Text grouping in patent analysis using adaptive K-means clustering algorithm

Tiara Shanie, Jadi Suprijadi, and Zulhanif

Citation: AIP Conference Proceedings **1827**, 020041 (2017); doi: 10.1063/1.4979457
View online: https://doi.org/10.1063/1.4979457
View Table of Contents: http://aip.scitation.org/toc/apc/1827/1
Published by the American Institute of Physics

Text Grouping in Patent Analysis using Adaptive K-Means Clustering Algorithm

Tiara Shanie[a)], Jadi Suprijadi[b)], and Zulhanif[c)]

Department of Statistics, Mathematics and Sciences Faculty of Universitas Padjadjaran, Bandung, Indonesia.

Corresponding author: [a)]tiarashanie13@gmail.com
[b)] jadisuprijadi@gmail.com
[c)] dzulhanif@gmail.com

Abstract. Patents are one of the Intellectual Property. Analyzing patent is one requirement in knowing well the development of technology in each country and in the world now. This study uses the patent document coming from the Espacenet server about Green Tea. Patent documents related to the technology in the field of tea is still widespread, so it will be difficult for users to information retrieval (IR). Therefore, it is necessary efforts to categorize documents in a specific group of related terms contained therein. This study uses titles patent text data with the proposed Green Tea in Statistical Text Mining methods consists of two phases: data preparation and data analysis stage. The data preparation phase uses Text Mining methods and data analysis stage is done by statistics. Statistical analysis in this study using a cluster analysis algorithm, the Adaptive K-Means Clustering Algorithm. Results from this study showed that based on the maximum value Silhouette, generate 87 clusters associated fifteen terms therein that can be utilized in the process of information retrieval needs.

Keywords: Patent Data, Green Tea, Statistical Text Mining and Silhoutte Width Index Measure

INTRODUCTION

Tea has been known belong to Indonesian people, and already entrenched in the lives of Indonesian people. Green tea is one of the most popular drinks and millions of cups consumed worldwide [1]. In addition to being beverages, green tea can contribute to the medical world. Green tea processing takes the role of technology, in an effort to save time as well as the expected content of interest in green tea. Declining agroindustrial Indonesian tea today occurs because not able to cope with the problems faced by Indonesian tea, low crop productivity due to the dominance of the tea plant people who are not yet using superior seed, limited mastery of technology products processing and yet inability of farmers to follow recommended technology as recommended (Good agriculture Practice / GAP; Good Manufacture Process / GMP) as well as product quality standards as required by the ISO [2].

Twiss(1992) said that one of the factors that affect the ability of a company are innovation, availability and management of information resources and knowledge both from within (internal) or outside (external) companies [3]. Accordingly, information concerning developments in technology and products in the fields of the tea industry of various countries need to be collected and analyzed to support efforts to advance technology in the field of the tea industry in Indonesia. Relevant information in this context is patent documents.

The information contained in patent documents is a title, inventor, applicant, international classification, abstract and so on. In this study, researchers focused on the patent titles to be analyzed. The title of the patent consist of text

Statistics and its Applications
AIP Conf. Proc. 1827, 020041-1–020041-9; doi: 10.1063/1.4979457
Published by AIP Publishing. 978-0-7354-1495-2/$30.00

data, making analysis of the patent can not use quantitative methods directly. So this research will combine Text Mining and Statistics, the Statistical Text Mining [5]. Terms of the title that appears to provide a set of patent documents related to the technology in the field of tea that is still widespread, so it will be difficult for users to information retrieval (IR). Therefore, it is necessary efforts to categorize documents in a specific group of related terms contained therein.

Non-hierarchical Clustering Techniques classic Statistics analysis commonly used in text mining is K-Means Clustering method. This study did not have a basis for determining upriori groups in grouping documents, so that researchers intend to give a new algorithm that does not directly specify the number of clusters by researchers. In this study the authors apply the expansion of K-Means Clustering called Adaptive K-Means Clustering in getting a group of documents related to information retrieval needs of the technology in the field of green tea industry is based on the titles of patent documents Green tea.

