connecting engineers... developing research

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Special Issue

National Conference on "Recent Trends, Advancement and Applications of Digital Image Processing" (NCDIP 2016)

Information Extraction from Microblog using Keyword Extraction and Clustering algorithms: A Survey

[1] Archana Sidhanti [2] Dr. S G Totad
[1][2] School of Computer Science and Engineering
KLE Technological University,Hubballi
Karnataka , India

***Abstract:-* With the rapid development of the Internet, Microblog is becoming one of the most popular social network platforms. Microblogging e.g. Twitter has attracted millions of users to share and disseminate most up-to-date information, resulting in large volumes of data produced every day. The large quantum of information makes it necessary to find out methods and tools to summarize them. This survey paper tackles a comprehensive overview of the last update in this field. Many recently proposed keyword extraction and clustering algorithms and various enhancements are investigated and presented briefly in this survey.**

***Index Terms*— clustering algorithms, information extraction, keyword extraction, microblog**

I. INTRODUCTION

The World Wide Web network is remarkable developing as a simple and inexpensive means to share information, and various people have been expressing their opinions from different standpoints on the web. Currently, micro-blog has emerged to be a popular web application fo billions of individuals sharing and spreading news and their emotions. Micro-blog post usually offers a large amount of short text about real events happened in social life.

Recently there have been increasing interests in events and topics detection over microblogging, especially on Twitter. Various systems with different purposes, structure and detection algorithms have been developed, such as detecting real-time earthquakes over Twitter , detecting emerging topics by modeling life cycle of key words , even classification approach with utilization of spatio-temporal information carried by microposts, 'TwitterMonitor' trend detection system that treats bursting keywords as entry points,

1) mining activity networks to identify interesting social events , as well as emerging topics early detection based on outlier and hashing techniques.

2) With more and more people using microblog, much information spread very fast in network. The amount of people who use Sina microblog is over 5 hundred million, and 45 million people use it every day, there are 2 hundred million microblogging content increased every day. It is really a nightmare to find information of interest through the huge amount of available posts that are often noisy and redundant.

3) In the era of Big Data, social media analytics services have caught increasing attention from both research and industry. The purpose of our work is to conduct a survey about existing research on information extraction from microblog i.e Twitter. We have discussed various keyword extraction and clustering algorithms and compared the same. We are using F1 score a measure of test's accuracy to compare keyword extraction algorithms. Clustering algorithms are compared using number of clusters and size and length of the dataset.

A. Information Extraction Process

Information extraction (IE) is the task of automatically extracting structured information from unstructured and/or semi-structured machine-readable documents. We are proposing various techniques to extract information from the social media specifically twitter. Twitter is currently the most popular micro-blogging service. It is a rich and real-time information source and a good way to discover interesting content or to follow recent developments.

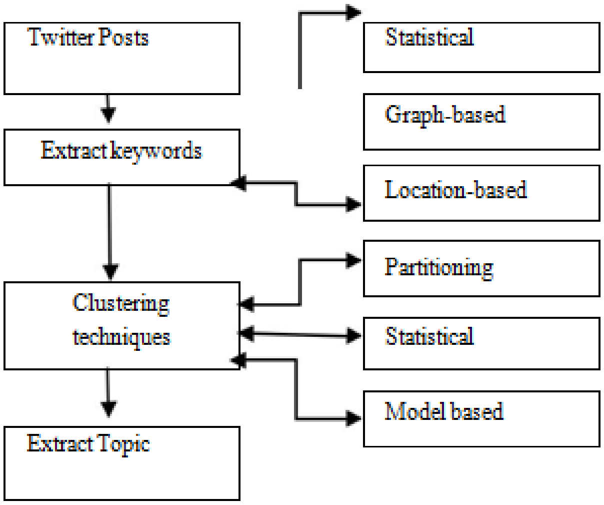


Fig1. Information extraction from Twitter

Information extraction from twitter can be considered as shown in the fig 1. It explains

1. Microposts generated within monitored subnet nodes are fetched using Twitter API.

2. Important keywords are extracted. Keyword extraction is a subtask of information extraction, with the goal to automatically extract relevant terms from a given corpus. Here we have categorized keyword extraction algorithms into 3 categories.

