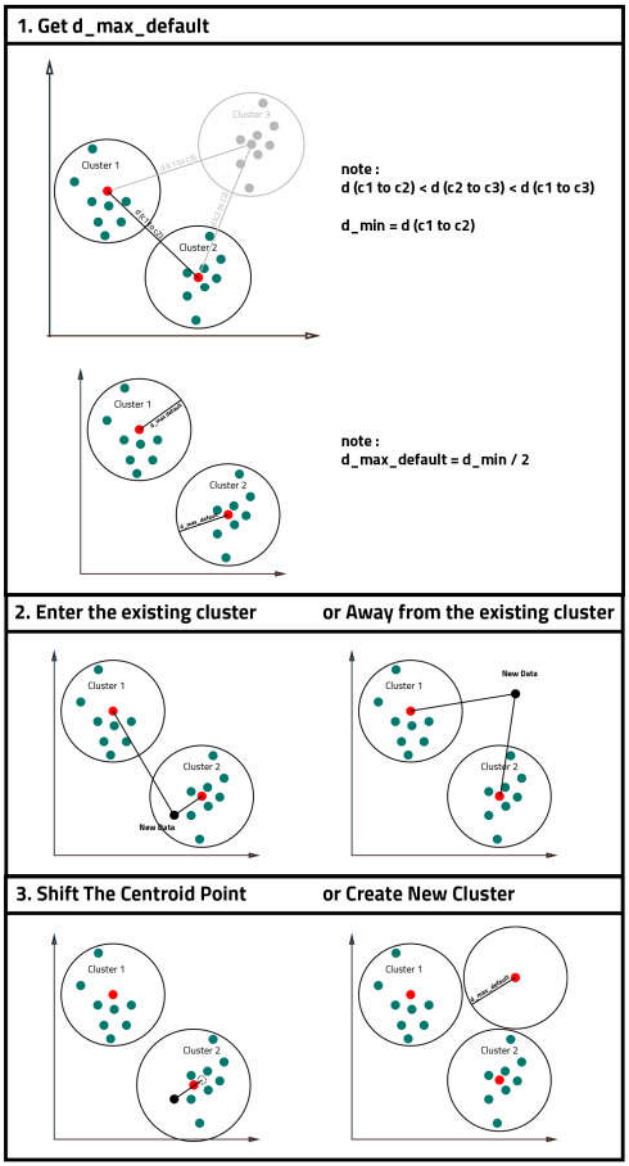

Figure 3. Illustration of Automatic incremental Clustering

The first process undertaken was the *d_max* Initialization process. This value was the value farthest from centroid as opposed to the cluster members inside. This value was obtained by finding the smallest value of distance among clusters, Automatic Clustering result, and division by two (2). On the first initialization, *d_max* value of each cluster ($d\_max\_c_1$, $d\_max\_c_2$, ..., $d\_max\_c_n$) was same and this value was used as *d_max* default on each new cluster (*d_max_default*). Then this value always changed or was continuously updated in accordance with new news entering into the built system (incremental).

$$d_{max} = \frac{\min\left(\sqrt[2]{(c_1)^2-(c_2)^2}, \sqrt[2]{(c_1)^2-(c_3)^2}, \ldots, \sqrt[2]{(c_m)^2-(c_n)^2},\ m \neq n\right)}{2} \quad (2)$$

Determining the response to new news was the next process. New news that entered this system had the following several stages:

1. Converting new news into news metadata format on the system. If there were new features, then the next process was to update the metadata by adding the new feature.
2. Calculating the distance of new news with the centroid on each cluster through the Euclidean Distance calculation method (dc_1, dc_2, dc_3, ..., dc_n).
3. Taking the shortest distance (*d_min*) of centroid on the cluster (c_x) against new news metadata. Then checking the process based on the condition of $d_{min} < (d\_\max\_c_x * \beta)$. Was the new news distance (*d_min*) shorter than the maximum centroid distance on the closest cluster (*d_max_cx*) with the multiplier β (threshold)? If yes, then go to number 5. If not, go to number 4.
4. Creating a new cluster, by initializing d_{max} with *d_max_default* value and the process of adding new news to the system had been completed.
5. The new news became a new member in the closest cluster.
6. Further, updating centroid cluster value on each feature/word (*C_new_w*$_n$, n was the first, second, ..., last feature/word) using Vector Quantization (3), namely processing centroid metadata value in each word (*C_old_w*$_n$) with a learning rate (*lr*, worth 1-10) and the word metadata value on the new news (*N_new_w*$_n$).

