International Journal of Innovative Research in Science and Engineering
Vol. No.2, Issue 04, April 2016
www.ijirse.com
ISSN: 2454-9665

CLUSTERING AND LABELING IN MICROBLOGGING USING NLP TECHNIQUES

H. Mohammed Sameer[1], Smt. M.Kavitha[2], Mr.Srinivas Karur[3]

[1] Mtech.,4th sem, [2,3]Asst. Professor, Dept. of CSE., S.I.T,Tumakuru, Karnataka, (India)

ABSTRACT

The explosive popularity of microblogging services produces a large volume of microblogging messages. It presents a great difficulty for the user to quickly gauge his/her followees opinions when the user interface is overwhelmed by large number of message. So we propose NLP techniques to organize the large amount of messages into clusters and label them to provide an overview of the content and to fulfil users needed information. Clustering and labeling of microblogging messages are challenging because the length of messages are much shorter than conventional text documents. As a result traditional text representation models lend to yield unsatisfactory performance. In this paper, we present the method that divides the documents into sentences, and perform clustering using Named entity recogniser which is a tool of Natural language processing and each cluster is assigned with label i.e., most frequently used word in the clusters. These labels facilitate the user to quick gauge his/her opinion. Experimental results verify that, this method is feasible and its effectiveness it is still under study.

Keywords: cluster, labels, Twitter, Facebook, LinkedIn, Social Media, Natural language Processing.

I. INTRODUCTION

Natural languages can be broadly defined as many informal languages that can be learnt by any person. They differ from formal languages like computer programming languages which have a proper structure and syntax. Natural Language Processing (NLP) is the understanding of the context of these natural languages and/or generating a text in natural language. Xia Hu[1]defines that the explosive popularity of microblogging services produce a large volume of microblogging messages which makes it difficult togauge his/her followees opinions. In recent years, Microblogging's such as Facebook and Twitter have become important communication tools for people across the globe. These websites are increasing the use for communicating breaking news, eyewitness accounts and organizing large groups of people. Users of these websites have become accustomed to receive timely updates on important events, both personal and global value. For instance, Twitter and Facebook is used to propagate information in real-time in many crisis situations such as floods, earthquakes and election details.

Many organizations and celebrities use their Twitter/Facebook accounts to connect to their customers and fans. Because of this increasing popularity, microblogging services produce a large amount of data every second. This information is aoverload and presents great difficulties for users to fulfil their needs. According to Twitter[3], many Twitter users only have the patience to glance the latest and sometimes redundant tweets. Many tweets of their interests may be buried in the large amount of streaming data. Given the huge number of tweets it is hard for users to efficiently gauge the main topics from their tweets.

ISSN: 2454-9665

Many valuable and interesting messages may be buried in unorganized large volume of data. To make this large collection of microblogging messages accessible to users as in current web systems style which provides accurate clusters in text and labels for each cluster, we perform clustering and labeling on large collection of unorganized data. With this users will be able to quickly identify messages of their interest like search engines. Under this scenario, our system is to explore clustering[1] and labeling in microblogging's. Clustering[2] is one of the tasks of data mining and it defines a process of finding groups in the data based on some similarity in data. Clustering is performed using NLP technique i.e., Named Entity Recognition and each cluster is assigned to relevant most frequently used word to form label.

II. RELATED WORK

With the increasing popularity of microblogging services, produces a large amount of data, **Xia Hu, Lei Tang, Huan Liu**[1]explains a method which form clustering and labeling using Wordnet and Wikiconcept. The limitation of this method is filtering of unstructured words appearing in the data.

Lidong Wang and Baogang Wei and JieYuan[2]propose a method to find the topics efficiently by combining the topic discovery and topic re-ranking techniques. Most topic models rely on the bag-of-words(BOW) assumption. The drawback of this method is difficulty in assuming the BOW.

Xiaohua Hu and xiaodanzhang[4]present a framework to improve the Clustering using word vector and category vector concepts. It works only on the basis of exact match and related match by using whole wiki data. Handling whole wiki data itself is difficult.

Y. Ko and J. Seo[12]described the automatic text Categorization method. This method can be used in areas where low-cost text categorization is required.

III. PROBLEM STATEMENT

In microblogging's websites messages of followees are listed in reverse chronological order for a user to read. With a large number of messages appearing in the interface, people often do not have patience to read every message. A collateral problem with many messages is that there are often different focal topics which includes diverse interests of users. It prevents a user from moving to what he/she is interested in.

We now formally define two major tasks in the problem of enhancing accessibility of microblogging's messages.

1)**Microblogging's Message Clustering:**Let $M = m_1, m_2 \ldots\ldots, m_n$ be a corpus of n microblogging's messages. Among these n messages, there are k different topics. We aim to cluster the n messages into k clusters c_1, $c_2 \ldots\ldots\ldots, c_k$.

2)**Cluster Labeling:**For each cluster c_i, we aim to generate human readable cluster labels$l_1, l_2 \ldots\ldots l_k$ in which most frequently used word is treated as a cluster label.

IV. PROPOSED SYSTEM

To make a large collection of microblogging messages accessible to users, our system will provide clusters and also labels for each cluster. The users will be able to quickly identify messages of interest by examining the cluster labels.

