Analysing User Posts for Web Forum using K-means Clustering

K.Kanagavalli[*], S.T.Tharani[**]

[*] Dept. of Computer Science and Engineering, Jay Shriram Group of Institution
[**]Dept. of Computer Science and Engineering, Jay Shriram Group of Institution

Abstract- With the increasing sites for forums, online reviews and social networking, the current trend is to look up reviews, expert opinions and discussions on the Web, so that the user can make an informed decision. Sentiment analysis, also known as opinion mining is the computational study of opinions, sentiments and emotions expressed in natural language processing and text analysis. A basic task is classifying the polarity of a given text at the document or sentence, whether the expressed opinion in a document or a sentence is positive, negative, or neutral. It can help better understand the behavioral patterns of users in social media for applications. First, the behavior of individuals is collected through their unstructured posts in a forum. Second, they are classified as positive/negative posts and perform the clustering. Third, the cases are encoded in terms of features in some numerical form, requiring a transformation from text to numbers and assign the positive and negative values to each word to classify the word in the document. Data are collected from forums.digitalpoint.com which includes a range of 75 different topic forums.

Index Terms- Data Mining, Forum Posts, K-means Clustering, Sentiment analysis

I. INTRODUCTION

Data mining is a multidisciplinary field, drawing work from areas including database technology, machine learning, statistics, pattern recognition, information retrieval, neural networks, knowledge-based systems, artificial intelligence, high-performance computing and data visualization.

Users can request and exchange information to others with the help of web forums. Web forums (also called Internet forums) are important services for users. Due to the richness of information in forums, researchers are increasingly interested in mining knowledge from them. The advancement in computing and communication technologies enables people to get together and share information in innovative ways. Social networking sites (a recent phenomenon) empower people of different ages and backgrounds with new forms of collaboration, communication, and collective intelligence.

The study of collective behavior is to understand how individuals behave in a social networking environment. Oceans of data generated by social media like Facebook, Twitter, and YouTube. It presents opportunities and challenges to study collective behavior on a large scale. It aims to learn to predict collective behavior in social media [1].

In k-Means algorithm, only find a local optimum, and is commonly run multiple times with different random initializations. Variations of k-means often include such optimizations as choosing the best of multiple runs, but also restricting the centroid to members of the data set, choosing medians, choosing an initial center less randomly.The algorithm prefers clusters of approximately similar size, as they will always assign an object to the nearest centroid. This often leads to incorrectly cut borders in between of clusters

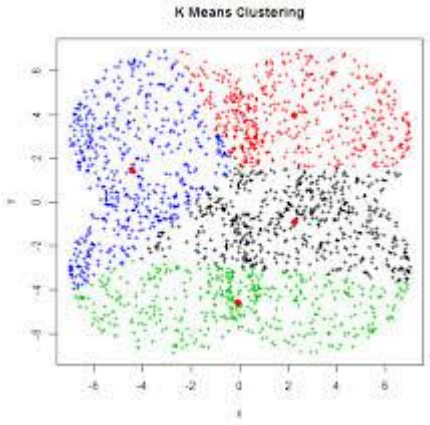

Fig.1. Simple K Means Clustering

K-means has a number of interesting theoretical properties. On the one hand, it partitions the data space into a structure known as a Voronoi diagram. On the other hand, it is conceptually close to nearest neighbor classification, and as such is popular in machine learning. Third, it can be seen as a variation of model based classification [2].

To address the scalability issue, the project proposes an edge-centric clustering scheme to extract sparse social dimensions. With sparse social dimensions, the proposed approach can efficiently handle networks of millions of actors while demonstrating a comparable prediction performance to other non-scalable methods [3]. The assessors developed guidelines for classifying user experiences and used them to interpret data from online forums. Also the researchers analyze the effects of similarity of users on the communities they join, and find two users who communicate more frequently or have common friends are more likely to be in the same set of communities.