a) Statistical method
b) Graph based method
c) Location based method

3. Clustering is the process of grouping similar objects. After extracting important keywords one needs to cluster the posts containing those keywords. This is very important to extract information from twitter. We have categorized clustering algorithms into 3 groups. They are

a) Partitioning
b) Statistical
c) Model-based

B. Applications of the information extracted from microblogs

Microblog owns potential value for company to increase communication with customer, improve the relationship between company and customer and become part of marketing strategy. Many companies, organizations and those who are interested in customer relationship have paid attention to microblog represented by Twitter. Some famous companies and organizations, such as Dell and Starbucks, are using twitter as a tool to communicate with customer effectively. Twitter has become the major platform for Dell to communicate with customer as Twitter provides Dell the opportunity to make instant feedback.

Twitter has become a dependable microblogging tool for real time information dissemination and newsworthy events broadcast. Its users sometimes break news on the network faster than traditional newsagents due to their presence at ongoing real life events at most times. Different topic detection methods are currently used to match Twitter posts to real life news of mainstream media. The information extracted from twitter can be utilized in many applications like detecting online trendy topics, to find out web public opinion[8] and to find out important events like Mariam Adedoyin-Olowe et al. tried to analyze tweets relating to the English FA Cup finals 2012. A recent survey found that consumers trusted more websites with reviews than professional guides and travel agencies blogs are often perceived to be more credible and trustworthy than traditional marketing communications for tourism.[11] Nicola Perra and Bruno Gonçalves describe tools that, using search queries, microblogging, or other web-based data, are able to predict the incidence of a wide range of diseases two weeks ahead respect to traditional surveillance.[12]

Twitter is a potential tool for monitoring and managing crisis and convergence events, if accurate spatiotemporal information related to the events can be derived. B. D. Longueville et al. tried this usage by mining tweets to track forest fires in Marseille, France and found that the timeline of tweets did accurately match the real-world spread of the fire, except for a lag time at the be- ginning of the fire". T. Sakaki et al. and M. Guy et al. took this application a step further by treating Twitter users as social sensors to realize early detection and warning of potential emergent situations. [6]

II KEYWORD EXTRACTION

Keywords can express the core content of a document. These words commonly are used to index the feature of thesis in computer systems, especially in information retrieval system. Automatic extraction of keywords can help people understand contents better on the web. This method is the basic job for the web resources management, and facilitates users to browse results or document collection after searching. A general keyword extraction mainly covers steps such as identifying keyword candidates, weighing each candidate, and selecting the keywords with the higher weights. Determining the weights of the candidate word has become the core of keyword extraction. The weight of the candidate word depends on the importance of the topic literature. The words which can reflect the topic of the literature will be given greater weight. The more the weight of a word is, the more independent the word can represent a category.[1]

Generally there are two kinds of keyword extraction methods : supervised methods and unsupervised methods. The main idea of the former is to train a keyword extraction model based on the part of speech, location, and so on. And then use the model to extract the keywords from the micro-blog texts. The shortcoming of supervised method is that it needs the training corpus, especially for micro-blog.

Generally speaking, keyword extraction can be categorized as

1. Statistical method based on word frequency (TFIDF)
2. Method based on word co-occurrence
3. Based on location

A. Statistical method based on word frequency

Here importance of a word is directly proportional to the number of occurrences of a word in the document. We are discussing three algorithms in this category TF-IDF, TF-PDF and Improved TF-IDF algorithms.

