ISSN: 2278-2419
Volume:04, Issue:02, December 2015

International Journal of Data Mining Techniques and Applications

Integrated Intelligent Research

International Journal of Data Mining Techniques and Applications

(IJDMTA)

Volume 04, Issue 02, December 2015

ISSN: 2278-2419

CONTENTS

International Journal of Data Mining Techniques and Applications
Volume: 04 Issue: 02 December 2015, Page No. 25-28
ISSN: 2278-2419

Applying Clustering Techniques for Efficient Text Mining in Twitter Data

K.T.Mathuna , I.Elizabeth Shanthi, K.Nandhini
Department of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore
Email: mathuna.thangaraj@gmail.com, shanthianto@gmail.com, nandhinik2@gmail.com,

Abstract- Knowledge is the ultimate output of decisions on a dataset. The revolution of the Internet has made the global distance closer with the touch on the hand held electronic devices. Usage of social media sites have increased in the past decades. One of the most popular social media micro blog is Twitter. Twitter has millions of users in the world. In this paper the analysis of Twitter data is performed through the text contained in hash tags. After Preprocessing clustering algorithms are applied on text data. The different clusters formed are compared through various parameters. Visualization techniques are used to portray the results from which inferences like time series and topic flow can be easily made. The observed results show that the hierarchical clustering algorithm performs better than other algorithms.

Keywords- Text mining, Visualization, Twitter, Clustering, Time series analysis, Social media.

I. INTRODUCTION

The communication of people has greatly changed in recent years, one of the most important being social media networking [1]. Nowadays, social media is not only used for personal networking but also for commercial purposes. Happenings in the real world is shared and communicated via Internet. Internet forms a bridge between the users irrespective of the global distance. The most popular social sites includes Face book, Twitter and You tube [2]. The social media has opened up many research opportunities because of the increased amount of information.

A. Background

This paper deals with the analysis of the topic and event detection on the social micro blogs. Particularly, Twitter which is widely used and fast growing in real world blog. More than 500 million users are assessing twitter and above 302 are active users which generate about 340 million messages everyday [1][3]. People upload their opinion and real world happenings in this public site. Current topics and trends are the main features of twitter. Twitter provides "hash tag (#)", which is used for providing topics for tweets. If this hash tag is used by many people then the topic becomes the current trending topic [4]. Hence gathering, mining and analyzing tweets has its own importance in all areas. This paper aims in analyzing the efficient text mining algorithms via clustering techniques.

Text Mining- Text mining, also known as text data mining or knowledge discovery process from the textual databases is generally the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents. Text Mining is similar to data mining, for storing information in structured manner and on the contrary text mining uses texts that are unstructured or semi-structured. The data is collected through World Wide Web, Linked In, government portals, face book, twitter, blog, news articles, digital libraries, electronic mail and so on. Approximately 80% of the organizational information is stored in unstructured form [5].

B. Applications and Issues in Text Mining

Text mining has its roots in almost all the areas. Text is used everywhere and has to be mined for information retrieval and knowledge gathering. It is applied in fields like publishing, media, telecommunications, research, banks, insurance, finance, government administration, legal documents, health care, business intelligence, national security, etc. All these fields have been improved and still waiting for betterment. The major challenge arises from the natural language processing of large information to extract required knowledge. Complexity of extraction and computational cost initiates the need for further improvement. Unstructured and multiple forms of text documents makes the retrieval process challenging. Research work is still in need on issues like text mining that uses different intermediate forms, integration of domain knowledge, analysis of social media network, etc [6].