In addition, the project includes a new concept called sentiment analysis. Since many automated prediction methods exist for extracting patterns from sample cases, these patterns can be used to classify new cases. The proposed system contains the method to transform these cases into a standard model of features

and classes. Then the behavior of individuals is collected through their posts in a forum and they are classified as positive/negative posts. The cases are encoded in terms of features in some numerical form, requiring a transformation from text to numbers and assign the positive and negative values to each word to classify the word in the document.

II. Forum Study

An Internet forum, or message board, is an online discussion site where people can hold conversations in the form of posted messages. A discussion forum is hierarchical or tree-like in structure: a forum can contain a number of sub forums, each of which may have several topics. Within a forum's topic, each new discussion started is called a thread, and can be replied to by as many people as so wish [4].

A forum consists of a tree like directory structure. The top end is "Categories". A forum can be divided into categories for the relevant discussions. Under the categories are sub-forums and these sub-forums can further have more sub-forums. The topics commonly called threads come under the lowest level of sub-forums and these are the places under which members can start their discussions or posts. Logically forums are organized into a finite set of generic topics and updated by a group known as members, and governed by a group known as moderators. It can also have a graph structure [5].

A. Administrator

The administrators manage the technical details required for running the site. As such, they may promote and depromote members to/from moderators, manage the rules, create sections and sub-sections, as well as perform any database operations. Administrators often also act as moderators. Administrators may also make forum-wide announcements, or change the appearance of a forum. There are also many forums where administrators share their knowledge.

B. Post

A post is a user-submitted message enclosed into a block containing the user's details and the date and time it was submitted. Members are usually allowed to edit or delete their own posts. Posts are contained in threads, where they appear as blocks one after another. The first post starts the thread; this may be called the TS (thread starter) or OP (original post). Posts that follow in the thread are meant to continue discussion about that post, or respond to other replies; it is not uncommon for discussions to be derailed [6].

Most forums keep track of a user's post count. The post count is a measurement of how many posts a certain user has made. Users with higher post counts are often considered more reputable than users with lower post counts.

C. Thread

A thread is a collection of posts, usually displayed from oldest to latest. A thread is defined by a title, an additional description that may summarize the intended discussion and an opening or original post which opens whatever dialogue or makes whatever announcement the poster wished. A thread can contain any number of posts, including multiple posts from the same members, even if they are one after the other.

A thread is contained in a forum, and may have an associated date that is taken as the date of the last post. When a member posts in a thread it will jump to the top since it is the latest updated thread. Similarly, other threads will jump in front of it when they receive posts. A thread's popularity is measured on forums in reply counts. Some forums also track page views. Threads meeting a set number of posts or a set number of views may receive a designation such as "hot thread" and be displayed with a different icon compared to other threads. This icon may stand out more to emphasize the thread. If the forum's users have lost interest in a particular thread, it becomes a dead thread.

D. User groups

Forums organize visitors and logged in members into user groups. Privileges and rights are given based on these groups. A user of the forum can automatically be promoted to a more privileged user group based on criteria set by the administrator. A person viewing a closed thread as a member will see a box saying he does not have the right to submit messages there, but a moderator will likely see the same box granting him access to more than just posting messages. An unregistered user of the site is commonly known as a guest or visitor. Guests are typically granted access to all functions that do not require database alterations or breach privacy.

A guest can usually view the contents of the forum or use such features as read marking, but occasionally an administrator will disallow visitors to read their forum as an incentive to become a registered member. A person who is a very frequent visitor of the forum, a section or even a thread is referred to as a lurker and the habit is referred to as lurking. Registered members often will refer to themselves as lurking in a particular location, which is to say they have no intention of participating in that section but enjoy reading the contributions to it.

E. Moderators

The moderators are users of the forum who are granted access to the posts and threads of all members for the purpose of moderating. Moderators also answer users' concerns about the forum, general questions, as well as respond to specific complaints. Common privileges of moderators include: deleting, merging, moving, and splitting of posts and threads, locking, renaming, sticking of threads, banning, suspending, unsuspending, unbanning, warning the members, or adding, editing, removing the polls of threads. Essentially, it is the duty of the moderator to manage the day-to-day affairs of a forum or board as it applies to the stream of user contributions and interactions. The relative effectiveness of this user management directly impacts the quality of a forum in general, its appeal, and its usefulness as a community of interrelated users.

