SITIA 2014

15th Seminar on Intelligent Technology and Its Applications

Volume 15
ISSN : 2355-8636

Thursday
May, 22nd 2014

Dept. of Electrical Engineering Building,
Institut Teknologi Sepuluh Nopember,
Surabaya, Indonesia

Full Proceeding

COMMITTEE OF SITIA 2014

Honorary Chaiman:
Prof. Dr. Ir. Triyogi Yuwono, DEA
Rector of Institut Teknologi Sepuluh Nopember (ITS)

General Chairman :
Dr. Ir. Yoyon Kusnendar Suprapto, M.Sc. (ITS)

Co-Chairman:
Ronny Mardiyanto, ST., MT., Ph.D. (ITS)

Technical Committee :
Prof. Dr. Mohammad Nuh (ITS)
Prof. Dr. Rukmi Sari Hartati(UNUD)
Prof. Dr. Ida Ayu Giriantari (UNUD)
Prof. Dr. Ontoseno Penangsang (ITS)
Prof. Dr. Ahmad A. Setiawan (UGM)
Prof. Dr. Mauridhi H. Purnomo (ITS)
Prof. Dr. Hanny H. Tumbelaka (UK Petra)
Prof. Dr. Abdullah Alkaff (ITS)
Prof. Dr. A.M. Shiddiq Yunus (PNUP)
Prof. Dr. Achmad Jazidie (ITS)
Prof. Dr. Mohammad Ashari (ITS)
Prof. Dr. Agus Uliniha (UMS)
Prof. Dr. Imam Robandi (ITS)
Prof. Dr. Parachai J (Sripatum University)
Prof. Dr. Adi Soeprijanto (ITS)
Prof. Dr. Gamantyo Hendrantoro (ITS)
Prof. Dr. Pei Yi Lim (UTS)
Prof. Dr. Iping Supriana (ITB)
Prof. Dr. Deddy Kurniadi (ITB)
Dr. Adel Elbaset (Minia University)
Dr. Engin Karatepe (Ege University)
Dr. Inggrid Nurtanio (UNHAS)
Dr. Arief Muntasa (Bangkalan University)
Dr. Lilik Anifah (Unesa Surabaya)
Dr. Syafaruddin (UNHAS)
Dr. Atris Suryantohadi(UGM)
Dr. Mochamad Hariadi (ITS)
Dr. Sentagi S. Utami (UGM)
Dr. Elyas Palantei (UNHAS)
Dr. Dian Savitri (UDINUS)
Dr. Rahmat Syam (UNM)

Organizing Committee:
Irwin Santoso S.
Mufiedah
Azmil Muftaqor
Dhuhari C.B
Hidayah
Fauzi Surya W.
Vigor Aryaditya
Surya Dwi K.
Ilham Laenur H.
Zulfikar Savero
Habibur Rohman
Novita Tristianti
Nafi Abdul H.
Hanif Fermanda
Dwi Haryanto
Hilman Andika
Wahyu Asrofi
Yudha Anughrah
Abdul Wakil
Fauziah Amin
Rakhmat O.
Roni Vayayang
Dylan Adhytia
Nugra A.
Mohamad Aziz
Fathan Nur H.
Naufal Rasviq
Taufani K.
Luqman H.
Viko Pujiantara
Nur Syarifuddin
Rahmadi R
Hitoshi Kusuma
Guntur Sadhiea
Tri Wahyu K.
Vincentius Raki
M. Fanani
Fachrul A.
Radifan Aiman
Novia Ayu I.
Fauziyah Amin
Adlia Difrianti
Nungki Dian S.D.
Sri Bayu Agus P.
Nani' L. Nada
Hasrul
Adit Dwinugraha
Lucky Andika N
M. Auliya Rasyid
Hadyan P.
Dhityo Y.
Reza Qashmal
Ana M. N.
Asa Femilsa
Charrel Naufal
Nindya A.F

ID: 303

Big Data Information Retrieval Based Search Engine Optimization Using K-Means Algorithm Optimization