C. Integration of Text Mining and Visualization

Visualization aids in better understanding of the extracted content from the raw database. It further gives a clear picture of the information that has to be delivered. Text mining integrated with visualization provides a better and fast understanding of interpreted results. Text visualization has two forms, *Topic based* and *Feature based.* In *Topic based* method, topics and events are visualized through visualization techniques. Some of the techniques include Tag clouds which depict the keywords or named entities. They use features like colour, size and layout based on usability and importance. Information landscape provides a geographical view of large set of documents for analysis. Text Flow method combines topic mining and interactive visualization techniques to visually analyze the evolution of topics in due course. In *Feature based* method, Word clouds are commonly generated to provide an intuitive visual summary of documents by displaying the keywords in a compact layout. Facet Atlas method integrates node-link diagram with density map to visually analyze the multifaceted relations of the document [7].

The organization of the paper is as follows. Section II deals with the recent research. Section III tells about the methodology and algorithms used. Section IV explains the experimental results and discussions and finally in Section V conclusion is given.

International Journal of Data Mining Techniques and Applications
Volume: 04 Issue: 02 December 2015, Page No. 25-28
ISSN: 2278-2419

II. LITERATURE REVIEW

Recent research work on Twitter data has been in various topics. Density based clustering, naïve based and other techniques used in research along with their observations are listed in Table 1.

Table 1: Review of the Related Work

Authors	*Technique(s) used*	*Observations*	*Year*
Chung-Hong Lee [8]	Density-based clustering method is used to mine micro blogging text streams.	Temporal and spatial features of real world events are analyzed and estimated.	2012
Rischan Mafrur et al.[4]	Naïve Bayes Classification technique was applied and word cloud visualization method was used.	Analysis about the election campaign through twitter shows the positive and negative comments from people.	2014
Farhan Hassan Khan et al.[9]	Proposed hybrid model for classification of tweets. Enhanced Emoticon Classifier (EEC), Improved Polarity Classifier (IPC) and SentiWordNet Classifier (SWNC) where used.	The hybrid model showed better accuracy. It decreases the neutral opinion by correctly classifying into positive or negative sentiment.	2014
Isti Surjandari et al.[10]	Support vector machine (SVM), naïve bayes and decision tree algorithms are used to classify. Chi squared test and Marascuilo procedure were performed for framing association rules.	Support vector machine resulted with higher accuracy in analyzing public opinion.	2015
Janez Kranjc et al.[11]	ClowdFlows, cloud-based scientific workflow platform with widget memory and halting mechanism was created. Word cloud and stream based visualization is given.	A web service is built to apply the models on unlabeled tweets.	2015
Shakira Banu Kaleel et al.[12]	Locality sensitive hashing (LSH) and K-means algorithm is used to form	Locality sensitive hashing (LSH) improved accuracy and	2015
	detecting and trending events form twitter clusters. Word cloud and Google maps are used.	runtime.	

Form the recent study, Support vector machine, naïve base classifier, density based clustering and k-means clustering algorithms were applied and the observations show that works are done on sentiment analysis to improve the accuracy. Works based on current issues and real world happenings still have a wider scope in research.

III. METHODOLOGY

A. *Overview of Methodology*

Tweets are gathered from twitter for text mining process. The user generated twitter data contains slang, noise and grammatical mistakes. These have to be cleaned for improving the quality of the tweet features [11].

Preprocessing in text data involves Stop word removal, Stemming, Converting upper cases to lower, removing punctuations and numbers. This makes the text more content specific. Stop word removal aims in removing stop words which has no meaning when it is single. Articles, prepositions, pro-nouns and conjunctions are the most common stop words which includes words like is, the, an, but, for etc. These words have to be removed to make text processing fewer complexes to facilitate the reduction in the number of words for retrieval [13, 14]. Stemming removes affixes in a word leaving the root word. For example the words study, studied, studying, studies gives the root word "*study*". This method reduces the indexing structure size as the numbers of distinct index terms are reduced [14, 15]. Finally converting upper case er, removing punctuations and numbers reduces the retrieval complexity. These methods of preprocessing make the text corpus less complex for text mining.