$$C\_new\_w_n = \frac{(lr-10) * C\_old\_w_n + (lr + N\_new\_w_n)}{10} \quad (3)$$

4.2 News Representatives

Representative news was news with the closest distance to the midpoint or centroid. Calculation of distance could be carried out using the Euclidean Distance formula.

5. EXPERIMENT AND ANALYSIS

Experiments and analysis consisted of the results of clustering on Automatic Clustering, Automatic Incremental Clustering, and News Representatives.

5.1 Automatic Clustering

In this study, there were 751 news from 95 news portals obtained. This data was processed through the keyword extraction and metadata aggregation to produce data in the form of a metadata matrix. Furthermore, these data were grouped using automatic clustering to determine initial cluster of the obtained data. This process generated 107 news clusters and the words used as features in each cluster are presented in Table 4. In the keywords column, the results show that features of the first cluster is the word "indonesia" as many as 60 news from 660 news (cluster member). If the number of keywords found was close to or equal to the total cluster members, then it was considered accurate. For example, in cluster_id of 0, this cluster had general keywords, such as "*Indonesia*", "*sebut*", and "2017". The word "*Indonesia*" was owned by 60 news, the word "*sebut*" by 49 news, etc. This indicates that the keywords generated as cluster features still can not represent 660 news because the number of keywords that appear in cluster members is still very small, as it can be seen in the word "*Indonesia*" at 60 compared to 660.

Table 4. Automatic clustering results

Cent-roid_id	Distance to zero	Number of members	Keywords
0	3.7247	660	indonesia (60), sebut (49), 2017 (43), main (40), laku (35), jakarta (31), anak (20), jual (19), presiden (19), negara (18), uang (17), perintah (16), menteri (16), harga (15), bangun (15), imlek (14), balap (13), tim (13), film (12), latih (12)
7	21.0631	9	yogyakarta (9), bandara (9), internasional (6), menteri (4), bangun (4), sebut (2), selesai (2), kapasitas (2), tumpang (2), pesawat (1), presiden (1), laku (1)
25	21.9829	4	indonesia (3), angpao (1), uob (1), imlek (1)

Different results were shown in cluster_id of 7. This cluster had the same number of keywords or close to the number of cluster members, e.g. the words "yogyakarta" and "*bandara*" appeared in each cluster member and the number of "*internasional*" was almost close to the number of cluster members. This indicates that cluster_id of 7 has three important keywords, namely "*bandara*", "*internasional*", and "yogyakarta".

Based on the results, cluster_id of 0 was a cluster with a very common keyword and only had a small portion of cluster members. Therefore, this cluster was not used in the next process of automatic incremental clustering.

5.2 Automatic Incremental Clustering

The Automatic Clustering process obtained the farthest initialization value among clusters divided by two (d_{max}). The value obtained in the process

was 10.520. This value was used as a threshold value for each cluster, whether the new news would create a new cluster or become a member of the existing cluster. If the distance of new news to a nearby cluster was less than d_{max} it would be entered in the existing cluster, otherwise it would create a new cluster. In addition, determination of new cluster was also influenced by the keywords generated from the news, alowing it to form a cluster, whether True or False, as the correct cluster. Some results can be seen in Table 5.