Imam Marzuki[1)] Christyowidiasmoro[2)] Mochammad Hariadi [3)]

4) Master Programs in Department of Electrical Engineering, Faculty of Industrial Technology ITS Surabaya Indonesia 60111, email: imam12@mhs.ee.its.ac.id

5) Department of Electrical Engineering, Faculty of Industrial Technology ITS Surabaya Indonesia 60111, email: christyowidiasmoro@ee.its.ac.id

6) Department of Electrical Engineering, Faculty of Industrial Technology ITS Surabaya Indonesia 60111, email: mochar@ee.its.ac.id

- A search engine is said to be efficient if it optimized. Optimization here is the process of proving search engine performance in terms of accuracy and speed . There are many techniques have proposed in search engine optimization . But, that techniques still leave problems of the big data search. This is due to big data is dominated by unstructured data . Unstructured data have properties that difficult to organize. So , we need a special technique to overcome it. In this study , the authors try to propose a search engine optimization techniques using k-means algorithm optimization. The output of the system is sorted according to some relevance document. Measurement is solved by knowing the value of precision, recall , and the travel time of the rated documents . It is used to know the level of accuracy and speed of the search . The measurements give conclusion that the system has to provide optimal results to finding back the big data information .

***Keyword**: search engine optimization, big data, information retrieval, k-means algorithm optimization*

1.INTRODUCTION

The development of information technology rapidly make human activity can not be separated from the digital data. This data comes from various sources, for example from social networking sites and news portals. Additionally handheld devices to daily communication is also a source of digital data. Everyone can create and send digital data every second. Therefore, the digital data on the network has increased massively (large-scale) [1]. Massive increase in digital data is dominated by unstructured data such as text, image, audio, video, email, presentation slides, animations, etc.. A world body analyzing digital data has increased exponentially every years [2]. Look at Figure 1.

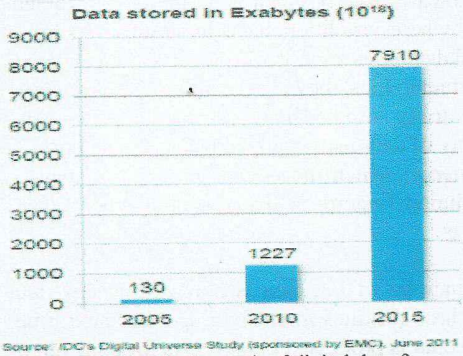

Figure 1. Predicted of growth of digital data from year to year

Figure 1 shows the increase of digital data in period 2005-2015 which rise exponentially. The development of these data has brought mankind to era of big data. One characteristic is indicated big data is unstructured data. Unstructured data does not have a relational hierarchy and may not be processed by relational databases.

Appropriate with the growth of big data with an abundance of unstructured data . Then we need fast and accurate techniques of information retrieval . This technique is called search engine optimization.

There are many techniques have been proposed in search engine optimization [3] [4] . But, that techniques still leave problems of the big data search. One solution that could be taken to solve this problem is add a clustering method into process of indexing information as in [6]. [6] Using the k-means method for clustering . overplus of the method application is clustering data and outliers quickly. But debility of this method is decrease of accuracy rate when it is used to process bigger information and still growing up. To make that debility not appear , then the k-means method needs to be optimized.

In this study , the authors try to perform an information retrieval techniques quickly and accurately from big data. In this system, a set of news comes from the Internet will be processed by text mining . The results of this process will be clustered using k-means

algorithm optimization method . So we will get a collection of documents are sorted according to their relevance levels .

2. LITERATURE REVIEW

This section will be discussed about the important theories that be a supporter and a reference for designing this study. The section covers the basic theory about big data, unstructured data, information retrieval, text mining, and the k-means algorithm optimization

2.1. Big Data

In a technical sense, big data is defined as a problem domain where traditional technologies such as relational databases are not able to serve. Big Data has three characteristics, they are the volume, the velocity, and the variety of data. Increase of volume, velocity and variety of data many happen because the adoption of the Internet which individuals produce content or leave a digital fingerprint that potentially be used to new things.