1) TF*IDF - TF-IDF (Term Frequency-Inverse Document Frequency) is widely used in information retrieval and text mining as a statistical method for assessing the importance of words in a set of files or a corpus. The importance of a word in the document is proportional to the number of occurrences in the document, and is inversely proportional to the count of documents containing this word in the corpus. The main idea of TF-IDF is: if a word or phrase appears frequently in an article,that means the TF value is high. However, this measure method is still inadequate in reality, mainly because it does not take the total number of keywords in the document into account. For example, the importance of a word appears 10 times in a 100-word document, should be higher than the word appears 20 times in a 1000-word document. In fact, if a term in one category of documents frequently appears, it indicates that the term can represent the characteristics of this type of text. Then this term should be given higher weight, and be chosen to be feature words to distinguish this document from other categories. [1]

2) TF-PDF (Term Frequency - Proportional Document Frequency). This is emerged in order to make up the shortcomings of TF-IDF. In TF*PDF algorithm, the weight of a term from a cluster is linearly proportional to the cluster’s frequency containing this term, and exponentially proportional to the frequency of document containing the term in the cluster. The total weight of a term will be the sum of term’s weight from each cluster.[1]

3) Improved TF-IDF Algorithm. There are many drawbacks in the conventional TF-IDF (term frequency - inverse document frequency) function. All the text in the text set to be considered as a whole in the conventional TF-IDF. Especially IDF part, just consider only the relationship between the keywords and the number of text that appears and ignores the distribution in a category. For example, if one word only appears in large numbers in individual documents within a certain category, but in most of the other documents within the category is rarely present, then does not rule out the word is a special case of this category. So that the words do not have representation, it stands to reason that the weights should be relatively low, but for this case, the traditional TF-IDF function is not better reflected. Through traditional TF-IDF weighting algorithm combined with within-class distribution of degrees proposed new improved TF-IDF weighting function: it considers both weight and degree of within-class distribution DIDC.[4]

B. Method based on word co-occurrence graph

Graph model is another commonly used unsupervised method, which has been proved effective in extracting traditional texts. So, in order to extract keyword effectively, graph model is adopted to express the relations between the words, and create the graph based on the co-occurrence.

1) Graph Model TFIDF is a commonly used method for finding term frequencies. But the term frequencies of many words are all equal to1 because the micro-blog is too short. So the performance of extraction is not so good when we only use TFIDF as the feature. A directed graph is adopted to represent the co-occurrence relation between the words. The graph is recorded as V, where V is the set of the vertexes: represent a word of the micro-blog; E is the set of the edges. Then weight is calculated based on the distance. If the difference between the sequence numbers of two words is smaller than distance, then the two words are co-occurrence, or else, they are not co-occurrence. [5]

c. Method based on location of a word

It is based on the position of a word in the microblog post. It has another additional advantage: it can be used to detect the wrongly segmented words, which can help the post-process after the extraction.

Table I

Keyword extraction methods	F-Score
Statistical method	0.5020
Graph based method	0.4999
Location method	0.5507

1) Semantic space and the location feature If words have the same weight, the graph model itself can't decide whether they are keywords. So it is necessary to adopt other features to help to extract the keywords.To solve this problem statistical weight based on the creation of the semantic space and the location feature is proposed. Due to the limitation of the length, most micro-blogers come straight to the main content when they write micro-blog. So, the words locating in the former part is usually more important than the words locating in the latter part. Based on this, location is adopted as an important extraction. The word with a smaller location number will get a larger ranked value. Then TOP-N is adopted to extract the keywords based on the ranked values firstly sort the keywords in descending order according to the ranked values, and then select the first N words from the beginning of the queue as the keywords.[5]

III CORPUS AND EVOLUATION METRIC

To evaluate these categories of algorithms we adopt recall and F1-score as the evaluation metrics to compare keyword extraction algorithms and are computed as follows

Precisioni=c/m (1)
Recalli=c/n (2)
F1i=(2* Precisioni* Recalli)/(Precisioni+ Recalli) (3)

where, Precisioni, Recalli and F1i are precision, recall, and F1-score the system got when it extract the ith entry of micro-blog. c is the number of keywords that the system extract correctly from the ith entry of micro-blog, m is the number of keywords that the system extract (correctly or wrongly) from the ith entry of micro-blog, n and is the number of the keywords of the ith entry of micro-blog tagged in the corpus. N is the total number of micro-blog texts.