PATENT DATA

Patents is one of Intellectual Property (IP) type, that refer to creation of the mind such as inventions. The patent system is based on the rules to gain a monopoly over an invention. Patent rights available to those who wants to apply for a patent, with at least 18 months after the submission. Ernst (2003) said that patents are representatives of the technological innovations of a country or an organization and are indeed an agreement between the inventor of the patent and government or any agency designated by the government [6]. Patent documents consist of structured and unstructured data. The first page of the patent document comprising structured bibliographic data including the title and abstract which categorized as unstructured data [4]. Structured data are contain of Number and Dates, Assigness, Inventor and Classifications. Furthermore, patent analysis is useful to identifying the future technological trends in a specific field of technology, to promote ideas and to view the background of an invention [6] [7].

STATISTICAL TEXT MINING

Text mining is used to process text data through retrieval and indexing. While Statistics is used to summarize, visualize numerical data, and infer the population based on the estimation and hypothesis testing. Statistical approaches Text Mining hereinafter called STM, broadly divided into two stages of data preparation and data analysis. Preparation of the data include document pre-processing and feature selection, which is done through text mining methods. While the data analysis stage, divided into two basic analysis and advanced analysis includes descriptive statistics and multivariate analysis. Both of these steps will be explained as follows.

1. Preparation Data

The preparation phase of data used to transform unstructured data into numeric data are ready to enter the analysis phase. This stage is the task of text mining which consists of two steps, namely document pre-processing and feature selection. Document Pre-Processing is a step in the Text Mining to transform data into a format that the process is more easily and effectively to the needs analysis. There are three stages in Document Pre-Processing [8], namely:

1. Tokenizing is the decomposition of the original description in the form sentences into words, changing all words to lowercase, and eliminate a period (.), Comma (,), spaces and numeric characters that exist in the word [8].
2. Filtering is the process of selecting tokens that are considered unimportant or stopword. Stopword is a vocabulary that is not a characteristic (unique word) of a document [9].
3. The process stemming aim to eliminate prefixes and suffixes that exist in every word [8].

Feature selection stage is one of the important functions provided by this process is to be able to choose any term or word that can serve as an important representative for the document to be analyzed. The method can be used is the Term Frequency - Inverse Document Frequency with the following formula [10].

$$w_{i,j} = tf_{i,j}\ x\ log(\frac{N}{df_i}) \quad (1)$$

With :

$w_{i,j}$: The weight for the i term in the document j

$tf_{i,j}$: The number of times a term to-i in document j

N : The number of documents in the database
df_i : The number of documents containing term i

2. Data Analysis Phase

STM approach in conducting statistical analysis of the data consists of two types of basic analysis and advanced analysis. The basic analysis is based on the Statistical Text Mining categorized into two types: Statistics summary and meta-analysis. Broadly speaking, at this stage the data summary and visualization of the patent, including the visualization of the technology sector and the choice of words or terms using Zipf curve. While further analysis or statistical analysis of structured data that can be done is find an association, regression, path analysis, and clustering text or document. The use of advanced analysis adjusted for purposes of research on the text data, which in this study wanted to cluster or group obtained the titles of patent documents related to terms that arise from patent data green tea that has been the object of this study.

ADAPTIVE K-MEANS ALGORITHM

Adaptive K-Means Algorithm is a modification of the algorithm K-Means. Adaptive K-Means Clustering identified K clusters by determining the initial threshold at the beginning. The purpose of this algorithm is based on a range of clusters to build a better partition when a new element should be added. There is trouble in implementing the K-Means method on big data to determine the number of clusters randomly. In addressing these issues, Adaptive K-Means recommended its use. Adaptive K-Means Clustering Algorithm, the number of K clusters will change dynamically depending on the data and determined threshold value [11]. This algorithm aims to minimize the sum of squared errors. The process of adding a cluster with the iteration will continue to run when the value of the difference SSE each cluster pair is greater than a threshold value, and the iteration will stop when the differences of each pair SSE cluster is smaller than the threshold value. Therefore, if the threshold value is small, the more clusters are formed, but if the threshold value will be fewer, large clusters formed.