Some principles of big data is not wasting any data because these residues might be important next time. Next, the data is processed quickly. As for confront of the high variation in the data, big data create a structure by extraction, transformation, without having to discard the raw data before.

2.2. Unstructured Data

One of the challenges in processing of big data is unstructured data which is considered have not a relational hierarchy and not fit with traditional databases as Relational Database Management System (RDBMS). Some characteristics of unstructured data, among others, as follows:

1. Containing objects or documents that free size and free type of data.
2. Disorganized.
3. Organization and information are inconsistent.
4. Containing text, images, audio, video, email and powerpoint presentation.
5. The data is displayed on web page.

2.3. Information Retrieval

Information retrieval is part of computer science about take information from documents based on content and context of that documents. The reference explain that information retrieval is a search of information based on a query that is expected to satisfy desire of the user from the collection of documents. . Information or data that be sought is text , image , audio , video and others. Collection of data that also can be used as a search source is text messages , such as e - mail , fax , news documents , and documents on internet . With large capacity of documents collection as a search sources , then we need a system that can help users find relevant documents in a short time and accurate.

In terms of information technology there is a data retrieval , besides information retrieval . These two things are very different . Data retrieval in general determine the appropriate documents from a collection of data , the contents of the documents containing keywords in a user query , it will never be enough to satisfy information requirement of the user. Different with retrieval of data , users of information retrieval systems more attention in getting (retrieving) information by subject , rather than data retrieval based on a query is given, because the user does not want to know how that process is underway .

2.4. Text Mining

Definition of text mining has often given by many researchers and practitioners. As data mining, text mining is the process to find information that not revealed previously with process and analyze large amounts of data. In analyzing part or all of unstructured text data, text mining tries to associate one part with the other parts of the text based on certain rules. The expected result is a new information not previously revealed clear.

Text created not to be used by the machine, but for direct human consumption. Text mining has adopted the techniques used in the field of natural language processing and computational techniques in computational linguistics linguistics. Although the techniques in computational linguistics can be spelled forward and accurate enough to extract the information, text mining destination not only extract information. But to find patterns and new information that has not been revealed.

Text mining process includes tokenizing process, wordlist or stoplist and stemming.

1. Tokenizing
   Tokenizing is the process removal punctuation of sentences in the document to produce words that stand on their own.
2. Wordlist or stoplist
   Wordlist or stoplist is filtering of the words that are not feasible to be used as a differentiator or a keyword in process to search of documents and these words can be removed from the document.
3. Stemming
   Stemming is one of the manufacturing process retrieval system, where the stemming process will be carried out after the filtering process. This stemming process makes the term from tables filtering into base word, with remove all existing suffix in the word (affixes meng-, me-, kan-, di-, i, pe, peng-, a-, etc.)
   The Importance of Stemming in the

manufacturing process of retrieval system where remove affixes in a word, had to be considered. Because the important process of stemming process is remove the prefix after that the affix. When we try the opposite process, then the appropriate basis words and according to the dictionary will not find. From the results of the process will get an information about the number of terms that appear in a document after calculating term frequency.

K-Means Algorithm Optimization

eans clustering included in the partitioning which lso called exclusive clustering that separate data k separate areas and each data must belong to a rticular cluster and allows for any data that included ecific cluster on a stage of the process, the clusters ve to the next stage of the other. K-Means rithm is very popular because of the ease and ility to clustering big data and outliers data very kly. In K-Means, each data should be included a particular cluster in a stage of the process and ve to another cluster in the next stage.

eans Algorithm on Clustering can be done with following steps:

1. Define k as the number of clusters to be formed.
2. Generate k centroids (cluster center point) beginning at random.
3. Calculate the distance of each data into their respective centroids.
4. Every data choose closest centroids.
5. Determine the position of the new centroids by calculating the average value of the data that choosing the same centroid.
6. Go back to step 3 if the position of the new centroids with longer centroids not same.

racteristics of the K-Means algorithm is as ows:

- K-Means clustering is very fast in the process
- K-Means is very sensitive to the random generation of initial centroids.
- Allow a cluster does not have a member.
- The results of the K-Means clustering is not unique (always changing) - sometimes good, sometimes bad.
- K-Means is very difficult to reach the global optimum.

e weakness of the algorithm K-Means clustering is e result of K-Means algorithm depends heavily on itialization of the initial centroids randomly erated, therefore allowing for any data that cluded specific cluster on a stage of the process and ove to another cluster in the next stage. olutions to overcome the drawbacks of K-Means orithm that difficult to reach the global optimum is perform the optimization algorithm. That is by changing the point of determining new cluster center that obtained at random previously by defining a cluster center with the following algorithm:

If $X = \{xi \mid i = 1, ..., n\}$ is the data, and k is the number of clusters, then the stage of K-Means optimization:

1. Set C = as initial from the centroid to be formed
2. Set DM = [] as the number of accumulated distance of matrix.
3. Define m as the average of X
4. Calculate DM (X, m) as the distance of matrix between X to m
5. Set i = 1 as a first timer to determine the centroid
6. DM = DM + D
7. select xj ← ArgMax (DM) as ci
8. $C = C \cup ci$
9. Set DM (xj, C) = 0
10. Calculate again D (X, ci) as the distance matrix between X to ci
11. i = i +1
12. If $i \leq k$, go back to step 6
13. C is the solution as the initial centroid

3. RESEARCH METHODOLOGY

This section will discuss the system design and implementation. In this discuss there is a block flow diagram of the overall system and explanation of the detail processes is illustrated with a flowchart.

3.1 System Design

This research will be applied use clustering to classify documents containing specific information from a collection of documents directly retrieved from the Internet on several sites that have been included. From the retrieval process document of the internet , then the system will save the documents at a particular site is taken online and then in a txt format in a directory . Furthermore txt files in the folder experiencing text mining process consists of 3 stages which include tokenizing , wordlist or stoplist and stemming. After generating a collection of words of text mining results it will be known how many existing keywords in each document . From the each amount then performed clustering with k -means algorithm optimization of the coordinates on the point which shows the number of keywords each these documents . Each cluster document value is calculated using matrix multiplication . After the comparison between the value of each cluster , a cluster of documents that have highest value is the number of clusters selected documents as a result of a search for documents according to keywords. To display the results in a web page with find the value of each document by multiplying each value matrix of the number of keywords in each document with matrix transpose then displays the results in order of documents that have a number of keywords at the most until the least.

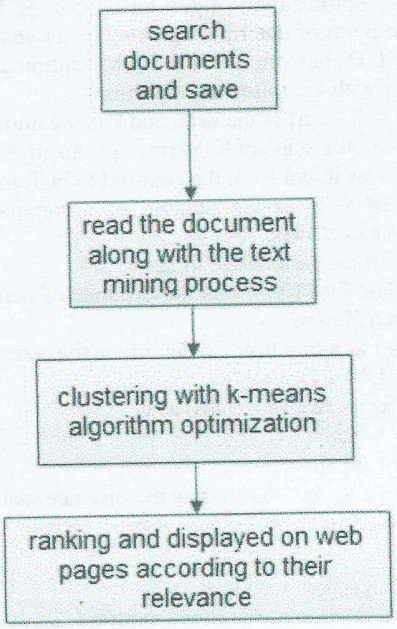

Figure 2. Block diagram of the global research

3.2 Analysis of Results

To evaluate the results, we need a measurement. In This Section describes how to measure the results. There are several measures is used in this research, that is the precision recall, and travel time.

