Asian Journal of Applied Science and Technology (AJAST)
(Open Access Quarterly International Journal) Volume 2, Issue 3, Pages 94-101, July-September 2018

Intensive Learning for Recognition and Ranking common News Topics using Social Media Factors

S.Hariprabha[1] and Mrs.R.Vidhya[2]

[1]*ME CSE, PG Scholar, Nandha College of Technology, Erode, India.*
[2]*Assistant Professor, CSE, Nandha College of Technology, Erode, India.*

Article Received: 20 April 2018 *Article Accepted: 29 July 2018* *Article Published: 24 August 2018*

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.

1. 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 of 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 an 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.

The social media such as blogs and social networks are fueled by the interest in sentiment analysis. With the proliferation of reviews, ratings, recommendations and other forms of online expression, online opinion has turned into a kind of virtual currency for businesses looking to market their products, identify new opportunities and manage their reputations. In business looks to automate the process of filtering out the noise, understanding the conversations, and also identifying the relevant content. Now many are looking to the field of sentiment analysis.

The connectivity information between users to access, but there is no idea why they are connected to each other. This heterogeneity of connections limits the effectiveness of a commonly used technique-collective inference for network classification. A recent framework based on social dimensions and it to be effective in addressing this heterogeneity.

The Proposed framework suggests a novel way of network classification: First, capture the latent affiliations of actors by extracting social dimensions based on network connectivity, and next, apply extant data mining techniques to classification based on the extracted dimensions. In the initial study, modularity maximization was employed to extract social dimensions. The previous framework is not scalable to handle the large size of networks because the extracted social dimensions are rather dense.

2. 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

- It is focus-only ranking by utilizing results obtained from a manual voting method as the ground truth
- 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
- It fails to extract informative social dimensions for classification.
- Not suitable for objects of heterogeneous nature.
- It is not scalable to handle networks of colossal sizes because the extracted social dimensions are rather dense.

3. 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.
- Not only forums are clustered based on sentiment values, but also posts are clustered to find the number of items belongs to the individual clusters.
- It is suitable for the objects of heterogeneous nature.

4. 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.

A social-dimension-based approach has been shown effective in addressing the heterogeneity of connections presented in social media. However, the networks in social media are normally of colossal size, involving hundreds of thousands of actors.

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.

Despite the strong empirical success of discriminative methods is a wide range of applications, when the structures to be learned become more complex than the amount of training data (e.g., in machine translation, scene understanding, biological process discovery), some other source of information must be used to constrain the space of candidate models (e.g., unlabeled examples, related data sources or human prior knowledge). The discriminative

learning procedure will determine the social dimension correlates with the targeted behavior and assign the proper weights.

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

5. PERFORMANCES ANALYSIS

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

Table 5.1 Positive Forum Command Analysis (Count)

S.NO	FORUM ID	POSITIVE PERCENT
1	1	486
2	2	5036
3	3	3832
4	4	2180
5	5	1552
6	6	4696
7	7	3796
8	8	1824
9	9	2012
10	10	3320
11	11	4616
12	12	2410
13	13	2322
14	14	2286
15	15	2676
16	16	2742
17	17	1959
18	18	1662
19	19	3918
20	45	1904

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

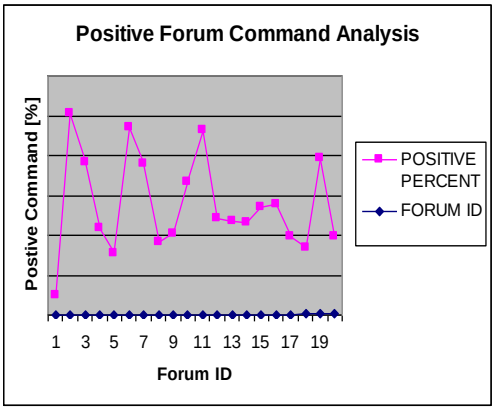

Fig 5.1 Positive Forum Command Analysis

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.

The following **Table 5.2** describes experimental result for proposed system for downloading the negative command analysis details. The table contains forum id and corresponding average number of negative command details are shown.