Once we get the evaluation result for each entry of micro-blog, the final evaluation result of the system is the average of the evaluation results of the 200 entries of micro-blogs, which are computed as follows:

$$\text{Precision}=\sum^{200}\text{Precision}_i \quad (4)$$

$$\text{Recall}=\sum_{i=1}^{200}\text{Recall}_i \quad (5)$$

$$\text{F1-Score}=\sum_{i=1}^{200}\text{F1}_i \quad (6)$$

Table I shows comparison of three categories of algorithms by considering F-Score and N i.e. number of microblog texts as constant N=8

By this we can say that the location feature is a very effective feature in the keyword extraction. The reason is that due to limitations of words, most micro-blog users are usually come straight to the main content, that is to say that the keywords of a micro-blog will usually locate in the beginning of the micro-blog.

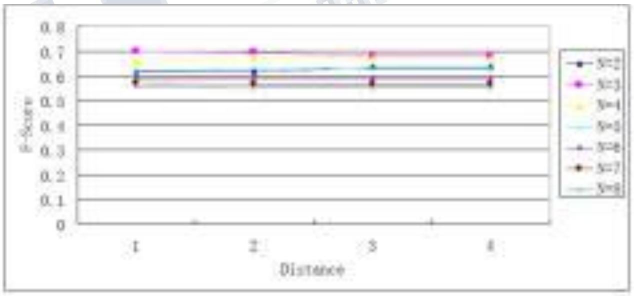


Fig 2 Graph to compare Distance and F-Score

In order to do some further analysis to our results, we now give the following comparison diagram (Figure 3). From the diagram, we can see that the system performs better when N=3, 4, the reason is that the micro-blog is relatively shorter, and it does not need to use more words to express its content. On the other hand, the influence of distance on the performance is small. Here distance is threshold value that is defined as if the difference between the sequence numbers of two words is smaller than distance, then the two words are co-occurrence, or else, they are not co-occurrence.[5]

IV CLUSTERING

To discover the user interest, only keyword extraction is not sufficient. One needs to cluster those tweets. Clustering is a process of creating groups of similar objects.

Clustering algorithms are categorized into three major categories namely,

1. Partitioning techniques
2. Statistical techniques
3. Model based techniques

A. Partitioning techniques

These are most popular class of clustering. These algorithms minimize a given clustering criterion by iteratively relocating data points between clusters until a (locally) optimal partition is attained.

1) k-means algorithm The k-means algorithm is a partitioning based clustering algorithm. It takes an input parameter, k i.e. the number of clusters to be formed, which partitions a set of n objects to generate the k clusters. The algorithm works in three steps. In the first step, k number of the objects are selected randomly, each of which represents the initial mean or center of the cluster. In the second step, the remaining objects are assigned to the cluster with minimum distance from cluster center or mean. In the third step, the new mean for each cluster is computed and the process iterates until the criterion function converges. K-means clustering algorithm is a method of vector quantization, aiming to partition n observations into k clusters in which each observation belongs to one cluster with the nearest mean. That is to say, the k-means algorithm uses the initial data to construct the closable cluster members, and then adjust the center until the sum of squares reaches the minimum.[1]

2) 2 - phase iteration feature selection Aiming at the problem of k-means that the value k is inputted manually, a method about how to find an optimal k value is proposed according the variable and feature of dataset. Because the center of a cluster based on k-means is easily effected, average similarity of a cluster is used as a parameter. Using the external knowledge from WordNet, an improvement of standard k-means is present. Rather than only using standard clustering algorithm, a new text cluster strategy is proposed in order to remove the non-discriminative general terms. The k-means only considers the relationship between an observation point and the corresponding cluster's center. In other word, it ignores the global information of the cluster. Another drawback of the k-means is that it is sensitive to outliers.

In order to avoid the above drawbacks of the k-means, 2-phase iteration feature selection is introduced to improve the performance. This is feasible for the clustering, especially for the mass data. Here after extracting the features, it computes the distance between the observation point and the feature set. If there are several mutual exclusive feature sets, an observation must have a minimum distance among the sets. Obviously, these feature sets can represent the corresponding cluster, as they are extracted from different clusters.[2]

B. Model based techniques

Typically the data are clustered using some assumed mixture modeling structure. Then the group memberships are ‘learned’ in an unsupervised fashion.

