The Application of Centroid Linkage Hierarchical Method and Hill Climbing Method in Comments Clustering Online Discussion Forum

Okfalisa[1], Joni Iskandar[2]
[1,2]Department of Informatics Engineering
Universitas Islam Negeri Sultan Syarif Kasim Riau
HR Subrantas Street KM. 15, Rimba Panjang, Riau, 28293, Indonesia

***Abstract* – Several problems are rised in order to enhance the effectiveness of communication in online discussion. The similarity and repetition of comments in terms of questions in the sentences or text meanings as well as triggers the emerging of miscommunication amongst participants in a forum discussion are investigated. Moreover, some comments seems are ignored or not been touched by other participants and in advance the effective used of forum discussion as knowledge acquisition and sharing can not be achieved. This paper studies the application of Centroid Linkage Hierarchical Method (CLHM) Algorithm and Hill Climbing methods in findings the similarity value of participants comments and clustering based on it. The analysis follows the text mining process including text processing, text transformation, attribute selection and pattern discovery. In order to test the validity and accuracy of both application methods, confusion matrix in euclidean and consine similarity were calculated. As the results, from variety numbers of comments groups, including Bersosial.com in 17 comments, Indowebster.com in 27 comments and Teknojurnal.com in 51 data comments provided the value of well-separated clusters performed. This testing also defined that the alteration of threshold and altitude did not affect the clustering process. From the calculation of F-measure values in confusion matrix explained that consine similarity provided better result that euclidean distance where teknojurnal 0.89, indowebster,com 0.71 and bersosial.com 0.57. This showed that CLHM algorithm and Hill climbing methods are effective approaches and have been successfully applied in comments clustering of online discussion.**

Keywords: Centroid Lingkage Hierarchical Method (CLHM), Online Discussion, Hill-Climbing, Text Mining, Clustering.

I. INTRODUCTION

Nowadays, the advancing of information technology is as straight as arrow with the human needs on it. One of the impacts showed the emergence of high usefulness of websites, portals, blogs and many social media forum for life activities, such as education, work, business, and socialization. Based on Indonesia Internet Services Provider Association (APJII) internet user reached 88.1 million people in 2014 and it brought Indonesia as top 8 ranking (number 8) in the world [1]. This also triggers the availability of modern features in supporting information quality provided, for example Frequently Asking Question (F.A.Q) which can be used to predict some questions that might be asked based on data history of former user question. For online discussion forums such as Kaskus, Quora, Teknojurnal, Indowebster, Bersosial and Islamic News Portals, the availability of features which can improve information quality becomes an obligation in ensuring the right information accepted by the participant. The recommendation system as an information filtering tool to exact the most relevant information for online user is important [2 and 3]. Even thought there are various recommendation algorithm designed to take account on information filtering none are able to achieve both high recommendation accuracy and diversity [4 and 5]. Online discussion forum always fullfill with large volumes of comments everyday. Online commenting forums are important for readers in sharing their opinions and knowledge thus all group members involved would gain the benefit from a large set of comments such as faster access, acess multiple conversation, focused discusion, and contribution at relevant place in discussion [6]. The need of moderator as recommendation system in interventing expanded conversation and making the discussion focused is usefull. Thus allowing users to get an overview of the conversation and quickly understand what comments discussing about.

The objective of this research is to develop "the comments" feature in online discussion forum as a discussion interaction tool between participants and administrators. Commonly, the website usually restricts on the number of comments appear in the main page and focuses on the newest participant posted, for example Indowebster.com and Bersosial.com limited only 20 comments for each page. This triggers the participant to respond based on the latest comments without reviewing back previous discussion conducted. This lacking causes the possibility of similar repetition comments towards an ineffective forum discussion. This paper answered the question on how to classify the similarity large number of comments in order to provide an effective online discussion.