3.2.1 Precision

Precision is ratio on the number of relevant documents obtained by the system with total number of documents is picked up by the system either relevant or irrelevant. According to this definition can be expressed in equation (1).

$$P = \frac{TP}{TP + FP} \quad (1)$$

Where :
P (Precision) = The level of precision search
TP (True Positive) = Relevant documents found
FP (False Positive) = Irrelevant documents found
TN (True Negative) = Relevant documents were not found
FN (False Negative) = Irrelevant documents that can not be found

3.2.2 Recall

Recall is ratio on the number of relevant documents obtained by the system with sum of all relevant documents in the collection of documents (drawn or not drawn by the system). According to this definition can be expressed in equation (2).

$$R = \frac{TP}{TP + FN} \quad (2)$$

Where :
R (Recall) = Recall level search

3.3.3 Retrieval Time

Measurement of the retrieval time is solved by knowing the retrieval time between the input and output

4. RESULTS AND DISCUSSIONS

For testing done by taking documents from several documents related to the keyword "scoring" and documents that are not related to the keyword. Testing is solved by two methods: the k-means algorithm and k-means algorithm optimization. It aims to find out the optimization that occurs when using the k-means algorithm optimization

Table 1. Testing the k-means algorithm

n	K	P	R	T
50	mencetak gol	73 %	80 %	0.5 sec
100	mencetak gol	75 %	78 %	0.5 sec
200	mencetak gol	70 %	79 %	0.5 sec
400	mencetak gol	78 %	76 %	0.5 sec

where :
n = Number of Documents
K = Keyword
P = Precision
R = Recall
T = Retrieval Time

Table 2. Testing the k-means algorithm optimization

n	K	P	R	T
50	mencetak gol	82 %	90 %	0.5 sec
100	mencetak gol	85 %	93 %	0.5 sec
200	mencetak gol	90 %	95 %	0.5 sec
400	mencetak gol	92 %	96 %	0.5 sec

where :
n = Number of Documents
K = Keyword
P = Precision
R = Recall
T = Retrieval Time

From Table 1 and Table 2 above we can see that there are differences in the test with k-means algorithm and k-means algorithm optimization. Especially in terms of precision and recall. In testing the k-means clustering algorithm (Table 1) a decline in the value of precision and recall on a greater number of documents. This decrease was due to the determination of the centroid randomly generated. While the test with k-means algorithm optimization (Table 2) an increase in the

of precision and recall for more data. This means the system is accurate enough for victory when an dance of data is increasing.

rms of retrieval time required by the search engine rithms k-means and k-means algorithm ization is no difference. This shows that the speed -means has not changed despite optimizations.

5. CONCLUSION

d on the testing and analysis of results can be luded:

1. Level of precision and recall documents with the K-Means algorithm without optimizations can vary depending on the determination of the centroid randomly generated.
2. Levels precision and recall of document clustering results with k-means algorithm optimization is more optimal than the k-means algorithm without optimization.
3. Necessary travel time clustering using k-means algorithm with the k-means algorithm optimization shows the same value as the amount of data increases. This means that the speed of k-means has not changed though has undergone optimization and increasing the amount of data.

REFERENCES

[1] Gantz, John Gantz. Diverse Exploding Digital Universe. *IDC*. [Daring] Maret 2011. http://www.emc.com/collateral/analystrepots/diverse-exploding-digital-universe.pdf
[2] Hariadi, Mochammad, "From Infrastructured To Analytics", Telematics Lab, Multimedia Network Dept. ITS, 2013
[3] K.K. Kattamuri, R. Chiramdasu, "Search Engine With Parallel Processing And Incremental K-Means For Fast Search And Retrieval", *International Journal of Advances in Engineering & Technology, Jan. 2013*
[4] Cahyono, H.D (2013), "Temu Kembali Citra dan Teks Dengan Pencarian Tekstual Berbasis Information Gain, Latent Semantic Analysis, dan Wieghted Tree Similarity", Master Program of Informatics Engeneering ITS, 2013
[5] Barakbah, A.R., Arai, K., "A New Algorithm For Optimization Of K-Means Clustering With Determining Maximum Distance Between Centroids", *In. IES 2006, Politeknik Elektronika Negeri Surabaya, ITS.*
[6] K. Supreet, K. Usvir, "An Optimizing Technique for Weighted Page Rank With K-Means Clustering", International Journal of Advanced Research in Computer Science and Software Engineering, July 2013
[7] Ridho Barakbah, Ali, Clustering, Soft Computing Research Group EEPIS-ITS, Surabaya 2006