The data are collected from a finite collection of populations. The data within each population can be modeled using a standard statistical model.

1) Latent Dirichlet allocation Latent Dirichlet Allocation (LDA) is a popular topic modeling technique which models text documents as mixtures of latent topics, which are key concepts presented in the text. A topic model is a probability distribution technique over the collection of text documents, where each document is modeled as a combination of topics, which represents groups of words that tend to occur together. LDA estimates the topic-term distribution and the document topic distribution from an unlabelled collection of documents using Dirichlet priors for the distributions over a fixed number of topics.

2) Author – Topic During the LDA learning process, distribution of words into each topic is estimated automatically. Nonetheless, the location associated with the tweet is not directly taken into account in the topic model. As a result, such a system considers separately the tweet content (words), to learn a topic model, and the labels (location) to train a classifier. Thus, the relation between the tweet content and its location (country) is crucial to efficiently locate (unknown) new tweets.

To solve these issues new topic model is built called Author Topic(AT).It takes into consideration all information contained in a tweet: the content itself (words), the label (country) and the relation between the distribution of words into the tweet and the location, considered as a latent relation. From this model, a vector representation in a continuous space is built for each tweet.[9]

C. Statistical techniques

It works on variables. It can cluster variables together in a manner somewhat similar to factor analysis. In addition, it can handle nominal, ordinal, and scale data, however it is not recommended to mix different levels of measurement.

1) TDA (Topic Detection using AGF) The aim of this algorithm is for generating the topics by clustering the frequent patterns. Frequent patterns are detected by using tf-idf. After finding out the frequent patterns using tf-idf, evaluate the AGF1 (Associative Gravity Force) values between each pattern pairs using two other parameters, Kr1 (Keyword rating) and CIMAWA1 (Concept for the Imitation of the Mental Ability of Word Association). Kr(x) and Kr(y) evaluate the importance of x and y respectively in one document, and CIMAWA(x(y)) evaluates the probability of co-occurrence of x and y together. AGF evaluates the attraction between x and y. If the AGF(x(y)) is large, that means, the attraction between x and y is very high, i.e., the chances of occurrences of x and y together is very high. So we cluster the words (frequent patterns) using these AGF values. In this way they are clustered. By using TDA, we created clusters of frequent patterns, which are named as topics. But these clusters may also suffer from the wrong correlation problem.[3]

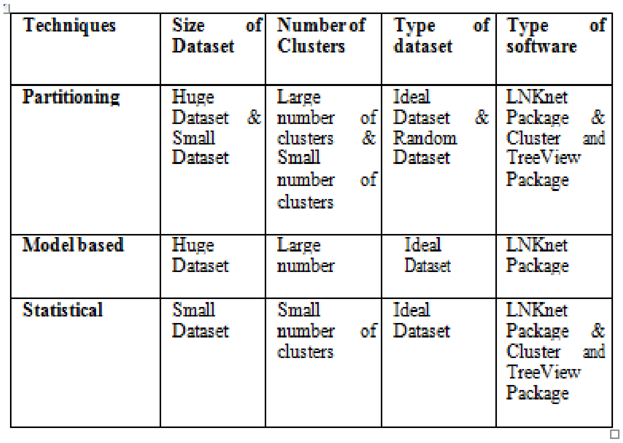
2) TCTR (Topic Clustering and Tweet Retrieval). This is the improvement of TDA. The aim of TCTR is to Cluster the topics and return the corresponding tweets as final outputs. Firstly sort all topics in the descending order of their number of words. The sorting is done to obtain a better clustering. Then clustering is done by finding the common number of words in each pair of topics. Avoid the wrong correlation of patterns by retrieving the tweets, which contain all the words in one topic. Lastly repeated tweets are removed.[3]

V COMPARISON OF CLUSTERING TECHNIQUES

Clustering algorithms are compared according to the following factors:

1. The size of the dataset
2. Number of the clusters
3. Type of dataset
4. Type of software

Table II



Techniques	Size of Dataset	Number of Clusters	Type of dataset	Type of software
Partitioning	Huge Dataset & Small Dataset	Large number of clusters & Small number of clusters	Ideal Dataset & Random Dataset	LNKnet Package & Cluster and TreeView Package
Model based	Huge Dataset	Large number	Ideal Dataset	LNKnet Package
Statistical	Small Dataset	Small number of clusters	Ideal Dataset	LNKnet Package & Cluster and TreeView Package