CLUSTER VALIDATION

Measurement of cluster validation is one of the steps that need to be done after the clustering analysis. The validity of the cluster is done by evaluating clustering algorithms are used, so that it can be seen that form clusters matches the natural partition. This study uses an internal validation. This is because the dataset that is owned in this study do not have the knowledge or information that is already known in advance, but uses the information residing in the data. There are three types of indices to determine the optimal cluster of internal validation, one of them is Silhoutte Width Measure Index [12]. The resulting value of Silhoutte Width Measure Index is the result of measuring the degree of confidence in determining the clustering of the data they hold. Value close to 1 indicates the sample has been well clustered, and the best number of clusters. If silhouette value of nearly zero, this indicates that the sample does not correspond to the closest group. Then, if the silhouette value -1 figures show the sample produce clusters wrong or "misclassified". So it can be said that the maximum silhouette value shows the best clusters. The equation used to calculate Silhouette Width Measure Index are as follows.

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)} \quad (2)$$

$$a_i = \frac{1}{n(C(i))} \sum_{j \in C(i)} dist\ (i,j) \quad (3)$$

$$b_i = \min \sum_{j \in C_k} \frac{dist\ (i,j)}{n(C_k)} \quad (4)$$

Where :

a_i = The average distance between i and all other observations in the same cluster
b_i = The average distance between i and observations at the nearest neighbor clusters.
$C_{(i)}$ = Cluster containing observations i.
Dist(i, j) = Distance between observation i and j.
$n_{(C)}$ = Cardinality of cluster C

ZIPF CURVE

Zipf curve is a tool to describe the frequency of a set of words. Zipf curve describes the dispersion of "term" arranged from the greatest frequency of occurrence of the term. Quoniam (1992) said that Zipf curve can repartition into three different zones[13], namely:

1. Zone 1- Trivial Information

Words or terms are included in this zone has the trivial or the information defined as a word that belongs to the central or main theme.

2. Zone II - Interesting Information

Based on Zipf curve, words or terms that are in this zone peripheral topics that describe the innovative potential of information.

3. Zone III - Noise Information

Last Zone in Zipf curve is the term the noise zone. Term contained in this zone can be said is not included in the concept of the new emerging technologies. In this zone can not be said whether the term will be a new idea or concept into a term that just noise or annoying.

Implementation of Zipf curve partition is to select words that are very important among a large number of words that are in the whole document. In this case the words are considered to be very important is the words that are in zone II (Interesting Zone). However, the justification of its importance is the domain expert.

RESEARCH METHODOLOGY

This study uses a Text Mining Statistical methods, wherein the data preparation phase is done by the method of text mining and data analysis using statistics analysis is clustering text. These steps are described in detail in Figure 3.1 below.