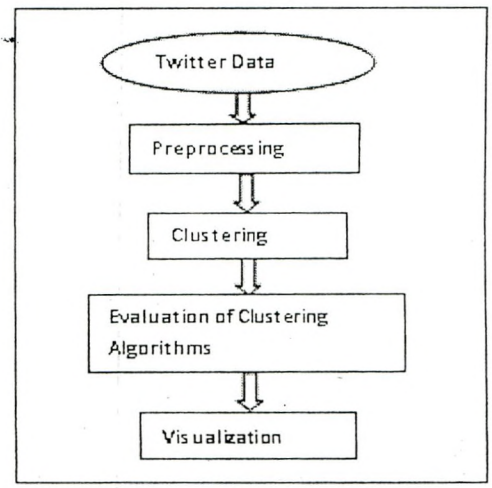


Figure1. Overview of Methodology

International Journal of Data Mining Techniques and Applications
Volume: 04 Issue: 02 December 2015, Page No. 25-28
ISSN: 2278-2419

The preprocessed text data is given as input to the clustering algorithms. Clustering is an unsupervised technique; it aims in grouping similar words in the same cluster and dissimilar words in different clusters [16]. The K-means, Entropy weighted K-means and Hierarchical algorithms are some of the most commonly used algorithms for text mining process. K-means algorithm is the basic and popular algorithms in text mining and used for large set of data. Entropy weighted K means (EWKM) algorithm is mainly used in Rattle data mining. Ewkm algorithm includes weighting scheme for clustering data and used for high dimensional data [17, 18]. Hierarchical Clustering algorithm forms clusters based on the dissimilarity between the observations. This algorithm decides when a cluster has to be splitted and when clusters have to be joined [16].

These algorithms are applied in Rattle environment in R language. Visualization is the communication of data through the use of interactive graphical user interfaces. In text mining ualization methods can improve and simplify the discovery extraction of relevant patterns of information [19].

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

The twitter dataset is taken from the Sentiment 140 website for academics [20]. R language and environment for statistical computing and graphics is used to provide a wide variety of statistical and graphical techniques, and is highly extensible.

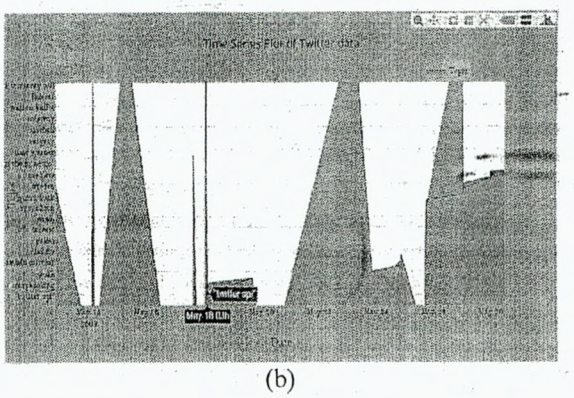
Tweets are preprocessed through text preprocessing methods and then applied on the clustering algorithms. K-means, Entropy Weighted K-means and Hierarchical clustering algorithms form clusters of similar data. The words that are similar in tweets are grouped together in single cluster and dissimilar words are grouped in different clusters. These clusters formed are evaluated based on the internal cluster validation parameters that are commonly used. Parameters like Average Between, Average Silhouette Width, Pearson gamma, Dunn and Dunn2 are calculated for all the three algorithms and mpared to find the best applying algorithm. The experiment as performed on cluster size increasing from 2 to 20, on preprocessed twitter data but the optimal and standard values for the data set have been obtained for cluster size 6 to 10. The optimal cluster size was obtained based on experiments conducted. Table 2 lists observed values of these parameters for the clustering algorithms used.

