International Journal of Innovative Research in Engineering Science and Technology

APRIL 2018 ISSN 2320 –981X

SCALABLE LEARNING FOR IDENTIFYING AND RANKING PREVALENT NEWS TOPICS USING SOCIAL MEDIA FACTORS

[1]Hariprabha S , [2]Vidhya R

[1]Department of Computer Science and Engineering,

Nandha College of technology,

Erode – 638052, India.

harisadha54@gmail.com

[2]Department of Computer Science and Engineering,

Nandha College of technology,

Erode – 638052, India.

Vidhya.raj@nandhatech.org

Abstract

A valuable information from online sources has become a famous research area in latest technology. In recent period, social media services provide a vast amount of user-generated data, which have great potential to contain informative news-related content. For these resources to be useful, must find a way to filter noise and only capture the content that, based on its similarity to the news media is considered valuable. 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. As a result, the behavior of individuals is collected through their posts in a forum and then 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.

INTRODUCTION

Data mining is a computer-facilitated process of digging through and analyzing the large

sets of data and then filtering the meaning of the data. Data mining tools predict the behavior. The future trends and allowing businesses to make proactive, knowledge-driven decisions. Text mining is concerned with the task of extracting relevant information from natural language text and to search for interesting relationships between the extracted entities. Text classification is one of the basic techniques in the area of text mining. It is one of the more difficult data-mining problems, since it deals with very high-dimensional data sets with arbitrary patterns of missing data. Text mining is a extraction of data mining to textual data and concerned with various tasks, such as extraction of information from the collection of documents. In existing system text collection is structure of traditional database. Traditional information retrieval techniques become inadequate for the increasingly vast amount of text data. Text expresses a vast range of information but encodes the information in a form is difficult to automatically.

EXISTING SYSTEM

The SociRank which identifies the news topics are prevalent in both social media and the news media, and it will be ranked by relevance of three factors there are Media Focus (MF), User Attention (UA), and User Interaction (UI). It is integrating the techniques, such as keyword extraction, measures of similarity, graph clustering and social network analysis.

SociRank uses keywords from news media sources for a specified period of time to identify the overlap with social media from that same period. Then built a graph whose nodes represent these keywords and whose edges depict their co-occurrences in social media. The graph is then clustered to clearly identify distinct topics. After obtaining well-separated topic clusters (TCs), the factors that

indicate their importance are calculated: MF, UA, and UI. Finally, the topics are ranked by an overall measure that combines these three factors.

2.1 Drawbacks

- Search engine click-through rates is not considered or implemented to provide even more insight into the true interest of users
- The clustering approach is not employed in order to obtain overlapping topic clusters
- The topics are not presented differently to each individual user popularity or prevalence

PROPOSED SYSTEM

The proposed system includes all the existing system aspects. The latent social dimensions are extracted based on network topology to capture the potential affiliations of actors. These extracted social dimensions represent how each actor is involved in diverse affiliations. The entries in this table denote the degree of one user involving in an affiliation. These social dimensions can be treated as features of actors for subsequent discriminative learning.

The project includes online forums hotspot detection and forecast using sentiment analysis and text mining approaches. This is developed in two stages: emotional polarity computation and integrated sentiment analysis based on K-means clustering.

The text-mining approach is used to group the forums into various clusters, and a hotspot forum within the current time span. Educationforum.ipbhost.com which includes

a large amount of different topics. Computation indicates within the same time window and forecasting achieves highly consistent results with K-means clustering.

3.1 Advantages

- It extracts the informative social dimensions for classification.

- Online data is taken for mining.
- Sparsifying social dimensions can be effective in eliminating the scalability bottleneck.
- K-Means clustering is implemented for obtaining the topics as clusters.

SOCAL RANK MODEL

Online social networks play an important role in everyday life for many people. Social media has reshaped the way in which people interact with each other. The rapid development of participatory web and social networking sites like YouTube, Twitter, and Face book also brings about many data mining opportunities and novel challenges.

The social dimensions are extracted based on network topology to capture the potential affiliations of actors. These extracted social dimensions represent how each actor is involved in diverse affiliations. The SocioDim framework demonstrates toward the predicting collective behavior. However, many challenges required for further research. This dynamic nature of networks involves efficient update of the model for collective behavior prediction. It is also fascinating to consider temporal fluctuation into the problem of collective behavior prediction.

- Need to determine a suitable dimensionality automatically which is not present in existing system.
- It is not scalable to handle networks of colossal sizes because the extracted social dimensions are rather dense.

A huge number of actors, extracted dense social dimensions so it can't hold even in memory and it also causing serious problem. To predict collective behavior in social media is being done by understanding how individuals behave in a social networking environment. In particular, given information about some individuals, how to infer the behavior of unobserved individuals in the same network. A social-dimension-based approach has been shown effective in addressing the heterogeneity of connections presented in social media.

SOCIAL RANK ANALYSIS

5.1 Create Graph

In this section, nodes are created flexibly. The name of the node is coined automatically and it should be unique. 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. The graph represents the connections in social media are not homogeneous. People can connect to their family, colleagues, college classmates, or buddies met online. Some relations are helpful in determining a targeted behavior while others are not. This relation-type information, however, is often not readily available in social media.

5.2 Convert to Line Graph

In this section, from the previous module's graph data, line graph is created. The edge details are gathered and constructed as nodes. The nodes with same id in them are connected as edges.In a line graph L(G), each node corresponds to an edge in the original network G, and edges in the line graph represent the adjacency between two edges in the original graph. The set of communities in the line graph corresponds to a disjoint edge partition in the original graph.