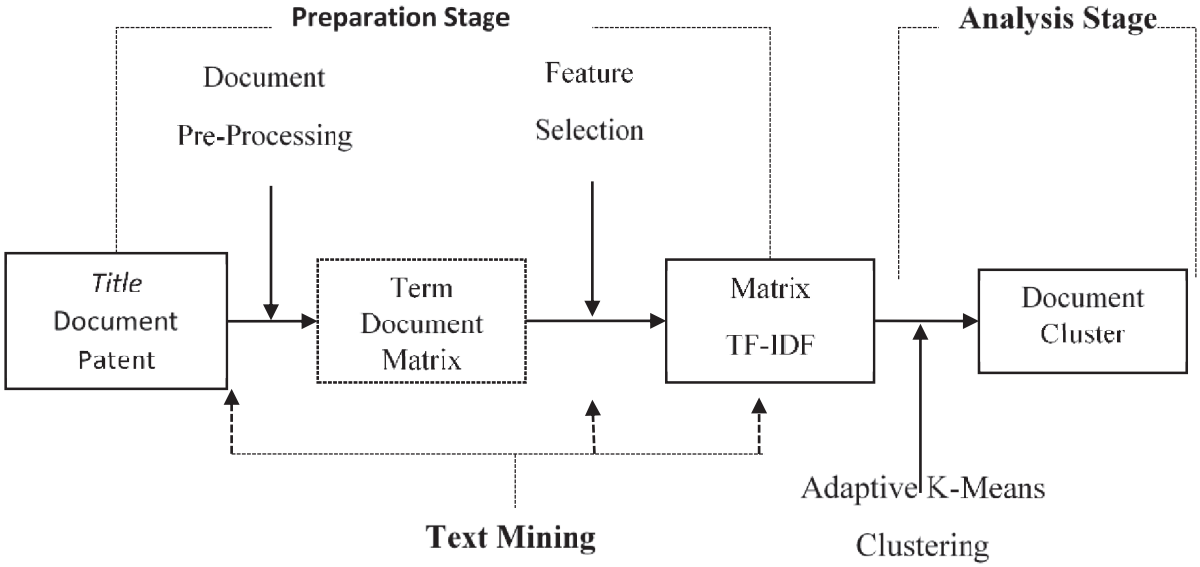

FIGURE 1. The chain grouping document of green tea patent data

Based on figure 1, analysis steps can be carried out as follows.

1. Obtain the titles of patent documents downloaded green tea on Espacenet server.
2. Document Pre-Processing phase which consists of tokenizing, filtering, and stemming thus obtained Term Document Matrix is a matrix of the frequency of occurrence of each word in each document.
3. Feature Selection phase by using Term Frequency-Inverse Document frequency, so we get a superbly weighted matrix.

4. Perform data analysis stage using Adaptive K-Means Clustering methods with R software on akmeans packages, so we get the following document groups term characteristic. This algorithm is described as having a data set X with n and p-dimensional objects as well as C1, C2, ..., Ck is a separate cluster k of X. Stages on analysis using Adaptive K-Means Clustering is as follows [11]:

1. Determine the number of clusters k minimum desired.
2. Determine the threshold value.
3. Allocate each data object into the nearest cluster to form a new partition.
4. Calculate the centroid or average of the data the new partition.
5. Calculate the distance between each object to each centroid using the Euclidean distance as follows [14]:

$$d_{ik} = \sqrt{\sum_{j=1}^{J}(y_{ij} - y_{kj})^2} \quad (5)$$

Where:

d_{ik} = Euclidean distance between object i with k

y_{ij} = i-th object in the j variable

y_{kj} = all k object of the j variables

6. For each cluster calculated value of SSE with the following functions:

Sum of Squared Error equation is as follows:

$$SSE = \sum_{k=1}^{K} \sum_{x_{i \in C_k}} \|x_i - c_k\|^2 \quad (6)$$

Where :

$x\hat{}_i$ = value or data on the object I

c_k = Centroid on cluster Ck

7. Compare the value of SSE each partner cluster. If the difference between the value of SSE each pair of cluster exceeds a threshold value, will increase the number of clusters, and the most distant object in the cluster will be elected as the new cluster.
8. Repeat steps 4 s.d 7.
9. Calculation is finished if the difference between the value of SSE each pair of clusters smaller than the threshold value for the whole cluster pairs, and it is the optimal number of clusters.

5. Validation cluster using Silhoutte Width Measure Index, where the optimum number of clusters determined based on the maximum value or the greatest silhouette shown.

RESULTS AND DISCUSSIONS

This research use title text patents green tea data totaled 421 patent documents from Espacenet servers using Adaptive K-Means Clustering. The discussions were held on the data analysis at every stage of the Statistical Text Mining as follows.

Data Preparation phase

Researchers must first choose the technology sector based patent document green tea are downloaded with the keyword "green and tea" in the title. Here are the results of the frequency of each of the technology sector in the form of the Pareto chart.