As problem solving, this paper proposed text processing method to recognize the similarity of contexts in the list of comments. Many previous research have been studied and applied the use of text mining and clustering method in finding pattern and connection in semi-structured and un-structured databases text, such as Application of Automatic Clustering in Document Searching Machine [7]; The

Application of Hill Climbing Automatic Clustering in Web mining for English Document Searching [8] and Comments-Oriented Document Summarization: Understanding Documents with Reader's Feedback [9]; A late acceptance Hill climbing algorithm for balancing two-sided assembly lines with multiple constraints [10]. The above researches found the successful of CLHM Algorithm in text database clustering. In order to analyze variance pattern of automatic clustering, Valley-tracing and Hill Climbing methods are commonly used. However, it is found that Hill Climbing is a faster and effective method in clustering and searching a large amount of document retrieval and long term keywords [7, 11, and 12]. This also triggers the emergence of innovation Hill Climbing method research [13, 14, 15 and 16]. This paper tried to study the application of CLHM algorithm which is combined with Hill Climbing method to automatic cluster based on the similarity of participant comments in online discussion forum. Therefore, the effective use of online discussion through the quality information filtering can be achieved. The structure of this paper consists of introduction stage which contained background of study, objectives, problem statement and significance. Furthermore theory backround on CLHM, Clustering Analysis Hill Climbing algorithm, Cluster validity and Measurement Standard in Document Clustering were explained. Methodology was discribed as research flow. Next, analysis and discusion were clarified as result of this paper. Finally conclusion remarks were given to end this paper.

II. THEORY BACKGROUND

A. CLHM Algorithm.

Text mining consists of interdisclinary studies including information retrieval, data mining, machine learning, statistics and linguistics computing [17, 18, 19, and 20]. The application of it can be seen in spam filtering; sentiment analytical; customer preferences measurement; document summarization and clustering research topics [21, 22, 23, and 24]. One of text mining technique is clustering that uses to identify a similar characteristics group of object that formed homogeneous group [25].

CLHM is clustering process algorithm which works based on the distance of it's centroid [26, 27]. This algorithm is effective for clustering normal dataset distribution. The first step algorithm is to assume each data regarded as a cluster, if *n*= number of data and c= number of cluster, so c=n; second is to find the distance between data and cluster with cosine similarity and Euclidean distance; next is finding 2 cluster with the largest in cosine similarity and the smallest distance in Euclidean, merge it into new cluster, so c= c-1; then repeat until it reaches optimum. It is ending by distancing calculation among data or through the calculation of Euclidean distance (Equation 1) and cosine similarity (Equation 2).

$$d(x,y) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \text{ ..... (1)}$$

$$sim(x,y) = \frac{\vec{x} \bullet \vec{y}}{|\vec{x}| \bullet |\vec{y}|} = \frac{\sum_{i=1}^{t}(W_{ix} \bullet W_{iy})}{\sqrt{\sum_{i=1}^{t} W_{ix}^2} \bullet \sum_{i=1}^{t} W_{iy}^2} \text{... (2)}$$

B. Clustering Analysis

As one of multivariate analysis techniques, clustering analysis is capable to find and organize variables within document information to be then clustered homogeneously [9]. It is earned from cluster density through variance within cluster (Vw) and variance clusters (Vb). The variance can be calculated by applying Equation 3[7, 17, and 26]:

$$Vc^2 = \frac{1}{n_c - 1} \sum_{i=1}^{n} (y_i - \bar{y_c})^2 \quad \text{..... (3)}$$

Where:

Vc^2 = variance in cluster c
c =1...k, where k=number of cluster,
n_c =number of data in cluster c,
y_i =data y_i, where i= 1…n and n is total data of each clusters.
y_c = centroid of cluster c.