III. System Design

A. Forum crawling

Download the content from website “forums.digitalpoint.com”. In this step, the information are extracted and downloaded from specified website. URL type recognition consists of two major parts: the entry part and the

online crawling. Forum crawler first finds its entry URL using the Entry URL module. Then, it uses the List/Thread URL module to detect list URLs and thread URLs on the entry page; the detected list URLs and thread URLs are saved to the URL training sets. Next, the destination pages of the detected list URLs are fed into this module again to detect more list and thread URLs until no more list URL is detected. After that, the Turning page URL module tries to find turning page URLs from both list pages and thread pages. Crawler performs online crawling as follows: starting from the entry URL, and follows all URLs matched with any learned URLs.Crawler continues to crawl until no page could be retrieved or other condition is satisfied.

B. Preprocessing of the forum data

In Preprocess, the downloaded forum pages web content are preprocessed and assign the attributes like forumid, forum subid, forum topic, forumurl. Here the preprocessing is used to process the download content in to dissertation application for further classification process.

The content are parse in to different concept such as forum topics, forum subtopics, forum post as HTML file.After download the content, the html files are converted into text file for the purpose of obtaining different forum topic and subtopics which are reside in the html file.This step is used to convert text file data in to dissertation dataset.When the parsing process is accomplished, data cleaning process is applied to the downloaded post sets. In this phase, automatically remove noise data and irrelevant data. Bag of words like stem word, stop word and synonym words are used to remove the noise and irrelevant data.

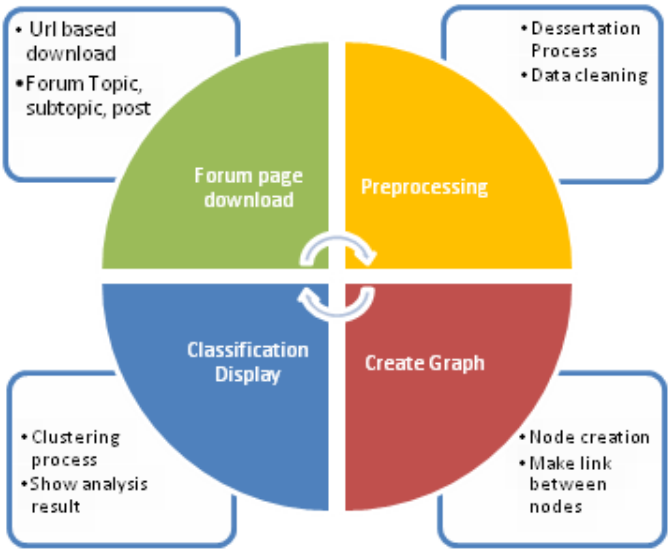

Fig 2. System Design Architecture

C. Create Graph

In Create Graph, forum topics are created as a node. The name of the node is coined automatically. The name should be unique. Links between the nodes represent the relationship between the topics. The link can be created by selecting starting and ending node; a node is linked with a direction. The link name given cannot be repeated. The constructed graph is stored in database. Previous constructed graph can be retrieved when ever from the database.

D. Classification Display

The pre processed forum data is clustered using four different dimensions. In K-Means Algorithm, the data instances are given as input along with number of clusters, and clusters are retrieved as output. First it is required to construct a mapping from features to instances. Then cluster centroids are initialized. Then maximum similarity is given and looping is worked out. When the change is objective value falls above the 'Epsilon' value then the loop is terminated.