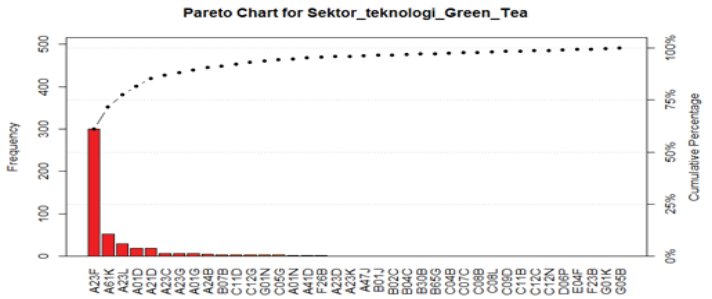

FIGURE 2. Pareto chart of technology sectors of green tea patent data

Based on the picture above, A23F sector has the highest frequency. A23F technology sector is engaged in tea and coffee drinks, so this field does provide benefits of green tea as a beverage or as a primary function. Frequency of the

patent document green tea A23F sector to A61K sector is very far, but A61K sector has the highest frequency among the other sectors after A23F. This has been an interest of researchers to choose the A61K technology sector. The technology sector fields used of A61K are preparations for medical, dental, and toilet pupose. Researchers re-downloading patent documents with the keyword "green and tea" in the title and add search criteria, namely the technology sector "A61K".

Document Pre-Processing

The results of this phase matrix form called Term Document Matrix (TDM), wherein the value that appears is the frequency of each term that appears in every document. Term Document Matrix consists of terms as row and column documents. Here is a TDM already in the filter using sparse value of 0.95, meaning that allow the zero value as much as 95% of all the data, which is described in Table 4.1.

TABLE 1. Term Document Matrix of Green Tea Patent Data

Terms	The Document of Green Tea Patent							
	Doc 1	Doc 2	Doc 3	Doc 4	Doc 5	…..	Doc 420	Doc 421
Composition	0	0	1	1	1	…..	0	0
Comprising	0	0	0	0	0	…..	0	0
Cosmetic	0	0	0	0	0	…..	0	0
Extract	1	1	1	1	1	…..	0	1
Extracts	0	0	0	0	0	…..	0	0
Ginseng	0	0	0	0	0	…..	0	0
Method	1	1	0	0	0	…..	1	0
Preparation	0	0	0	0	0	…..	0	0
Preparing	0	0	0	0	0	…..	0	0
Skin	0	0	0	0	0	…..	0	0
Thereof	0	0	0	0	0	…..	0	1
Treating	0	0	0	0	1	…..	0	1
Treatment	0	0	0	0	0	…..	0	0
Using	0	0	0	0	0	…..	0	0

Based on table 1, there are several documents that do not have the associated frequency fifteen elected term, therefore do filtering of such documents. Documents that do not have a frequency related term in figure 4.1 will be removed to avoid the formation of groups on the zero value. This step makes the data amounted to 351 titles of patent documents green tea.

Zipf curve contributes in describing the terms that may be a concern of researchers. This curve divides the terms or words into three zones that can be described as follows.

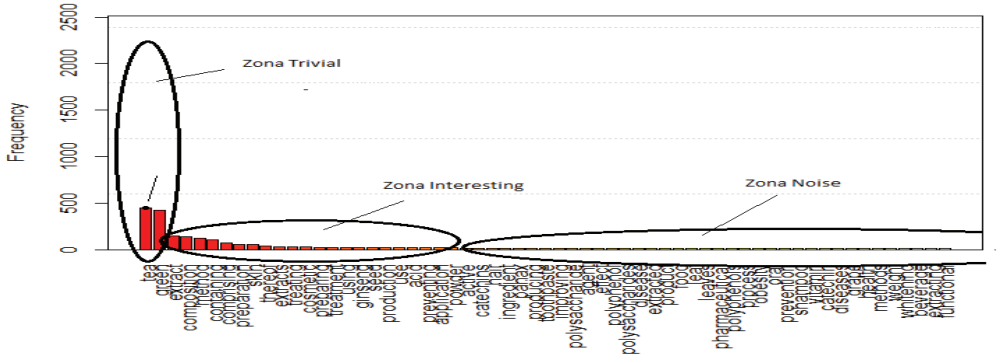

FIGURE 3. Zipf curve of terms

Based on the theory of the Zipf curve, the interesting zone is the innovative potential of information, so that the terms chosen to be object of this study. Terms contained in Table 4.1 contains the terms contained in this interesting zone, so the theory on Zipf curve is consistent with the purpose of research. The visualization of term frequency can also be viewed using wordcloud as follows.