Then, Vw can be calculated by Equation (4)[7] :

$$Vw = \frac{1}{N-c} \sum_{i=1}^{c} (n_i - 1) V_i^2 \text{ .... (4)}$$

where:

N = total number of data in all cluster,
n_i = number of data in cluster i,
V_i = variance of cluster i.

and to calculate Vb Equation (5)[7] is applied.

$$Vb = \frac{1}{c-1} \sum_{i=1}^{c} n_i (\bar{y_i} - \bar{y})^{2} \text{..... (5)}$$

where:

$\bar{y} = \text{mean of } \bar{y_i}$

C. Hill-climbing

In order to recognize the variance cluster pattern, Hill-climbing method is better than valley tracing method. This method provides large number of cluster construction due to its high sensitivity. The minimum running time of cluster in retrieving a document causes this method is suitable for long keyword and numerous data [7, 8, 17, and 19]. The global optimum of this method can be calculated by Equation (6) [7]:

$$V_{i+1} > a.V_i \text{ ... (6)}$$

where a is altitude value.

D. Cluster validity

There are two testing stages in identifing the cluster validity, including cluster accuracy and external validity. The cluster accuracy can be calculated by Equation (7) [7 and 17] with the condition if ($\varphi \geq 2$) means that the cluster is well-separated.

$$\varphi = \frac{\text{Max}(\partial)}{\text{nilai terdekat ke Max}(\partial)} \quad \text{... (7)}$$

where φ = phi value
∂ = high value differences

Meanwhile the external validity can be measured through external information within confusion matrix in terms of relevant or not relevant and retrieved or not retrieved comments.

E. Measurement Standard in Document Clustering

There are some standard measurement in document clustering, namely Recall, Precision and F-measure. Recall is the success level values in recognizing an event. It can be counting by formula below (Equation 8):

$$\left[Recall = \frac{|\{Relevant\} \cap \{Retrieved\}|}{|\{Relevant\}|}\right] \quad \text{... (8)}$$

where:
Recall = success values.
{Relevant} = number of document actually relevant.
{Retrieved} = number of document recognized by system.

Meanwhile Precision is precision level values of cluster against an event. The formula is as follows (Equation 9):

$$Precision = \frac{\{Relevant\} \cap \{Retrieved\}}{\{Retrieved\}} \quad \text{... (9)}$$

Where:
Precision = precision values
{Relevant} = number of document actually relevant
{Retrieved} = number of document recognized

F-Measure is the combination of Recall and Precision which is defined in Equation (10):

$$FMeasure \frac{2 x Recall x Precision}{Recall + Precision} \quad \text{... (10)}$$

III. METHODOLOGY

Based on the objectives of this paper, the stage conducted in research schema can be seen in Fig 1. It is started by data collection through several research literature reviews on topics text mining, information retrieval, document clustering, document summarization, CLHM algorithm method and Hill Climbing method. It then strengthened by web observation on online discussion forum such as Kaskus, Quora, Teknojurnal, Indowebster, Bersosial, Amazon and Islamic News Portals. The step analysis process initiated from text processing, text transformation, attribute selection and ending by pattern discovery through the application of CLHM clustering method and Hill Climbing as constraints for identifying optimum globalization of variance pattern in automatic clustering. Text data was dowloaded from three forum online discussion, including Teknojurnal 51 comments, Indowebster 27 comments, and Bersosial 17 comments. Through the document text structure, regular expression was identified for preprocessing stage. the application of tokenizing, case folding, filtering, spelling, normalization and stemming of data text were conducted in initiating text processing stage. In order to select an attribute, document representation then was calculated by using TF-IDF formula as follows (Equation 11)

$$W_{dt} = tf_{dt} * IDF_t \quad \text{...(11)}$$

Where:
d = document for-d
t = word for-t from the key
w = document weight for-d over the word for-t
tf = number of word in document searching
IDF= Inversed Documents Frequency
D = total document
df = number of document which contains the searching words.

