SURVEY ON CLUSTERING AND LABELING IN MICROBLOGGING

Hage Dipmala[1], Anil Kumar Ambore[2]
[1]Department of Computer Science and Engineering, REVA ITM, Bangalore
[2]Assistant Professor, Dept of CSE, REVA ITM, Bangalore

Abstract: The growing popularity of microblogging is also exponential to the generation of the data in the microblogging services. Hence, for example if we consider one user for reference purposes, it becomes very cumbersome for him to timely analyze the messages and inputs/opinions of his bloggers/fans or followers if a very large volume of data is generated. In this literature review we have viewed and compared different techniques used to give solution to organize the data into meaningful and convenient clusters so that an overview is provided for the user or the blog owner so that he can read and analyze his followers opinions in very short duration without much confusion and in a systematic way. Now, since blog messages are generally shorter and more complex than other text documents and messages it is quite cumbersome to cluster and label these. Another reason for the difficulty in clustering the blog messages are since they are short, semantic associations become vague due to the lower term recurrence in the messages. Different techniques and methods have been proposed and implemented to solve the problem in microblogging so that a user can easily access his required data without having to go through corpus. Hence, in this survey we talk about different types of proposed methods for clustering and labeling of short text that can be applied to microblogging messages and provide the merits and demerits in each approach.

Keywords: Blogging, Clustering, labeling, knowledge base, semantic mapping, semantic meaning.

I. INTRODUCTION

Microblogging services have become very inherent to today’s masses and internet users where users publish snippets of information about their daily life and activities which has become the latest trend or fad. Online micro-blogging and social network services such as twitter, tumblr, facebook etc provides us an alternative way of communicating with friends and people. In these sites millions of people can see any posts i.e. mostly in short texts also called micropost or tweets or status that the blog owner post. Blogger can blog on any topic which can range from their personal life and work to current news, events, mood and thoughts. The tweets or microposts are published on the authors personal page or blog for their followers to see and also where they can comment on each update. Among others twitter has become one of the most notable microblogging or social networking service today with millions of users generating millions of tweets per day. In short such micro blogging services provide a real time environment to reflect what people are thinking and what inputs/opinions they want to publish. Hence, these services provide a new platform for understanding human behavioral processes and also a new dimension for assessing or analyzing these behaviors. With the increase in the number of users in micro blogging services number of people blogging increase and their followers commenting on each tweets or status increases large amount of data. One drawback of popular modern microblogging services is that since a large amount of data is generated every minute, it becomes quite difficult for the user to find the useful and important data which can be hidden and obscured in the large amount of live streaming data. Hence, to solve this issue, web systems need to provide a meaningful method of grouping or clustering of these obscured data and subtopics and also give them cluster label name using a suitable and convenient method of labeling. So that a user can quickly identify messages of his interest and need by just examining an overview of the subtopic from their label. The characteristics of these live streamed data in the micro blogging sites is very different from a normal text document so it becomes difficult to apply traditional text analytic methods on them. Firstly, because microblogging messages are very short in length, this it becomes difficult for conducting similarity measurement. For example, twitter allows tweets of length not more than 140 characters. Secondly because, unstructured data is very often used in microblogging sites such as slang words or message lingos makes it difficult for text processing. Using such informal words makes for convenient and easy communication but makes it difficult to apply traditional text processing method. Methods such as Bag of words (BOW) [1] and text mining, used to represent text for topic modeling has been extended to micro blogs also [2]. However these methods are not sophisticated enough to perform textual analysis on microblogging messages where data are unstructured and are of complex semantic meaning. This paper looks into and discusses many methods and approaches that were used to enhance the message presentation by using techniques such as mapping techniques to map messages from an unstructured form to a semantically knowledgeable space, feature generation for text categorization, using probabilistic knowledge base for short text conceptualization and many more which are discussed in the next section related work and studies.

II. RELATED WORK AND STUDIES

Xia Hu et al. [3] proposed a paper trying to solve the problem where a blogger has to scan through whole data set that he get as a response from his/her followers for posts/blogs that he updates. Now going through whole data set every single time can be overwhelming, time consuming and tedious. This can lead him to stop reading through all those comments midway and thus he might lose onto some useful comments. This approach presents or defines a text representation framework by utilizing the power of semantic knowledge base of Wikipedia and wordnet too in which related texts are matched with the semantic representation. Using this technique improves and enhances the performance measure of the clustering and labeling of short texts which are usually hard to do using traditional text categorizing techniques or existing system. The proposed approach involves decomposing and parsing of text into different pieces and integrates the corresponding external knowledge base so that the performance of clustering for microblogging messages can be improved.