Partitioning technique is usually suited for both huge dataset and small datasets. In model based technique data are clustered using some assumed mixture modeling structure, it is suited for huge data set. Statistical method works on variables and it uses factor analysis so it is suited for small datasets. In Partitioning technique large and small clusters are formed. But in model based technique large number of clusters is formed and in statistical technique small number of clusters is formed. All techniques work on ideal datasets and use LNKnet Package & Cluster software. Clustering is a descriptive technique. The solution is not unique and it strongly depends upon the analyst's choices. When applying a cluster analysis we are hypothesizing that the groups exist. But this assumption may be false or weak. Clustering results should not be generalized. Cases in the same cluster are similar only with respect to the information cluster analysis was based on i.e., dimensions/variables inducing the dissimilarities.[10]

VI CONCLUSION

News spreading capability of social networking sites is really high. In order to harness this ability of social networking sites there is lot of research work going on in this field. This paper starts at the basic applications of microblogs especially Twitter, how to extract information from Twitter, basic keyword extraction and clustering algorithms and analysis the keyword extraction algorithms.

The main purpose of this paper is to introduce basic and core ideas of commonly used algorithms specify the source of each one, and analyze the advantages and disadvantages of each one. It is hard to present a complete list of all algorithms due to the diversity of information, the intersection of research fields and the development of modern computer technology. So important categories of the commonly used keyword extraction and clustering algorithms, with high practical value and well studied, are selected and one or several typical algorithm(s) of each category is(are) discussed in detail so as to give readers a systematical and clear view of the importance of information extraction from Twitter.

ACKNOWLEDGMENT

This work is supported by School of Computer Science and Engineering, KLE Technological University, Hubballi, Karnataka, India.

REFERENCES

1. Zhiqi Fang1, Yue Ning2,Tingshao Zhu2 "Hot Keyword Identification for ExtractingWeb Public Opinion" 2010 IEEE.

2. Kai Gao, Bao-quan Zhang "Modelling on Clustering Algorithm Based on Iteration Feature Selection for Micro-blog Posts" 20 14 International Conference on "Modelling, Identification and Control " IEEE.

3. Amrutha Bennya,*, Mintu Philipb "Keyword Based Tweet Extraction and Detection of Related Topics" International Conference on Information and Communication Technologies (ICICT 2014) ELSEVIER

4. Xing Huang, Qing Wu "Micro-blog Commercial Word Extraction Based On Improved TF-IDF Algorithm" 2013 IEEE

5. Hua Zhao and Qingtian Zeng " Micro-blog Keyword Extraction Method Based on Graph Model and Semantic Space", JOURNAL OF MULTIMEDIA 2013 IEEE.

6. SUI Yue, YANG Xuecheng "The Potential Marketing Power of Microblog" 201O Second International Conference on Communication Systems, Networks and Applications .IEEE

7. Neethu Kuriana*, Shimmi Asokana "Summarizing User Opinions: A Method for Labeled-Data Scarce Product Domains" International Conference on Information and Communication Technologies (ICICT 2014) ELSEVIER

8. Kushal Bafna, Durga Toshniwal "Feature Based Summarization of Customers' Reviews of Online Products" 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems -KES2013 ELSEVIER

9. Mohamed Morchida, Yonathan Portillaa,b, Didier Josselinc,a, Richard Dufoura, Eitan Altmanb,a "An Author-Topic based Approach to Cluster Tweets and Mine their Location" Spatial Statistics 2015: Emerging Patterns ELSEVIER

10. T. Soni Madhulatha "An overview on Clustering methods" IOSR Journal of Engineering Apr. 2012

11. Gary Akehurst "User generated content: the use of blogs for tourism organisations and tourism consumers" 2009 Springer

12. Nicola Perra and Bruno Gonçalves "Modeling and Predicting Human Infectious Diseases" Springer