Table 5.2 Negative Forum Command Analysis (Count)

S.NO	FORUM ID	NEGATIVE PERCENT
1	1	18
2	2	4
3	3	0
4	4	0
5	5	0
6	6	0
7	7	3
8	8	6
9	9	3
10	10	0
11	11	3

12	12	0
13	13	3
14	14	0
15	15	15
16	16	6
17	17	6
18	18	6
19	19	6
20	20	0

The following **Fig 5.2** describes experimental result for proposed system for downloading the negative command analysis details. The table contains forum id and corresponding average number of negative command details are shown.

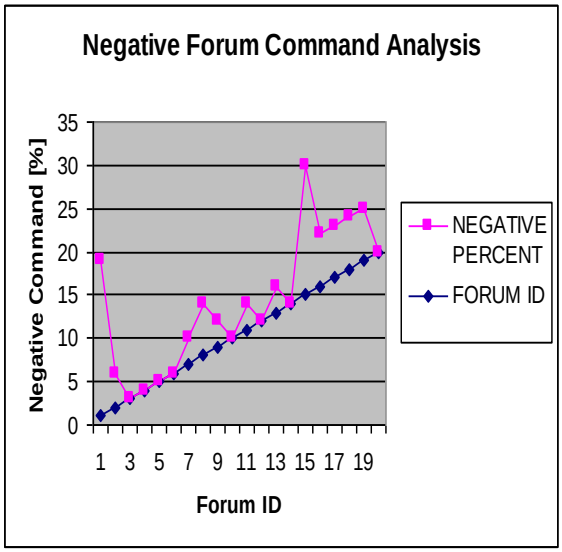

Figure 5.2 Negative Forum Command Analysis (Count)

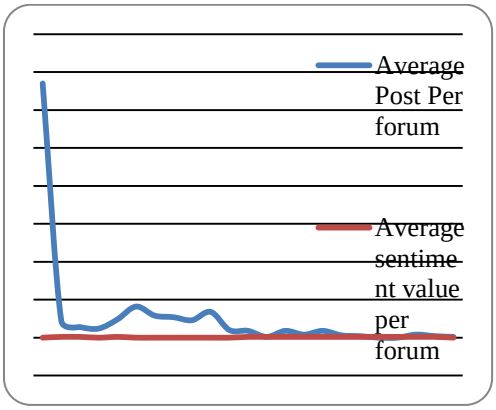

Fig 5.3 Analyzing Average Post Per Forum And Average Sentimental Value

Online ISSN: 2456-883X

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.

6. CONCLUSION AND FUTURE ENHANCEMENT

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.

Using the hotspot predicting approaches can help the education institutions understand what their specific customer's timely concerns regarding goods and services information. Results generated from the approach can be also combined to competitor analysis to yield comprehensive decision support information. The new system become useful if the below enhancements are made in future.

- At present, number of posts/forum, average sentiment values/forums, positive % of posts/forum and negative % of posts/forums are taken as feature spaces for K-Means clustering. In future, neutral replies, multiple-languages based replies can also be taken as dimensions for clustering purpose.
- In addition, currently forums are taken for hot spot detection. Live Text streams such as chatting messages can be tracked and classification can be adopted.

The new system is designed such that those enhancements can be integrated with current modules easily with less integration work and it becomes useful if the above enhancements are made in future.

REFERENCES

[1] L. Tang and H. Liu, “Toward predicting collective behavior via social dimension extraction,” IEEE Intelligent Systems, vol. 25, pp. 19–25, 2010.

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

[3] T. C. Schelling. Dynamic models of segregation. Journal of Mathematical Sociology, 1:143186, 1971.

[4] M. Newman, "Finding community structure in networks using the eigenvectors of matrices," Physical Review E (Statistical, Nonlinear, and Soft Matter Physics), vol.74, no.3, 2006.

[5] M. E. J. Newman, The structure and function of complex networks. SIAM Review 45, 167–256 (2003).

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

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

[8] G. W. Flake, S. R. Lawrence, C. L. Giles, and F. M. Coetzee, Self-organization and identification of Web communities. IEEE Computer 35, 66–71 (2002).

[9] 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).

[10] P. Singla and M. Richardson, "Yes, there is a correlation: - from social networks to personal behavior on the web," in WWW '08: Proceeding of the 17th international conference on World Wide Web. New York, NY, USA: ACM, 2008, pp. 655–664.

[11] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," Annual Review of Sociology, vol. 27, pp. 415–444, 2001.

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