IJCSN - International Journal of Computer Science and Network, Volume 7, Issue 2, April 2018
ISSN (Online) : 2277-5420
www.IJCSN.org
Impact Factor: 1.5

Data Processing Using Clustering Algorithm

[1] Hindu Sindhura.Y; [2]Anand.R; [3]Dr.Rajashekar M Patil

[1] Computer Science and Engineering, Visvesvaraya Technology University,
BMS Institute Of Technology And Management
Bangalore, Karnataka, India

[2] Asst. Professor Computer Science and Engineering,
BMS Institute Of Technology And Management
Bangalore, Karnataka , India

[3] Professor Information Science Engineering,
HKBK College Of Engineering
Bangalore, Karnataka, India

Abstract - Increasing importance of 'networking' in practical management guides and also by the proliferation of 'social networking' Social networks produce an enormous quantity of data. Facebook consists of over 400 million active users sharing over 5 billion pieces of information each month. People are interconnected through online social networks such as Twitter, Facebook, LinkedIn Instagram etc. Social network analysis has gained prominence due to its use in different applications and analyzing how the members of network interact, share information or establish relationships, useful knowledge about them and their relations can be extracted. In this paper we present an approach to analyze the Twitter and Facebook profiles based on the location. The locations of these users should selected by the user. The proposed analysis is comparing the Facebook and Twitter profiles based on the location and extracting the tweets or comments posted on the network based on the users interested area. Thus the work is related to the big data in the area of big data analytics.

***Keywords* - Big data Analytics, Twitter and Facebook mining, social networking Analysis**

1. Introduction

Social network is used to define web -based services that allow individuals to generate a public/semi-public profile within a domain such that they can connect with other users within the network.

Facebook is a multi-purposed social networking platform. Facebook users posts their views, opinions and their point of perception on different topic. It may include political issue, religious issue, technology, product, movie review and much more daily gossiping issues flooded in their surroundings. Usage of social networking sites like Facebook, Twitter, Myspace, Google+, LinkedIn has shown a rapid increase over years. Today, users of social networking sites are more Twitter called followers friends too in the very early days .where as in face book called friends friending someone in the face book carries a much deeper connection then following on twitter. You can build a significant relationship on twitter in the same as on face book. Twitter following can be interested based on the users twit the facebook interact with family and friends in real life but in twitter talks about friends interested in similar things face book allows users to select from various privacy settings, from a completely visible profile to one that is not even except by acknowledged friends. Twitter has two secure settings personal messages can only seen by the people that the users follow .individual messages does not have different privacy messages. The Facebook and twitter can be merge. Tweets can be posted on to Facebook automatically using the twitter app. Here we get the retrieve data from the twitter and the Facebook based on location. Based on those area users related information and the tweets or post posted by the users based on the interested domin. By using the keyword extraction is the process of selecting words and phrases from the text document. For extracting the key we use the LDA algorithm we use the topic clustering to get the number of users interested area in the certain location to plot the graph based on the users focused region by using supervised and unsupervised algorithm. After that we get

a notification to your mail. Through that we have link through then graph is generated.

2. Related work

2.1 Data Retrieval:

To retrieve the data from the twitter we use the API4j. First you should retrieve the set of tweets using that API by using the some users in that geographical area based on the user focused region. To retrieve the data from the Facebook we API4J (JAVA) .Retrieve the data based on the geographical users and the comments or post posted by the users on the interested domin.

2.2 Topic identification:

Two traditional methods for detecting topics are LDA and PLSA. LDA is a generative probabilistic model that can be applied to different tasks, including topic identification. PLSA, similarly, is a statistical technique, which can also be applied to topic modeling. In these approaches, though temporal information is lost, which is paramount in identifying prevalent topics and is an important characteristic of social media data. Facebook and Twitter-LDA model designed to identify topics in tweets and post. Their work, however, only considers the confidential interests of users, and not prevalent topics at a global scale

2.3 Keyword Extraction:

Keyword or informative term extraction, many unsupervised and supervised methods have been proposed. Unsupervised methods for keyword extraction rely solely on implicit information found in individual texts or in a text corpus. Supervised methods, on the other hand, make use of training datasets that have already been classified. Keyword extraction using supervised and hybrid approaches. Two traditional supervised frameworks are KEA and GenEx, which use machine learning algorithms for the effective extraction of keywords. Other innovative approaches for keyword extraction have been proposed in recent years, including the application of neural networks. Extract keywords from the news media sources.