The weight normalization can be counting by formula below (Equation 12).

$$w'_{i,j} = \frac{w_{i,j}}{\sqrt{w_{i,j}^2}} \quad \text{...(12)}$$

Where:
$w'_{i,j}$ =weight of word for-i with normalization
$w_{i,j}$ = weight of word for-i with un-normalization

The selection was conducted based on the similarity of words and the weight values of DF. This document index will be run and clustered in CLHM and Hill Climbing algorithm. In order to investigate the significance of analysis, three forum online discussions are tested for validity and accuracy test through the values of Euclidean, Cosine similarity, Recall, Precision and F measures. Finally, the conclusion and documentation are produced as the result.

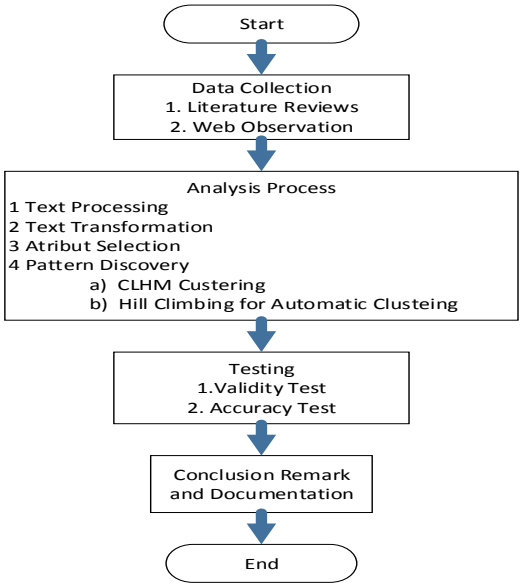

Fig 1. Research Flow Process

IV. Analysis and Discussion

Analysis is initiated by clustering the comments from three online discussion forums such as teknojurnal.com, indowebster.com and bersosial.com (See Table I). Online forum in Table I was classified based on the avarage quantity of comments in general. It was devided into three group classification, including high with 51 comments, avarage with 27 comments, and the lowest with 17 comments. This clustering automatically places the significant similarity of comments into ideal cluster. The above comments must followed the web clustering process as explained in Fig 2. Fig 2 described that data acquisition comes from participant comments based on topics classification. Then, the system analyzed web page forum discussion by converting html page format downloaded into *plaintext* then parsing it as well as document structure identification.

As the result, document repository is established which contains the string parsing topics and comments database. Before clustering, document database proceeds into text pre-processing including tokenizing, case folding, filtering, spelling, normalization and stemming. It is then re-presented the document weighting TF-IDF method and Vector Space Model for normalization.

Table II explains the final result of attribute selection from teknojurnal forum where 50 words that passed text processing was decreased into 7 words normalization. This is due to the value of frequency document (DF) index (Equation 11 and 12) less than 2. From 7 words in Table II, the word "things" and "internet" had the highest values. It showed the strong lingkage of those words to comments.

TABLE I
COMMENT DATA OF WEB ONLINE FORUM

No	Website/ Forum	Topic	No of comment
1	http://teknojurnal.com/	definition-internet-of-things	51
2	Forum.idws.id (Forum Indowebster)	Machinery - Google Glass, Technology Hands free Bluetooth Intelligence	27
3	https://www.bersosial.com	7 Things that lecturers avoid to talk to students _ Bersosial.com	17

Furthermore, unsupervised learning of pattern discovery with CLHM clustering and Hill Climbing method runs Equation 1-7 to find Euclidean distance and cosine similarity values as explained in Table III for three web testings. The above websites provided the global optimum position of cluster automatically until it is constructed into well-separated cluster. Well separated cluster is identified by Equation 7 where euclidean and cosine similarity $\varphi \geq 2$. Table III showed Teknojurnal web is well separated in φ = 7.0519 and 41 clusters performed.

After cluster performed, confusion matrix is then calculated to identify the cluster group. Based on Equation 8, 9 and 10, validity and accuracy tests in confusion matrix are conducted to calculate the value of Precision, Recall and F-Measure from three web test systems. As the result, Table IV and V are produced as validity testing.

Table IV explained that from three web tests, teknojurnal.com provided the best performance. It can be showed by the number of data retrieved and relevant in highest score (12 data in Euclidean and cosine similarity calculation). This is also supported by the highest number of data not retrieved and not relevant (32 in Euclidean and 36 in cosine similarity calculation).