FIGURE 4. Wordcloud Terms Has in Filter

Feature Selection

Having obtained the document term matrix, then weighted the terms that appear by using TF-IDF. After weighting by using the software R result of TF-IDF matrix obtained as follows.

TABLE 2. Matrix TF-IDF of Technology Patent Document Green Tea

Terms	The Document of Green Tea Patent							
	[,1]	[,2]	[,3]	[,4]	[,5]	.....	[,350]	[,351]
Composition	2.326	2.326	2.326	2.326	2.326	.....	0	0
Comprising	0	0	0	0	3.326	.....	0	0
Containing	0	2.811	2.811	2.811	0	.....	0	0
Cosmetic	0	0	0	0	0	.....	0	0
Extract	0	2.536	2.536	2.536	0	.....	0	2.536
Extracts	0	0	0	0	0	.....	0	0
Ginseng	0	0	0	0	0	.....	0	0
Method	2.584	0	0	0	0	.....	2.584	0
Preparation	0	0	0	0	0	.....	0	0
Preparing	0	0	0	0	0	.....	0	0
Skin	0	0	0	0	3.995	.....	0	0
Thereof	0	0	0	0	0	.....	0	4.097
Treating	4.501	0	0	4.501	0	.....	0	4.501
Treatment	0	0	0	0	0	.....	0	0
Using	0	0	0	0	0	.....	0	0

Data Analysis Phase

Step-up analysis conducted by using Adaptive K-Means Clustering. The threshold value used in the method of K-Means Clustering Adaptive this is 0.2, which is the default threshold at akmeans in software packages R. In addition, however, require a clustering analysis cluster validation measurements to determine how much the number of clusters can describe a good group. In this experiment, validation measures silhouette, with the following results.

TABLE 3. Results of the Patent Document Validation Silhouette Green Tea

Adaptive K-Means Clustering	
Threshold	**Silhouette Value**
0.2	0.66
0.3	0.64
0.4	0.56
0.5	0.59
0.6	0.56
0.7	0.51
0.8	0.46
0.9	0.42

Based on the theory of the validity of silhouette, the optimum number of clusters is determined based on the maximum silhouette. In Table 4.3, the most large silhouette value is 0.66 by using a threshold value of 0.2. Adaptive K-Means Clustering Results obtained 87 clusters were formed of 351 documents. Each cluster can characteristic with the terms contained therein based on the weight of these terms within each cluster. Below is a table of documents each cluster associated membership term are contained.

TABLE 4. Terms Weights and Cluster Membership

Cluster	Term at every cluster which has been organized from the highest weight until the lowest	Member of Clusters	Cluster	Term at every cluster which has been organized from the highest weight until the lowest	Member of Clusters
1	treatment comprising composition extract method	Document of 161, 162, 289, 292	45	extract	Document of 7, 19, 40, 52, 89, 192, 213, 240, 241, 250, 254, 257, 265, 314, 356
2	skin comprising composition	Document of 8, 120, 146, 199	46	containing extract	Document of 181, 263, 267, 359, 366
3	extracts thereof skin containing method composition	Document of 177	47	extracts cosmetic composition skin containing	Document of 210 dan 360
4	ginseng comprising extract composition treating method	Document of 57, 76, 123	48	treatment extracts cosmetic skin containing	Document of 368
5	preparing cosmetic comprising method skin extract composition	Document of 129 dan 90	49	treating method composition	Document of 17 dan 396
. . .	. . .	. . .	. . .	. . .	. . .
41	ginseng preparation extract	Document of 99	85	extracts extract	Document of 122 dan 256
42	ginseng thereof	Document of 345 dan 418	86	skin	Document of 173, 246, 374
43	containing method extract	Document of 108	87	skin extract	Document of 358
44	method extract	Document of 1, 2, 166, 200, 247, 260, 327, 339			

Table 4. shows the groups of documents are generated in this study. Obtained 87 groups or clusters based on adaptive k-means clustering method, but not all groups has performed in that table. The results of the cluster can be used for information retrieval which shows the appearance of a document when researchers call with two pair or better terms in the patent document green tea. Weighting cluster at the top make the order terms are called first. For example, when a user searches for technology use green tea with the theme treatment Comprising the Cluster 1 will be called.