5.3 Algorithm of Scalable K-Means Variant

In order to partition edges into disjoint sets, treated that the edges as data instances with their terminal nodes as features. Then a typical clustering algorithm like k-means clustering can be applied

to find disjoint partitions. One concern with this scheme is that the total number of edges might be too huge. Owing to the power law distribution of node degrees presented in social networks, the total number of edges is normally linear, rather than square, with respect to the number of nodes in the network.

5.4 Algorithm for Learning of Collective Behavior

In this section, the network data, labels of some nodes and number of social dimensions are submitted to the system as input; output is label of unlabeled nodes. The following steps are worked out.

- Convert network into edge-centric view.
- Edge clustering is performed.

PERFORMANCES ANALYSIS

The following **Fig 6.1** describes the experimental result for downloading the positive command details. The figures contains forum id and corresponding average number of positive details are shown

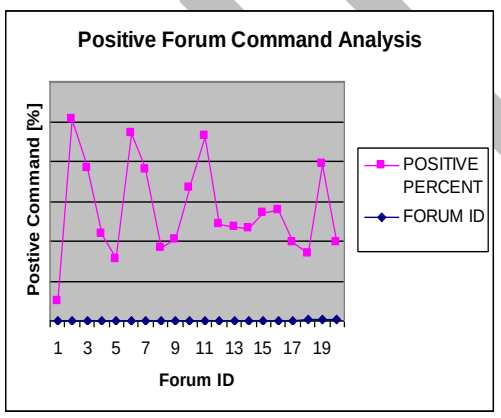

Fig 6.1 Positive Forum Command Analysis

The proposed methodology efficiently analyzes their sentiments. An incomparable advantage of the proposed model is that it easily scales to handle networks with millions of posts. Since the proposed model is sensitive to the number of social dimensions as shown in the experiment, further research is needed to determine a suitable dimensionality automatically.

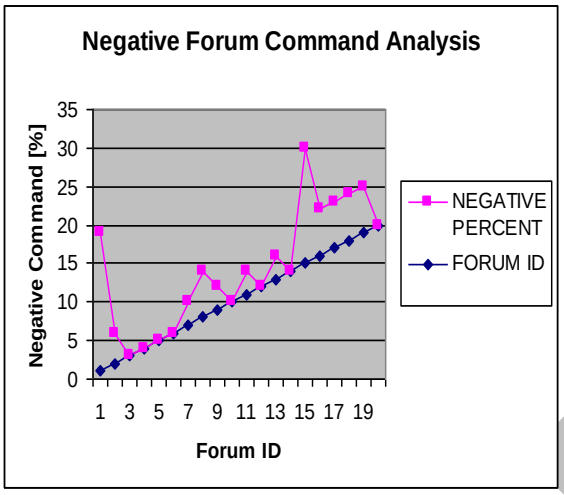

Table 6.2 Negative Forum Command Analysis (Count)

This approach includes the group of forums into various clusters using emotional polarity computation and integrated sentiment analysis based on K-means clustering. Also positive and negative replies are clustered. Using scalable learning the relationship among the topics are identified and represent it as a graph. Data are collected from forums.digitalpoint.com which includes a range of 75 different topic forums. Computation indicates that within the same time window, forecasting achieves highly consistent results with K-means clustering.

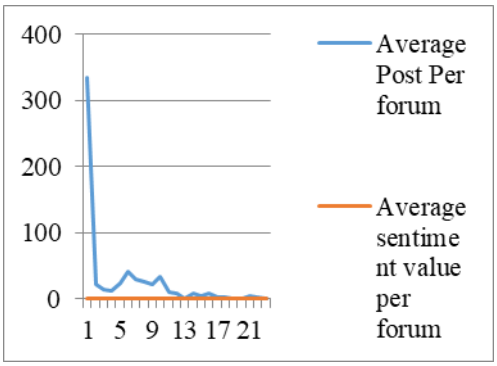

Fig 6.3 Analyzing Average Post Per Forum And Average Sentimental Value

CONCLUSION

SVM is applied to developed 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. This K-means clustering is applied to develop integrated approach for online sports forums cluster analysis. Clustering algorithm is applied to group the forums into various clusters, with the center of each cluster representing a hotspot forum within the current time span. In addition to clustering the forums based on data from the current time window, it is also conducted forecast for the next time window. Empirical studies present strong proof of the existence of correlations between post text sentiment and hotspot distribution. Education Institutions, as information seekers can benefit from the hotspot predicting approaches in several ways. They should follow the same rules as the academic objectives, and be measurable, quantifiable, and time specific. However, in practice parents and students behavior are always hard to be explored and captured.

REFERENCES

[1] M. Granovetter. Threshold models of collective behavior. American journal of sociology , 83(6):1420, 1978.

[2] M. Girvan and M. E. J. Newman, Community structure in social and biological networks. Proc. Natl . Acad. Sci USA 99, 7821–7826 (2002).

[3] R. Guimer`a and L. A. N. Amaral, Functional cartography of complex metabolic networks. Nature 433, 895–900 (2005).

[4] S. Gupta, R. M. Anderson, and R. M. May, Networks of sexual contacts:Implications for the pattern of spread of HIV. AIDS 3, 807–817 (1989).

[5] H. W. Lauw, J. C. Shafer, R. Agrawal, and A. Ntoulas, "Homophily in the digital world: A LiveJournal case study," IEEE Internet Computing, vol. 14, pp. 15–23, 2010.

List of Figures

Figure1: Positive Forum Command Analysis

Figure2: Negative Forum Command Analysis

Figure3: Analyzing Average Post Per Forum And Average Sentimental Value