Table 5. Automatic incremental clustering result of new news

News id	Title	Cluster Exist	New Cluster	Correct Cluster
1916	Facebook, YouTube, Microsoft dan Twitter Bersatu Lawan Terorisme		164	True
1934	Wisatawan yang Liburan ke Pantai Gunungkidul Harus Baca Ini!		166	True
2756	Sevel Tutup, Kemenperin: Ekspansi Agresif Tapi Kurang Perencanaan	147		True
2897	Harga Tiket Masuk TMII Naik Rp 5.000 Selama Libur Lebaran	206		True
2898	Jokowi ke Susi: Sus, Jangan Terus-terusan Urusi Cantrang		221	True
2899	Jepang Garap Kereta Kencang JKT-SBY, Luhut: Pilihan yang Bagus		222	True
2916	Diskon Aneka Peralatan Kantor dan Alat Tulis di Transmart Carrefour		230	False
3088	5 Waterpark Ini Bisa Bikin Kamu Segar Usai Mudik Lebaran	150		False
2925	Aksi 12 Jam, Demo Sopir Truk Tangki Pertamina Akhirnya Bubar		237	False
2923	Jokowi: Kalau IHSG 6.000, Dirut BEI Jalan Kaki dari Mana ke Mana?		236	True
2931	Pemindahan Ibu Kota RI Akan Dimulai 2018?	239		True
3090	Bisnis Kedai Kopi Digandrungi Anak Muda, Ini Pemciunya	146		True
3152	Racikan Es Kopi Susu yang Enak dengan Harga Terjangkau	146		True
3153	Ruang pamer Jatim di Tianjin dibuka Agustus		266	True

For example, on news_id news with a value of 1916, a news entitled "*Facebook, YouTube, Microsoft dan Twitter Bersatu Lawan Terorisme*", was a new news that produced a new cluster. This happened because the news had a distance between clusters exceeding d_{max} at 13,047, while this news was categorized as True in the correct cluster column. Meanwhile, on news_id news with a value of 3088, an article entitled "*5 Waterpark ini Bisa Bikin kamu Segar Usai Mudik Lebaran*", was new news that did not produce a new cluster and

became a member of an existing cluster of cluster_id at a value of 150. In this news, the value in correct cluster clumn was False because the "waterpark" keyword was not found in the cluster.

Table 5 shows that there were 8 news in the existing cluster with True value and 6 news with False value, while the new news producing a new cluster with True value was at 78 news and False value was at 9 news. Based on the results, 86 of 101 news belonged to correct cluster with True value, thus the precision ratio value was 85.14%.

5.3 News Representatives

The Automatic Incremental Clustering process shows 70 clusters with inaccurate news representatives and 695 clusters with accurate news representatives. Accurate means that contents of the news representative could represent the entire news in a cluster, and on the contrary, inaccurate means that the contents of the news representative could not represent the entire news in one cluster.

The way to determine the news including representative news is to read it manually and read other news that is still in one cluster, to determine whether the news is representative or not.

Table 6. News representatives result

Cluster id	Title	News Representatives
25	Survei UOB: Rayakan Imlek, Masyarakat Belanja Lebih Banyak	False
575	Perpres Percepatan Mobil Listrik Indonesia Sedang Disusun	True
455	Pentingnya Advokasi Anti-Perundungan yang Menyasar Anak-anak	True
65	Pembangunan bandara Kulon Progo sudah dimulai	True
284	Liburan Sekolah di Jakarta Juga Bisa Asyik, Ini Destinasinya	True
321	Raih Tempat Kedua di Brno, Pedrosa Senang Repsol Honda Kompetitif	True
146	Bisnis Kedai Kopi Digandrungi Anak Muda, Ini Pemciunya	True
380	20 Kelakuan balita ini bukti jangan tinggalkan mereka main sendirian	False

In Table 6, the "News Representatives" column shows whether the news closest to the cluster center point could be used as a news representative. If the value in the "News Representatives" column was False, then the news could not be used as a news representative while True value indicated that the news could be a news representative.

6. CONCLUSION

This study conducts a set of experimental studies to verify the system performance. Automatic Incremental Clustering was very dependent on the previous process, namely initial cluster formation using the Automatic Clustering algorithm. As many as 107 clusters were formed out of 751 news. Each cluster had a varied number of members, ranging from one to 9. The average member of each cluster was 2 news, while there was one cluster with 660 news. This cluster contained general keywords such as "Indonesia" and "2017" and was close to zero, thus the cluster was not used in the next process.

The next stage was analyzing the accuracy of the Automatic Incremental Clustering algorithm. Observation of 110 new news resulted to a classification accuracy of 85.14%. Out of 110 new news, 87 news formed a new cluster and produced accuracy in cluster formation at 78 clusters. Inaccuracies of cluster formation/determination were caused by the news containing keywords that did not exist in the initial cluster.