This technique addresses the problem of enhancing accessibility of a huge quantity of microblogging messages/comments and proposes to utilize flat clustering algorithm and labeling to solve the problem. It is one of the best way to instantly generate labels for each cluster by ranking concepts which are structured from the knowledge base of Wikipedia. But use of Flat clustering method in this approach produces way too many clusters and hence, many comments on each cluster as well. And also the method of labeling which is done in this technique is purely based on the most occurring term and is not actually done with any reference to the content of the cluster.

H.-J. Zeng et al. [4] proposed a method to address the problem of clustering in web search results so that a user can quickly browse through all the search results that is displayed when he is searching for something of his required data in the internet. This approach characterizes the problem of clustering as a salient phrase ranking problem. The method used here comprises of converting the unsupervised clustering problem to a supervised clustering problem which is a concept from machine learning. Suppose we give a query and its corresponding ranked list of document which is nothing but a list of titles and snippets which have been returned by certain web search engines, then what this method does is first extract and rank the salient phrases as candidate cluster labels, based on regression model learned from or given by a human labeled training data. The merits in this approach is that the cluster generated here are more readable, understandable and shorter when compared to cluster generated by the existing systems. But the demerit here is that it can work efficiently only with supervised learning method i.e. it requires additional training data which generates substantial overhead.

H. Chim and X. Deng [5] in this paper presents a concept-based document similarity to determine the similarity of documents by making use of a model which encodes the information about word order called suffix tree document model (STD). This is done by mapping each and every node in the suffix tree into a unique feature term in vector space document (VSD) model. The concept-based document similarity make use of and inherit the terms ctf(conceptual term frequency), tf(term frequency) and df(document frequency) weighting scheme in computing the document similarity with the help of concept. In this approach the aforementioned technique is applied to an algorithm called hierarchical agglomerative clustering (HAC) to define a new type approach for document clustering. This model analyzes the term based on its occurrence in the sentence, document and in mass/corpus level. The similarity is determined based on the new concept-based similarity models (Euclidean distance measure). The algorithm used here is called concept based analysis algorithm. The main advantage of this approach when compared to other existing approach is that the technique used here is solely based on semantic structure i.e. it can produce significant concept based clusters. Thus resulting in clusters with better meaning since it determines similar concepts between documents which in turn leads to better clustering results. But this approach is not so efficient for brief textual messages and fails to provide significant labels to the clusters.

A. Tagarelli and G. Karypis [6] in their paper, address the problem of grouping or clustering multiple topic document. This approach make use of natural composition of document in text segment which can consist of multiple topic on their own i.e. they proposed a segment-based document clustering framework intended to activate documents classification firstly by identification of connected group of segment-based portion from its original document. Based on the underlying 'n' number of topics of the document each document in the corpus is modeled with a set of segments, which have been identified. Next step, using a document clustering algorithm the segment sets from all documents are clustered based on segment-based approach. Again from the segment set clustering a possible overlaying classification of the original documents can be or is induced. Using this approach every document can be assigned to more than one cluster according to their topic. This method has efficiently improved the identification of the various topics of each document and allows the assignment of documents into multiple clusters based on their topic. But the solution resulting from this approach can deteriorates when overlap of segment set increases and gives low interpretability.

Y. Song et al. [7] proposed a technique in which understanding of text is improved by making use of probabilistic knowledge base that is as wealthy as our metal world which contains the concepts of worldly facts. This approach use a method of conceptualizing short text using a probabilistic knowledge base where terms are mapped and detected in short text to the attributes and characteristics in the knowledge base. Then using Bayesian inference the most likely concept is derived. Here Bayesian inference is used to conceptualize short texts or words. This conceptualization techniques can be used to cluster data's in microblogs and social network such as tumblr, twitter, facebook etc and other microblogs. It provides high interpretability of the clusters generated but has drawback where multiple classes of concepts make the feature indiscriminative.