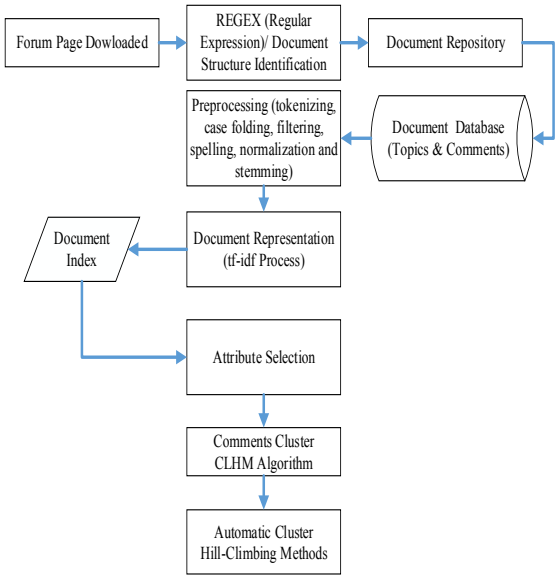

Fig 2. Flowchart Of Web Clustering Process

TABLE II
DOCUMENT REPRESENTATION AFTER ATTRIBUTE SELECTION

Term	tf(D1)	w(D1)	tf(D2)	w(D2)	tf(D3)	w(D3)	.....	tf(d10)	w(d10)	IDF	DF
benda	0	0	0	0	1	0,522879		0	0	0,522879	3
berita	0	0	0	0	0	0		0	0	0,69897	2
saji	0	0	0	0	0	0		0	0	0,69897	2
things	0	0	0	0	0	0		1	0,30103	0,30103	5
inovasi	0	0	0	0	0	0		0	0	0,69897	2
internet	0	0	0	0	0	0		1	0,30103	0,30103	5
iot	0	0	0	0	0	0		0	0	0,69897	2

Web test validity and accuracy check in Table V indicated teknojurnal.com as the web best performance in clustering process. It is found well-separated= 7.05 (Phi (φ) $\geq$ 2); precision = 0.8; recall= 1; F-measure= 0.89 (validity and accuracy values are nearly close to 1 in cosine similarity). Performance is then followed by indowebster.com (well-separated=∞; precision = 0.56; recall= 1; F-measure= 0.71)

and bersosial.com respectively (well-separated= ∞; precision = 0.5; recall= 1; F-measure= 0.57). This showed that the alteration of threshold (λ) and altitude (α) did not affect the cluster performed (See Table IV where cluster retrieved and relevant are similar). However, from three web testing F-measure values in consine similarity provided better result than euclidean distance where teknojurnal.com 0.89, indowebster.com 0.71 and bersosial.com 0.57.

TABLE III
EUCLIDEAN DISTANCE AND COSINE SIMILARITY VALUES

Webs Test	Euclidean			cosine similarity	
	No. Group Comments Before	No. Group Comments After	φ	No. Group Comments After	φ
Teknojurnal.com	51	40	7,0519 (well - separated)	41	~ (well - separated)
Indowebster.com	27	9	~ (well - separated)	23	~ (well - separated)
Bersosial.com	17	7	~ (well - separated)	13	~ (well - separated)

TABLE IV
CONFUSION MATRIX

Webs Test Teknojurnal.com	Euclidean		Cosine similarity	
	Relevant	Not Relevant	Relevant	Not Relevant
Retrieved	12	4	12	0
Not Retrieved	3	32	3	36
Webs Test Indowebster.com				
Retrieved	4	15	4	0
Not Retrieved	3	5	3	20
Webs Test Bersosial.com				
Retrieved	4	5	4	0
Not Retrieved	4	4	6	7