It is clear from Table 2 that Hierarchical algorithm performs better than the other two clustering algorithms for clusters of size ranging from 6 to 10. In this research work visualization methods are applied to the results interpreted from Hierarchical clustering algorithm. *Word cloud* is used to display the frequency of words that occur in twitter data. Time series analysis is one important issue in the current scenario. The day to day happenings are recorded and interpreted for knowledge discovery, information retrieval and decision making. Area plot is created using Plotly package in R language. This shows the time line of the topic that is tweeted. The frequency of a topic that is currently tweeted decides on the current trending news which is given through *2D Histograms*. Figure 2(a) shows the word cloud of tweets, Figure 2(b) shows the time series visualization of twitter data and Figure 2(c) shows the topic frequency of twitter data. The interpreted results are visualized through visualization methods.

Table 2. Comparison of Clustering Algorithms

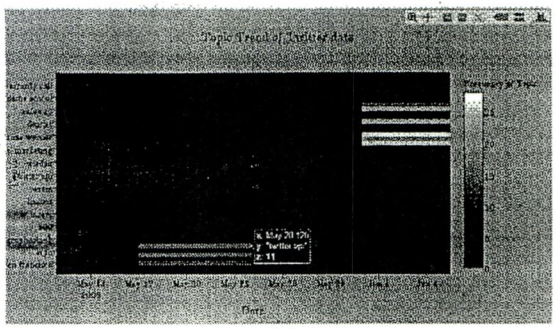
Number of clusters	*Parameter (min²)*	*K-Means*	*Ewkm*	*Hierarchical*
	dunn2	0.2144205	0.04087694	0.5501868
8	avg.between	2707.922	2681.809	2934.221
	avg.shilwidth	-0.07800377	-0.3615832	0.7646878
	pearsongamma	0.2291186	0.173456	0.1163658
	dunn	0.000371132	0.000185366	0.06829618
	dunn2	0.1996394	0.04817217	0.6920821
9	avg.between	2680.734	2712.621	2861.714
	avg.shilwidth	-0.03648374	-0.2015375	0.7724293
	pearsongamma	0.2216804	0.2282056	0.2158955
	dunn	0.000371132	0.000338952	0.06829618
	dunn2	0.1996394	0.04849406	0.3927788
10	avg.between	2659.513	2585.059	2785.165
	avg.shilwidth	-0.09589604	-0.4249884	0.7874223
	pearsongamma	0.214121	0.14578	0.122558
	dunn	0.000371132	0.000185366	0.06442296
	dunn2	0.1996394	0.0353203	0.3750494

(a)



(b)

International Journal of Data Mining Techniques and Applications
Volume: 04 Issue: 02 December 2015, Page No. 25-28
ISSN: 2278-2419



(c)

Figure 2. (a) Word cloud Visualization, (b) Time series analysis, (c) Topic trends of Twitter data.

V. CONCLUSIONS AND FUTURE SCOPE

The results of this study show the clustering of twitter data. Twitter is fast growing and widely used social networking site on the World Wide Web. Mining of Twitter data has gained importance in the past decades. Integrating both text mining and visualization provides better knowledge on information discovery and decision making. This work can be further applied on medical text mining, government portals, etc so that they could be used for national security and rehabilitation due to natural disasters.

REFERENCES

[1] Samar M.Alqhtani, Suhuai Luo and Brain Regan, "Fusing Text and Image for Event Detection in Twitter," The International Journal of Multimedia and Its Applications (IJMA), vol.7, No.1, pp.27-35, February 2015.

[2] J.Mingers, "The paucity of multimethod research: a review of the information systems literature," Information Systems Journal, vol.13, pp.233-249, July 2003.

[3] Twitter: https://en.wikipedia.org/wiki/Twitter.

[4] Rischan Mafrur, M Fiqri Muthohar, Gi Hyun Bang, Do Kyeong Lee, Kyungbaek Kim and Deokjai Choi "Twitter Mining: The Case of 2014 Indonesian Legislative Elections," International Journal of Software Engineering and Its Applications, vol. 8, No. 10, pp. 191-202, 2014.