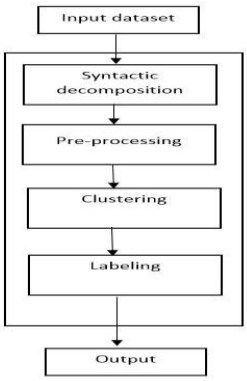

Fig 4 : Text clustering and labeling (TCL)

4.1 Input Dataset

Data is collected from Facebook and twitter as input to TCL system. Step to collect the data from face book

1) Register in Facebook.developer.com
2) Login with id and password and wait till binding process is complete
3) Token Id is generated
4) Collect the data of individual

4.2 Syntactic Decomposition

It is the task of breaking the sentence into pieces called tokens (words). Once the input data set is collected, each sentence is divided into tokens.

Input: Hello everyone, wish you happy New Year. This is Sameer

Output: Hello, everyone, wish, you, happy, New, Year, This, is, Sameer.

Further this will be useful in pre- processing.

4.3 Data Pre-Processing

Data pre-processing is a task of transforming the unstructured data into structured via predefined dataset. Ex: @=at, nt=night etc. The structured data is then used for clustering.

Algorithm:

{

1. ***Specify the corpus location***
2. ***Divide each sentence in the document into tokens***
3. ***Transform the unstructured data into structured via predefined dataset i.e., find and replace meaningless words with relevant meaningful words from predefined dataset***

}

4.4 Clustering

Natural language processing (**NLP**) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Using NLP technique clustering can be done.

Clustering[8][7]is the process of grouping set of objects in such a way that objects in the same group (called a cluste*r*) are most similar (in some sense or another) to each other than to those in other groups (clusters). In this

paper, concept of clustering for grouping the microblogging's site messages is done by using the NLP technique, **Named Entity Recognition (NER).** By using this, it is easy to group the sentences based on named entities (Person, Location, Date, Time, etc) .

Algorithm:

{

1. ***Collect the data from the pre-processing***

2. Perform cluster concept of NER

{

1. Tokenize sentence in to word and perform pos-tagging

2. Identify the entities

}

3. Classify text based on entities recognized so that each sentence will be in one or the other entity group

}

NER[10][11] have different category based person-per, Location-Loc Date and Time and Miscellaneous-Misc

Ex : Sam in Bangalore

Sam- Person, Bangalore—Location

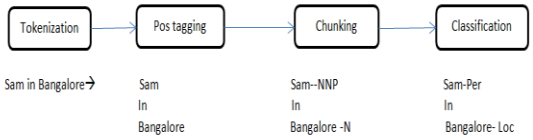

Fig 4.1: Example for NER

Fig explains clearly how to identify entities categories:

Tokenization - dividing the sentence into words

Ex: sam in Bangalore

Sam, in , Bangalore

Postagging - find NNP,NN,VP

Ex : Sam in Bangalore

Sam- NNP

Bangalore-NP

Chunking - After the POS-tags find Phrase that constitute entity

Classification : Words are tagged with named entity

Ex: Sam in Bangalore

Sam-Person

In-O

Bangalore-Location

(O-not any of the entity)

4.5 Labeling

Each cluster is assigned with a label which is the most frequently used word in the cluster. This enables the user to quickly identify the cluster and easily go through the text of his/her interest.

V. CONCLUSION

By analyzing the structure of microblogging messages, the original short and noisy texts are mapped into predefine BAG-OF-WORD (BOW) to improve the quality of text representation. Clustering and labeling helps the user to easily go through the microblogging messages.

VI. FUTURE ENHANCEMENT

Improve the quality of microblogging messagesfor clustering use good pre-processing algorithm.To getmeaningfulwords in cluster and use hierarchicalclustering algorithm to get accurate clusters.

REFERENCES

[1] Xia Hu, Lei Tang, Huan Liu, IEEE "Embracing Information Explosion withoutChoking: Clustering and Labeling in Microblogging," IEEE TRANSACTIONS ON BIG DATA, JANUARY 2015.

[2] Y. Song, S. Pan, S. Liu, M. X. Zhou, and W. Qian, "Topic and keyword re-ranking for lda-based topic modelling ," in Proceedingsof the 18th ACM conference on Information and knowledge management. ACM, 2009, pp. 1757–1760.

[3] Y. Hu, A. John, F. Wang, and S. Kambhampati, "Et-lda: Joint topic modeling for aligning events and their twitter feedback." in AAAI, vol. 12, 2012, pp. 59–65.

[4] L. Hong and B. D. Davison, "Empirical study of topic modeling in twitter," in Proceedings of the First Workshop on Social MediaAnalytics. ACM, 2010, pp. 80–88.

[5] D. Ramage, S. Dumais, and D. Liebling, "Characterizing microblogs with topic models," in International AAAI Conference on Weblogs and Social Media, vol. 5, no. 4, 2010, pp. 130–137.

[6] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Identifying influencers on twitter," in Fourth ACM International Conference on Web Seach and Data Mining (WSDM), 2011.

[7] D. D. Lewis and W. B. Croft, "Term clustering of syntacticphrases," in Proceedings of the 13th annual international ACM SIGIRconference on Research and development in information retrieval.ACM, 1989, pp. 385–404.

[8] Logamani.KandPunitha. S. C "Density Based Clustering using Enhanced KD Tree" in International Journal of Computer ScienceEngineering and Technology(IJCSET) | November 2014 | Vol 4, Issue 11,314-318.

[9] http://www.nltk.org/book/ch07.html

[10] Rami Al −Rfou'StevenSkiena "SpeedRead: A Fast Named Entity Recognition Pipeline"Proceedings of COLING 2012: Technical Papers, pages 51–66,COLING 2012, Mumbai, December 2012.

[11] Dan Roth and Wen-tau Yih "Global Inference for Entity and RelationIdentification via a Linear ProgrammingFormulation" ebook .

[12] YoungjoongKo and JungyunSeo"Automatic Text Categorization by Unsupervised Learning".