TABLE V
WEB TEST VALIDITY CHECK

Dataset	*Euclidean*	*Cosine Similarity*	α	(λ)
Teknojurnal.com	*Well-Sep*=7,0519	*Well-Sep*=~	α=2	0,05-1
	Precision=0,8	*Precision*=0,8		
	Recall=0,75	*Recall*=1		
	F-Measure=0,77	*F-Measure*=0,89		
Indowebster.com	*Well-Sep*=~	*Well-Sep*=~	α=3	0,05-1
	Precision=0,57	*Precision*=0,56		
	Recall=0,21	*Recall*=1		
	F-Measure=0,31	*F-Measure*=0,71		
Bersosial.com	*Well-Sep*=~	*Well-Sep*=~	α=4	0,05-1
	Precision=0,5	*Precision*=0,4		
	Recall=0,44	*Recall*=1		
	F-Measure=0,46	*F-Measure*=0,57		

V. CONCLUSION

1. The accuracy value of best well-separated cluster construction found 7.05 in 51 number of data comments in teknojurnal web. Meanwhile, the worst performance is obtained in 27 and 17, the number of data comments. This concludes that well-separated cluster is influenced by the number of cluster construction stage and the number of data comments (linear) in website.
2. In order to measure the similarity distance between comments, cosine similarity provides the higher accuracy value than euclidean distance (See Table 5).
3. CLHM algoritm has been successfully applied in clustering comments in online discussion forum. The effective used of Hill Climbing method in automatically cluster constructed has been proven enhancing CLHM algoritm performance.

ACKNOWLEDGMENT

We would like to thank the university State Islamic University of Sultan Syarif Kasim Riau (UIN Suska Riau) Riau, Indonesia as sponsorship and our team in Department Informatics Faculty Science and Technology, the reviewers, who have provided comments and suggestions to improve the manuscript. Many thanks for your cooperation and support.

REFERENCES

[1] Indonesia Association Internet Provider, "Indonesia Internet User Profile 2014, Puskakom UI Jakarta, First Edition, March 2015.

[2] Fu Guo Zhang and An Zheng. *"Information Filtering via Heterogeneous Diffusion in Online Bipartite Networks"*. Jornal Plos One DOI:10.1371/journal.pone.0129459, June 30, 2015.

[3] Adomavicius, G and Tuzhilin, A.*"Toward the next generation of recommender systems : a survey of the state of the art and possible extensions"*. IEEE Transcactions on Knowledge and Data Engineering, Vol. 17, Issue: 6, pp. 734-749.2005

[4] Tianqi Chen,Weinan Zhang, Qiuxia Chen, Zhao Zheng, and Yong Yu. *"SVDFeature: A Toolkit for Feature based Collaboration Filtering"*. Journal of Machine Learning Research, Vol. 13. Pp. 3619-3622. 2012

[5] Zhang F.G and Zeng A. "Improving Information Filtering via Network Manipulation" , EPL, 100:58005. 2012

[6] Aker Ahmet, Kurtic Emina, Balamurali, A.R., Paramita Monica, Barker Emma, Hepple Mark, and Rob Gaizauskas.*"A Graph Based Approach to Topic Clustering for Online Comments to News"*. Springer International Publishing Switzerland N. Ferro et al. (Eds): ECIR 2016 DOI: 10.1007/978-3-319-30671-1_2. 2016

[7] Entin Martiana, Nur Rosyid and Usmaida. Agusetia, "Application of Automatic Clustering in Document Searching Machine,"*Telekomnika*, Vol. 8, no. 1, 2010.

[8] Eldira, Hervilorra and K Entin. Martiana and Mubtada, Nur Rosyid, "The Application of Hill Climbing Automatic Clustering in Web mining for English Document Searching," *EEPIS Project*, 2011.

[9] Meishan Hu, Aixin Sun and Ee-Peng Lim, "Comments-Oriented Document Summarization: Understanding Documents with Readers' Feedback," *SIGIR 08 Proceeding of the 31 st annual international ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore,* 2008.