E. Gabrilovich and S. Markovitch [8] proposed a method which improves the machine learning algorithm for categorization of texts with generated feature based on domain specific and common sense knowledge. The aforementioned knowledge is represented using publicly available ideas that consist of huge number of concepts such as open directory, these ideas or logic are further improved by several orders of magnitude through controlled web crawling. This method is an alternative solution that makes use of the power of existing induction techniques whilst also enhancing the language of representation. Before doing text categorization, this approach employ a feature generator that uses common sense and domain specific knowledge to improve the bag of words with new and more information features. Feature generation is performed unsupervised i.e. on its own using machine readable hierarchical repositories of knowledge base. This knowledge based feature generation brings text categorization improves the performance to a very high level. But the disadvantage in this approach is that as the data increases more number of features are generated and use of more generated feature has adverse effect on the text categorization using this approach.

D. Carmel et al. [9] proposed a method to improve the phenomenon of clustering and labeling using knowledge base Wikipedia. The authors of this paper proposed a method in which the words in the documents are parsed, tokenized and represented as term vectors. Here the weight or value of each term is calculated using tf-idf. After weights are evaluated they are indexed using suitable indexes. Next step is to cluster the document. The most significant terms are extracted from the document and then the related Wikipedia pages are searched for the candidates for the most significant or most weighted terms. The last step in this method is extraction of label. The most important term in the cluster is taken as the cluster labels i.e. top ranking candidate is taken as the label. This approach is very robust and can withstand noise. But the problem with this approach is that the topics which are not covered by Wikipedia may affect the system performance since it won't have its reference.

X.-H. Phan et al. [10] proposed a framework for building classifiers that can be used with short or sparse text and web segments by making the most secluded topics discovered from a corpus of data. The underlying idea of framework here is that for each task of classification, they call a large scale external data collection called universal dataset, and then construct a classifier on both a small set of labeled training and a huge set of secluded topics discovered from the corpus. The framework is universal enough to be applied to various data genres ranging from educational to entertainment. This approach gives very high accuracy even with short and sparse data's and also this framework is easy to apply even though there is combination from different components. Since this framework works as semi-supervised model it works well only with small size of training data.

We summarize the advantages and disadvantages of each work below:

Work	Advantages	Disadvantages
[3]	Simultaneously generate textual label for each cluster	Cluster label don't reflect the content of the cluster
[4]	Generates clusters with highly readable/understandable/shorter labels	Requires additional training data which generates substantial overhead
[5]	Gives better and meaningful cluster	Fails to provide meaningful labels
[6]	Improved identification of various topic for better clustering	Low interpretability
[7]	High interpretability	undiscriminative feature due to multiple class
[8]	Feature generation done automatically	More data leads to detrimental effect on text categorization
[9]	Robust and can withstand noise	Dependent on Wikipedia
[10]	High accuracy even with data sparseness	Require training data

III. CONCLUSION

The paper summarizes the current works in the clustering and labeling of large number of text messages used in micro blogs into a cluster with meaningful labels. We have seen that there are some framework which has been proposed has successfully enhanced the accessibility of microblogging messages by utilizing semantic concepts, knowledge base and others. In some approach Wikipedia has been used as to map the original noisy texts into a semantic space to improve the quality of text representation. With the help of the knowledge base of Wikipedia, the task of cluster labeling was solved without much cost. All the paper discussed here tries to make retrieving of required data easy and quick among zillions of other data.

REFERENCES

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

[2] 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. Xia Hu, Member, Lei Tang, Member and Huan Liu,

[3] "Embracing Information Explosion without Choking: Clustering and Labeling in Microblogging," IEEE transactions on big data, January 2015, In Press

[4] H.-J. Zeng, Q.-C. He, Z. Chen, W.-Y. Ma, and J. Ma, "Learning to cluster web search results," in Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2004, pp. 210–217.

[5] H. Chim and X. Deng, "Efficient phrase-based document similarity for clustering," Knowledge and Data Engineering, IEEE Transactions on, vol. 20, no. 9, pp. 1217–1229, 2008.

[6] A. Tagarelli and G. Karypis, "A segment-based approach to clustering multi-topic documents," in Text Mining Workshop, SIAM Datamining Conference, 2008.

[7] Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen, "Short text conceptualization using a probabilistic knowledgebase," in Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three. AAAI Press, 2011, pp. 2330–2336.

[8] E. Gabrilovich and S. Markovitch, "Feature generation for text categorization using world knowledge," in International joint conference on artificial intelligence, vol. 19. LAWRENCE ERLBAUM ASSOCIATES LTD, 2005, p. 1048.

[9] D. Carmel, H. Roitman, and N. Zwerdling, "Enhancing cluster labeling using wikipedia," in Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2009, pp. 139–146.

[10] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in Proceedings of the 17th international conference on World Wide Web. ACM, 2008, pp. 91–100.