This system requires a better approach in the Keyword Extraction stage. The use of Term Frequency formula is less appropriate to be used as a feature of each news. In addition, initialization process of clusters is also an important part of this system.

REFERENCES

[1] J. Efendi and S., **Perbandingan Nilai Berita Halaman Depan Portal Berita riauterkini.com dengan Portal Berita goriau.com**, Jurnal Online Mahasiswa, vol. 2, Februari 2015.

[2] E. L. Lukman, **Laporan: inilah yang dilakukan 74,6 juta pengguna internet Indonesia ketika online**, 31 October 2003. [Online]. Available: https://id.techinasia.com/tingkah-laku-pengguna-internet-indonesia. [Accessed on 24 Desember 2015].

[3] R. Nistanto, **Pengguna Internet Indonesia Tembus 88 Juta**, Kompas, 26 Maret 2015. [Online]. Available: http://tekno.kompas.com/read/2015/03/26/14053597/pengguna. internet.indonesia.tembus.88.juta. [Accessed on 24 Desember 2015].

[4] D. Z. E. Puspitasari, A. R. Barakbah and I. Winarno, **Automatic Representative News Generation using Automatic Clustering**, Industrial Electronics Seminar (IES), Surabaya, 2012.

[5] M. Sigita, A. R. Barakbah, E. M. Kusumaningtyas and I. Winarno, **Automatic Representative News Generation Using On-Line Clustering**, EMITTER International Journal of Engineering Technology, vol. 1, p. 107, 2013.

[6] F. Cambi, P. Crescenzi, and dan L. Pagli, **Analyzing and Comparing On-Line News Sources via (Two-Layer) Incremental Clustering**, 8th International Conference on Fun with Algorithms, FUN 2016, a Maddalena, Italy, 2016.

[7] S. R. Fitriyani and H. Murfi, **The K-means with mini batch algorithm for topics detection on online news**, 2016 4th International Conference on Information and Communication Technology (ICoICT), Bandung, 2016, pp. 1-5.

[8] A. K. Sangaiah, A. E. Fakhry, M. Abdel-Basset, I. El-henawy, **Arabic text clustering using improved clustering algorithms with dimensionality reduction**. Cluster Computing, 2018, 1-15.

[9] A. R. Barakbah and K. Arai, **Pursuit Reinforcement Competitive Learning: an approach for on-line clusterin**g, The 2nd International Seminar on Information and Communication Technology Seminar (ICTS), Surabaya, 2006.

[10] A. R. Barakbah and K. Arai, **Determining constraints of moving variance to find global optimum and make automatic clustering**, Industrial Electronics Seminar (IES), Surabaya, 2004.

[11] K. Arai and A. R. Barakbah, **Cluster construction method based on global optimum cluster determination with the newly defined moving variance**, Japan, 2007.

[12] A. R. Barakbah and K. Arai, **Reversed pattern of moving variance for accelerating automatic clustering**, EEPIS journal, vol. 2, p. 15, 2004.

[13] A. R. Barakbah and K. Arai, "**Identifying moving variance to make automatic clustering for normal data set**," IECI Japan Workshop, Tokyo, 2004.

[14] J. Asian, **Effective Techniques for Indonesian Text Retrieval**, RMIT Research Repository, Australia, 2007.

[15] A. Z. Arifin, I. P. A. K. Mahendra and H. T. Ciptaningtyas, **Enhanced Confix Stripping Stemmer and Ants Algorithm for Classifying News Document in Indonesian Language**, Proceeding of International Conference on Information & Communication Technology and Systems (ICTS), Surabaya, 2009.

[16] A. D. Tahitoe and D. Purwitasari, **Implementasi Modifikasi Enhanced Confix Stripping Stemmer Untuk Bahasa Indonesia dengan Metode Corpus Based Stemming**, Surabaya, 2010.

[17] A. Z. Arifin and A. N. Setiono, **Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering**, Seminar on Intelligent Technology and Its Applications (SITIA), Surabaya, 2002.