[10] Biao Yuan, Chaoyong Zhang and Xinyu Shao,. "A Late Acceptance Hill-Climbing Algorithm for Balancing Two-Sided Assembly Lines with Multiple Constraints", *Journal of Intelligent Manufacturing*, Vol. 26, Issues 1, pp. 159-168, 2015.

[11] Jorg Hoffmann, "*A Heuristic for Domain Independent Planning and Its Use in an Enforced Hill Climbing Algorithm*". Foundation of Intelligent Systems Book Chapter, Lecture Notes in Computer

Science, Vol. 1932. 2002.
[12] Wei Liu and Wilson Wang, "Web Service Clustering Using Text Mining Techniques". *International Journal of Agent-Oriented Software Engineering*, Vol. 3, Issue 1. 2009.
[13] Lim A., Rodrigues B. and Zhang X,"A Simulated Annealing and Hill Climbing Algorithm for Traveling Tournament Problem". *European Journal of Operational Research*, Vol. 174, Pages 1459-1478. 2006.
[14] Hongfeng Wang, Dingwei Wang, and Shengxiang Yang, "A Memetic Algorithm with Adaptive Hill Climbing Strategy for Dynamic Optimization Problems". *Soft Computing Journal*, Vol. 13, Issue 8, pp. 763-780, 2009.
[15] Lee R Cooper, David W Corne and M. James C Crabbe, "Use of a Novel Hill Climbing Genetic Algorithm in Protein Folding Simulations", *Computational Biology and Chemistry*, Vol. 27, Issue 6, pp. 575-580, 2003.
[16] Ali Riza Yiliz, "An Effective Hybrid Immune-Hill Climbing Optimization Approach for Solving Design and Manufacturing Optimization Problems in Industry". *Journal of Materials Processing Technology*, Vol. 209, Issue 6, pp. 2773-27780, 2009.
[17] Jiawei, H., Kamber, M., & Pei, J, *Data Mining: Concepts and Techniques,* Third Edition. Waltham, MA: Morgan Kaufmann, 2012.
[18] Zhai, C., & Aggarwal, C. C, *Mining Text Data*, New York: Springer, 2012.
[19] Feldman, R., & Sanger, J, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, New York: Cambridge University Press, 2007.
[20] Gullo, F, "From Patterns in Data to Knowledge Discovery: What Data Mining Can Do," *Physics Procedia* 62 , pp.18-22, 2015.
[21] Oiaozhu Mei and ChengXiang Zhai, "Discovering Evolutionary Theme Pattern Form Text: An Exploration of Temporal Text Mining," *KDD 05 Proceedings of the elevent ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, Chicago, Illinois, USA, August 21-24, 2005. pp. 198-207, 2005.
[22] Krishna K. and Raghu Krishnapuram, "A Clustering Algorithm for Asymmetrically Related Data with Applications to Text Mining," CIKM *01 Proceeding of the Tenth International Conference on Information and Knowledge Management*. pp. 571-573, Atlanta, Georgia, USA, October 05-10, 2001.
[23] Shady Sehata, Fakhri Karray and Mohamed Kamel, "An Efficient Concet Based Mining Model for Enhancing Text Clustering," *IEEE Transactions on Knowledge and Data Engineering*. Vol. 22, Issue 10, 2009.
[24] Liritano S., and M. Ruffolo , "Managing the Knowledge Contained in Electronic Documents: a Clustering Method for Text Mining". *Database and Expert Systems Application, Proceeding 12th International Workshop* on 3-7 September, 2002.
[25] Charu C. Aggarwal and ChengXiang Zhai, *"Mining Text Data,"* Springer New York Dordrecht Heidelberg London, 2012.
[26] Christopher D Manning, Prabhakar Raghavan and Hinrich. Schutze, *"An Introduction to Information Retrieval,"* Online Edition, Cambridge University Press, Cambridge, England. 2009.
[27] Feldman, R., & Sanger, J. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. New York: Cambridge University Press. 2007