2.4 Co-Occurrence Similarity:

Co-occurrence relationship of frequent word pairs from a single document. Provide statistical information to aid in the identification of the document's keywords. They proposed that if the probability distribution of co-occurrence between a term x and all other terms in a document is biased to a particular subset of frequent terms, then term x is likely to be a keyword.

Co-occurrence similarity measure in which they measure the association of terms using snippets returned by Web searches. They refer to this measure as co-occurrence double checking (CODC). Proposed a method that uses Page counts and text snippets from Web searches to measure the similarity between words or entities

3. Implementation

3.1 Extraction of Twitter Profiles:

In this phase, the user selects the user community that s/he wants to analyze. This task is done by entering a search word based on location. In addition, the user can select the different geographical areas (locations) from which the tweets will be obtained. The different geographical areas are described in a csv file by its name and its coordinates.
The public Twitter API is a method that allows direct access to some Twitter data. Since the twitter profiles are public, we do not need to ask to the users for these data. For this task, we can use the library Twitter4J to handle communication with Twitter. This library can easily be integrated in our application with the Twitter service. By using this library, we can search tweets according to the search word and the geographical areas entered by the user. Then, the profiles of the users who posted these tweets (containing the search word) are collected.

3.2 Extraction of Facebook Profile:

The data is extracted from the Facebook using third party component Facebook Java APIs. Around 2000 posts are collected from different users which is considered as training set. Live news feeds are used as test data to classify using the different classifiers.

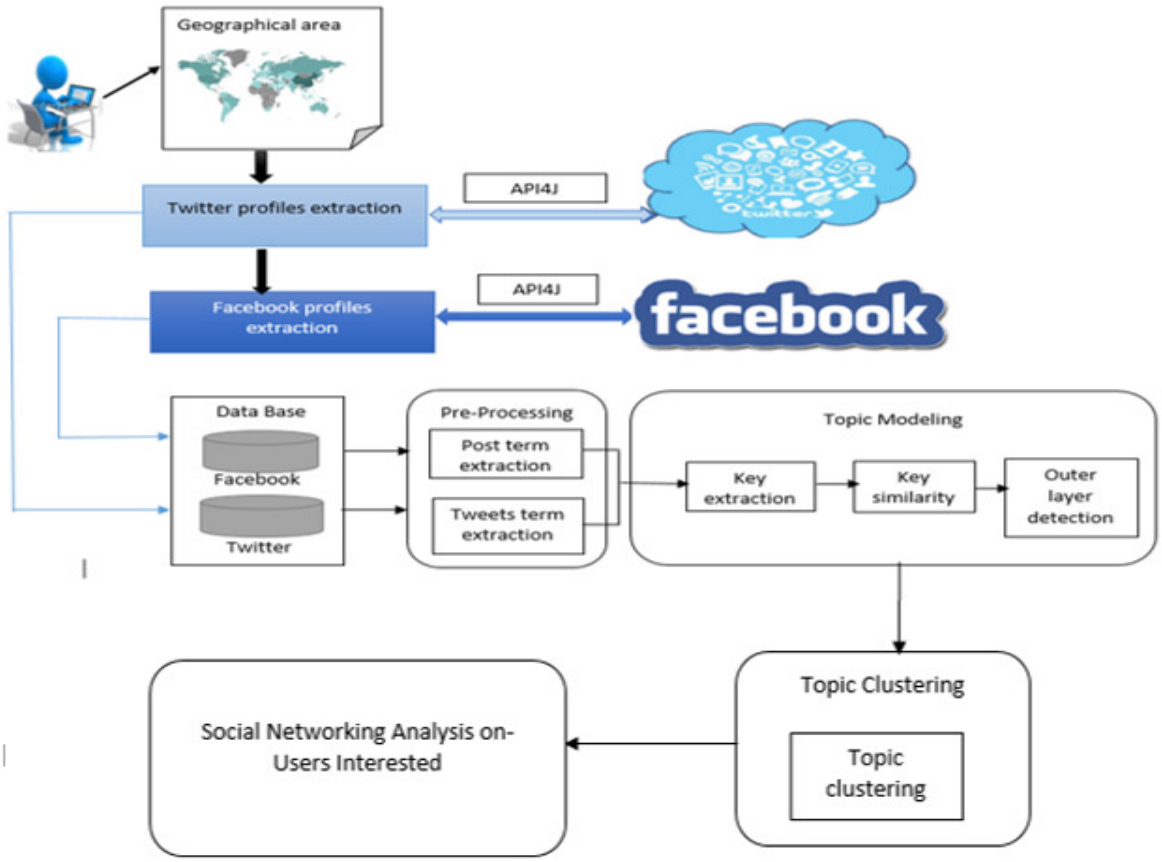

Figure 3.1: General structure of our approach

3.3 Preprocessing:

Key terms are extracted and filtered from news and social data corresponding to a particular period of time based on below methodology.

- Stop words
- Data filters

3.4 Topic Modeling:

The data extracted from the Twitter and Facebook must be analyzed on the content available and modelled. Topic Modelling is a methodology that is frequently used text-mining tool for discovery of hidden semantic structures in a text body. It is, the discovery of "topics" in text corpora by clustering together frequently co-occurring words. This is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents. We have variety of approaches for the topic modelling. Such as

a) Latent Dirichlet allocation

b) Probabilistic latent semantic analysis.

Even though many potentially unimportant terms have been excluded thus far, there are still too many terms (vertices) and co-occurrences (edges) in the graph. We wish to capture only the most significant term co-occurrences, that is, those with sufficiently high QS (Quotient of Similarity) values. To identify significant edges in the graph, irregular co-occurrence values (outliers) must be differentiated from regular ones

3.4.1 Keyword Extraction:

Collecting tweets and post is the first process for this tweets and post are extracted using keywords. For example, in our experiments we used "movies" as the keyword and then collect all tweets related to that keyword. So the corpus will be a collection of tweets. For collecting tweets from net used the Twitter and Facebook API4J