Algorithm of scalable k-means variant

Input: data instances $\{xi|1 \le i \le m\}$ number of clusters k
Output: $\{idxi\}$
1. construct a mapping from features to instances
2. initialize the centroid of cluster $\{Cj\ |1 \le j \le k\}$
3. **repeat**
4. Reset $\{MaxSimi\}$, $\{idxi\}$ 5. **for** j=1:k
6. identify relevant instances Sj to centroid Cj
7. **for** i in Sj
8. compute $sim(i,Cj)$ of instance i and Cj
9. **if** $sim(i,Cj) > MaxSimi$
10. $MaxSimi = sim(i,Cj)$
11. $idxi = j$;
12. **for** i=1:m
13. update centroid $Cidxi$
14. **until** change of objective value $< \epsilon$

The K-Means clustering is accomplished by using all topics and sub topics of the forum. The four dimensions of clustering are number of posts/topics, average sentiment values/topics, positive percentage of posts/topics and negative percentage of posts/topics. The posts/topics dimension are determined by number of replies for a post, the sentiment values of this topics are identified from user replies, it describe the user opinion, the positive and negative dimensions are determined from user replies, describe the user perception in the posts. After parsing positive and negative replies are determined. The positive and negative dimensions are also used to identifying the user attitude and pros and cons of the specific topics are discussed in the particular forum.

IV. Conclusion

Algorithms are implemented to automatically analyze the emotional polarity of a text, based on which a value for each piece of text is obtained. The absolute value of the text represents the influential power and the sign of the text denotes its emotional polarity.

The relationships among the topics are determined using scalable learning and represented as a graph. This is used to send the same thread in related topics and get the replies in various dimensions based on the topics

Clustering algorithm is applied to group the forums into various clusters. The obtained final clusters grouped based on the

topics with similar sentiment values and user opinions. Based on the sentiment values, the positive and negative posts are clustered for each thread. Information seekers, decision makers can benefit from this clustering. It simplifies the decision making process.

REFERENCES

[1] Lei Tang , Huan Liu , "Scalable learning of collective behavior based on sparse social dimensions", Proceeding CIKM '09 Proceedings of the 18th ACM conference

[2] Anoop Kumar Jain; Satyam Maheswari," Survey of Re cent Clustering Techniques in Data Mining", IAAST. Jun2012, Vol. 3 Issue 2, p68-74. 7p.

[3] [3] Umesh B.Shingote, "Study of Effective Behavior Prediction by Scalable Learning Method", International Journal, 2014

[4] M.Saravanakumar, T.Suganthalakshmi, "Social Media Marketing", Life Science Journal 2012,9(4).

[5] Oladosu O. A. et al.: "Social Interactions and Query Analy sis in an Online Forum" American Journal of Networks and Communications 2014; 3(1)

[6] Alexandra Raicu, "Knowledge step in advanced mate rials: Polymeric Wikia ", Proc. SPIE 8411.

[7] Jiawei Han and Micheline Kamber, "Data Mining Con cepts and Tech niques", Elseveir.

[8] Hu, Mingqin and Bing Liu. 2004. "Mining and Summarizing Customer Reviews". In Proceedings of KDD 2004.

[9] Wilson, Theresa, Janyce Wiebe and Paul Hoffmann. 2005. "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis". In Proceedings of HLT/EMNLP 2005.

[10] German cobo, David Garcla – solorzano, jose Antonio moran, Eugenia santamaria, carlos monzo, Javier melenchon "Using agglomerative hierarchical clustering to model learner participation profiles in online discussion forums".LAK'12, 29 April – 2May 2012.

[11] Chan, J, Hayes, C. and Daly, E. 2010. "Decomposing discussion forums and boards using user roles". In Proceedings of the WebSci10: Extending the Frontiers of Society On-Line (Raleigh, NC, USA, April 23 - 24, 2010).

AUTHORS

First Author – K.Kanagavalli received her B.Sc Computer Science from Avi-nashilingam University, Coimbatore, M.Sc Computer Science from Bharathiyar University, currently pursuing M.E Conputer Science and Engineering in Jay Shriram Group of Institution, Tirupur, Email-kkanagavallicse@gmail.com

Second Author – S.T.Tharani received her B.E Computer science and Engineer-ing from Avinasilingam University, Coimbatore, M.E degree in Computer Science and Engineering from Anna University. Now working as Assistant Professor in Jay Shriram Group of Institu-tions, Email-tharani.cse85@gmail.com