CONCLUSIONS

Patent documents successfully grouped into 87 clusters / groups where in each group are characterized by different terms based on the weight of their respective interests. Adaptive Methods K-Means Clustering is more appropriate in a document grouping in this study. This is because there is no foundation to the group upriori determination of patent documents, so that by using this method of determining the number of groups formed by itself and produce groups of patent documents following characteristic term.

ACKNOWLEDGMENTS

This research was supported by Department of Statistics, Faculty Mathematics and Natural Science, Universitas Padjadjaran Indonesia. I would like to express my great appreciation to Dr. Jadi Suprijadi , DEA and Zulhanif, S.Si., M.Sc for every support, blessings, correcting and knowledge to finished this research.

REFERENCES

1. I.M.Monirul and J.H.Han, "Perceived quality and attitude toward tea & coffee by consumers," 2012, *International Journal of Business Research and Management (IJBRM)*, 3(3)
2. Sudjarmoko, B., Indonesian Tea Market Development in Domestic Market and Market International, accessible from the market http://balittri.litbang.pertanian.go.id/index.php/component/content/article/49-infotekno/207-perkembangan-pasar-teh-indonesia-di-pasar-domestik-dan -internasional, on October 5, 2016.
3. Suprihatini, R., Sa'id, E., Marimin and Djumali M. 2005. Analysis of the condition of the components of the processing technology in Indonesia bulk tea industry. Thesis. Faculty of Agricultural Technology. Bogor Agricultural Institute. Bogor.
4. L.Ruotsalainen, "Data Mining Tools for Technology and Competitive Intelligence," Espoo 2008, VTT Tiedotteita Research Notes 2451, P.11.
5. S.Jun, "A Statistical Text Mining Method for Patent Analysis," 2013, *International Journal of Advancements in Computing Technology (IJACT)*, Vol.5, P.144.
6. A.Abbas, L.Zhang, and S.U.Khan." A literature review on the state-of-the-art in patent analysis." *In World patent Information (2014)* (Elsevier ,2013), p.1-11.
7. H.Dou, V.Leveillé, S.Manullang & J.M.Dou Jr, "Patent Analysis For Competitive Technical Intelligence And Innovative Thinking," 2005, *Data Science Journal*, Vol.4, P.209.
8. Manalu, B.U. 2014. Sentiment Analysis On Twitter Using Text Mining. Essay. Faculty of Computer Science and Information Technology. University of Northern Sumatra. Field.
9. E.Dragut, F.Fang, P.Sistla, C.Yu, and W.Meng. "Stop Word and Related Problems in Web Interface Integration. 2009."
10. Langgeni, D.P., Abdurrahman, B., and Yanur, F. 2010. Clustering Articles News Speak Indonesia Using Unsupervised Feature Selection. National Seminar on Informatics, 2010. UPN "Veteran". Yogyakarta.
11. Rajalakshmi K, Thilaka B, Rajeswari N, "An Adaptive K-Means Clustering Algorithm and its Application to Face Recognition." 2010, *Computer Science & Mathematics*, 9(4).P.89.
12. Brock, G., Pihus, V., Datta, S., and Somnath D. 2008. clValid: An R packages for cluster validation. Journal of Statistical Software, Volume 25, issue 4.
13. Tarapanoff, K., Quoniam, L., Junior, R., and Lillian A. 2001. Intelligence Obtained by applying data mining to a database of French theses on the subject of Brazil. Information Research, vol.7 No.1.
14. Johnson, Richard, and Wichern, D. 2002. Applied Multivariate Statistical Analysis: Third Edition.