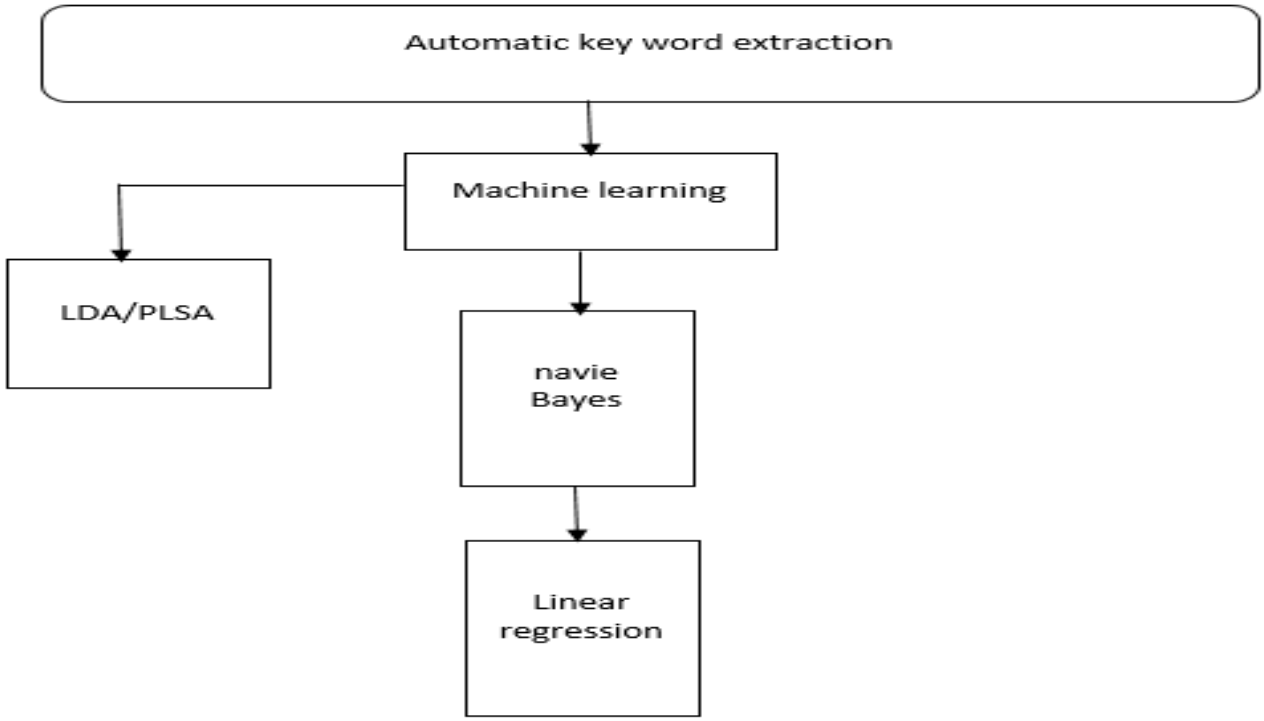

Figure 3.4.1: keyword Extraction

3.4.2 Keyword Similarity:

Similarity between keywords and under- stand the usage scope of keywords as entered by different users in their on-line social network profiles, we analyzed Facebook and Twitter profiles. Considered keywords that are available in the English dictionary. For this purpose, we used the entries present in the Interests fields of a Facebook and Twitter profile. We consider the positive statement of the Users list the activities they are passionate about or topics of which they are interested in this field.

3.4.3 Outliers Detection:

Outliers are identified based on the user's interested areas. The user can select the value of the parameter sigma that acts as a threshold in the detection of outliers. The output of this task is the list of outliers and a graphic with the density of the analyzed examples.

3.5 Topic Clustering:

When clustering is used with dynamic content, the traditional algorithms may not be appropriate. It has been argued that to perform effectively on large content, a clustering algorithm should be able to analyze and update the results incrementally as data are added through the streaming process. It works with limited main memory, process each data only once. We propose using a Hybrid approach that uses both supervised and unsupervised approaches together to process the streaming data and extract the relevant keywords and create the model.

3.6 Social Networking Analysis On-Users Interested:

After completing the clustering data based on the users interested areas .The email is sent to the specific mail id through URL that the graph is generated. Once the URL is generated then the graph is generated.

Social Networking Sites	Sports	Travel	Movies	Reading	Cooking	photography
Facebook	1478	734	1267	300	1267	548
Twitter	783	1245	800	240	800	974

Figure 4.45: location wise user interested areas

4. Results

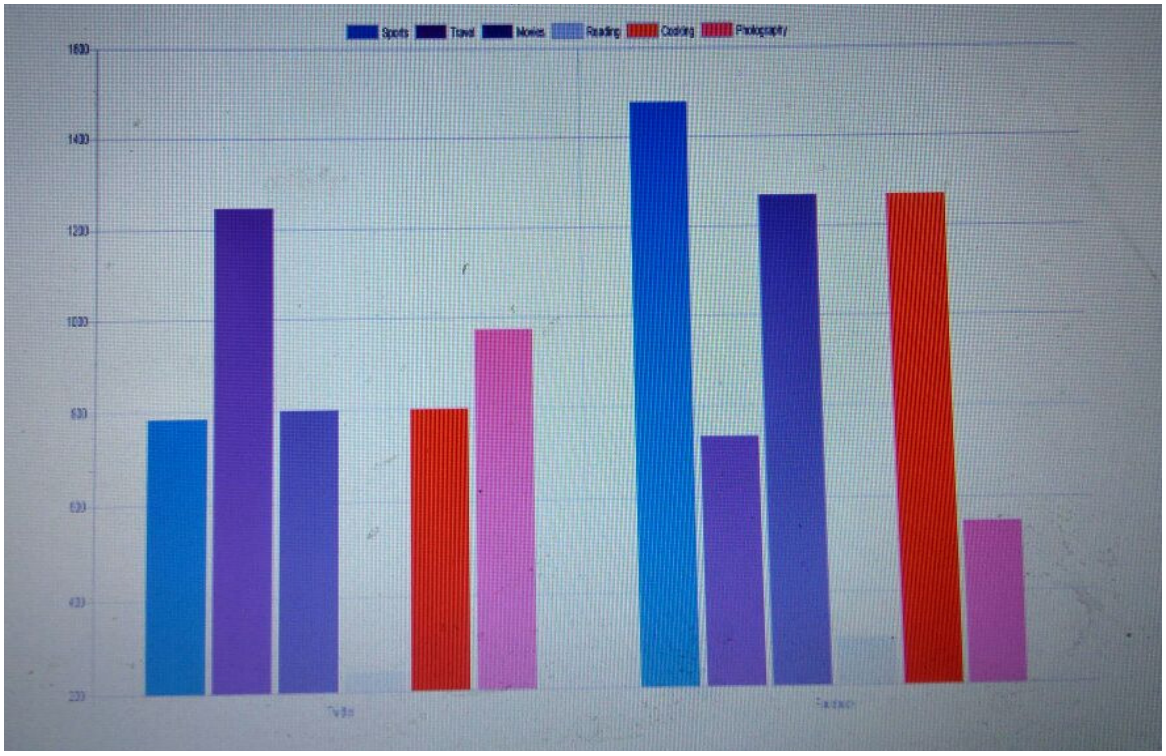

Figure 4: Graph Based On User Interested Area

5. Conclusions

We present an approach to analyze the Twitter and Facebook profiles based on the location. The locations of these users should selected by the user. The proposed analysis is comparing the Facebook and twitter profiles based on the location and extracting the tweets or comments posted on the network based on the users interested area. To retrieve the data from the twitter and Facebook we use the API4j. Topic Modelling is a methodology that is frequently used text-mining tool for discovery of hidden semantic structures in a text body. To

detecting topics we use LDA and PLSA. LDA is a generative probabilistic model that can be applied to different tasks, including topic identification. In these approaches, though temporal information is lost, which is paramount in identifying prevalent topics. Keyword extraction is used by supervised and unsupervised algorithm. Clustering algorithm should be able to analyze and update the results incrementally as data are added through the streaming process. Then the graph is generated based on user's interested areas. Future can be implemented like user interested areas we can also have comparison on various other attributes across social networking comparison on the authenticity on the content usage by users per day this may be implemented by using the graph clustering algorithm. .

References

[1] P. P. Angelov and X. Zhou, "Evolving fuzzy classifier for novelty detection and landmark recognition by mobile robots," in Mobile Robots, 2007, pp. 89–118.

[2] T. Jo, M. Lee, and T. M. Gatton, "Keyword extraction from documents using a neural network model," in *Proc. Int. Conf. Hybrid Inf. Technol. (ICHIT)*, vol. 2. 2006, pp. 194–197.

[3] SociRank: Identifying and Ranking Prevalent News Topics Using Social Media Factors *Davis, Gerardo Figueroa, and Yi-Shin Chen* Derek 2016 IEEE. Personal use is permitted

[4] Priyanka, Anand.R1 Dr.Rajashekhar M.Patil"comparison of betweenness and closeness centralities using incremental algorithms in dynamically growing networks", IJACET, ISSN(PRINT):2394-3408,(ONLINE):2394-3416,VOLUME-3,ISSUE-2,2016.

[5] Anand.R, Pushpalatha .M, Dr Rajshekhar M Patil " A parallel algorithm for reading the different variables in social networks using data mining techniques" Journal of Advanced Computing and Communication Technologies (ISSN: 2347 - 2804) Volume No.4 Issue No.2, April 2016 .

[6] Bharathi M , S. N. Chandra shekara , Anil G.N , Anand R , Muneshwara M.S " K-Anonymity for Real-time Social Network Applications with Network Based anonymization and processing framework" 2012 IACSIT Coimbatore Conferences IPCSIT vol. 28 (2012) © (2012) IACSIT Press, Singapore.

.