[5] AurangzebKhan, Baharum Baharudin, Lam Hong Lee, Khairullah Khan"A Review of Machine Learning Algorithms for Text-Documents Classification," Journal of Advances in Information Techonology, vol.1, No.1, Feburary 2010.

[6] K.L.Sumathy ,M.Chidambaram, "Text Mining: Concepts, Applications, Tools and Issues – An Overview," International Journal of Computer Applications, vol. 80, No.4, pp. 29-32, October 2013.

[7] Guo-Dao Sun, Ying-Cai Wu, Rong-Hua Liang and Shi-Xia Liu "A Survey of Visual Analytics Techniques and Applications:State-of-the-Art Research and Future Challenges," Journal of Computer Science and Technology, vol. 28, No.5, pp. 852-86, September 2013.

[8] Chung-Hong Lee "Mining spatio-temporal information on microblogging streams using a density-based online clustering method," Expert Systems with Applications, Elsevier, vol. 39, No. 10, pp. 9623–9641, August 2012.

[9] Farhan Hassan Khan, Saba Bashir, Usman Qamar "TOM: Twitter opinion mining framework using hybrid classification scheme," Decision Support Systems, Elsevier, vol. 57, pp. 245–257, January 2012.

[10] Isti Surjandari, Muthia Szami Naffisah and M. Irfan Prawiradinata "Text Mining of Twitter Data for Public Sentiment Analysis of Staple Foods Price Changes," Journal of Industrial and Intelligent Information, Engineering and Technology Publishing ,vol. 3, No. 3, doi: 10.12720/jiii.3.3.253-257, September 2015.

[11] Janez Kranjc, Jasmina Smailovic, Vid Podpecan, Miha Grcar , Martin Znidarsic, Nada Lavrac "Active learning for sentiment analysis on data streams: Methodology and workflow implementation in the ClowdFlows platform," Information Processing and Management, Elsevier, vol. 51, No.2 , pp. 187–203, March 2015.

[12] Shakira Banu Kaleel, Abdolreza Abhari "Cluster-discovery of Twitter messages for event detectionand trending," Journal of Computational Science, Elsevier, vol. 6, pp. 47–57, January 2015.

[13] Dr. S. Vijayarani, Ms. J. Ilamathi, Ms. Nithya "Preprocessing Techniques for Text Mining - An Overview," International Journal of Computer Science & Communication Networks, vol 5(1), pp. 7-16.

[14] Ms. Nikita P.Katariya, Prof. M. S. Chaudhari "Text Preprocessing For Text Mining Using Side Information," International Journal of Computer Science and Mobile Applications, vol.3 Issue. 1, pp. 01-05, January 2015.

[15] C.Ramasubramanian, R.Ramya "Effective Pre-Processing Activities in Text Mining using Improved Porter's Stemming Algorithm," International Journal of Advanced Research in Computer and Communication Engineering, vol. 2, Issue 12, pp. 4536-4538, December 2013.

[16] Divya Nasa "Text Mining Techniques- A Survey," International Journal of Advanced Research in Computer Science and Software Engineering, vol 2, Issue 4, April 2012.

[17] Anil Kumar Tiwari , Lokesh Kumar Sharma , G. Rama Krishna "Entropy Weighting Genetic k-Means Algorithm for Subspace Clustering," International Journal of Computer Applications (0975 – 8887), vol 7– No.7, pp.27-30, October 2010.

[18] Zhongying Zhao, Shengzhong Feng, Qiang Wang, Joshua Zhexue Huang, Graham J.Williams, Jianping Fan "Topic oriented con ity detection through social objects and link analysis in social ne :s" Knowledge Based Systems, Elsevier, vol. 26, pp. 164-173, February 2012.

[19] Andreas Hotho, Andreas N¨urnberger, Gerhard Paaß "A Brief Survey of Text Mining," In *Ldv Forum* vol. 20, No. 1, pp. 19-62, May 2005.

[20] Sentiment 140 for academics: http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip.