ISSN:2348-2079

Volume-6 Issue-2

International Journal of Intellectual Advancements and Research in Engineering Computations

Scalable Learning for Identifying and Ranking Prevalent News Topics using Social Media Factors

[1]Mrs.K.E.Eswari, M.C.A., M.Phil., M.E., Associate Professor/MCA
[2]Mr. G. Sethupathi, III MCA
Department of MCA, Nandha Engineering College (Autonomous) Erode – 52.
Email ID: eswari.eswaramoorthy@nandhaengg.com, pathisethu143@gmail.com

***Abstract* - In this paper describe a valuable information from online sources has become a prominent research area in information technology in recent years. 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.**

***Keywords*— Big Social Mining, Data Mining opportunities, Support Vector Model, Cross Re Ranking Algorithm, Web Recommendation System**

I. INTRODUCTION

Data mining or knowledge discovery is the computer-assisted process of digging through and analyzing enormous sets of data and then extracting the meaning of the data. Data mining tools predict behaviors and future trends, allowing businesses to make proactive, knowledge-driven decisions. Data mining derives its name from the similarities between searching for valuable information in a large database and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find where the value resides.

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. Growth of volumes of text data, automated extraction of implicit, previously unknown and potentially useful information becomes more necessary to utilize this vast source of knowledge.

Text mining is extraction of data mining approach to textual data and concerned with various tasks, such as extraction of information implicitly contained in 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 rise of social media such as blogs and social networks has fueled 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. As businesses look to automate the process of filtering out the noise, understanding the conversations, identifying the relevant content and action it appropriately, many are now looking to the field of sentiment analysis.

Communal reinforcement is a social phenomenon in which a concept or idea is repeatedly asserted in a community, regardless of whether sufficient empirical evidence has been presented to support it. Over time, the concept or idea is

reinforced to become a strong belief in many people's minds, and may be regarded by the members of the community as fact. Connections in social media networks are not homogeneous. Different connections are associated with distinctive relations. For example, one user might maintain connections simultaneously to his friends, family, college classmates, and colleagues. This relationship information, however, is not always fully available in reality. 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 is shown 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.

II. LITERATURE SURVAY

In this paper "Toward Collective Behavior Prediction via Social Dimension Extraction" [1] the authors Lei Tang and Huan Liu, Arizona State University in the year of 2010 were stated that collective behavior refers to how individuals behave when they are exposed in a social network environment. In the paper, they examined how they could predict online behaviors of users in a network, given the behavior information of some actors in the network.In this paper "Finding community structure in networks using the eigenvectors of matrices" [2] the author M. E. J. Newman considered the problem in the year of 2006 were detecting communities or modules in networks, groups of vertices with a higher-than-average density of edges connecting them.

In this paper "Yes, There is a Correlation - From Social Networks to Personal Behavior on the Web" [3] the authors ParagSingla and Matthew Richardson stated that characterizing the relationship that exists between a person's social group and personal behavior has been a long standing goal of social network analysts. They applied data mining techniques to study this relationship for a population of over 10 million people, by turning to online sources of data.

In this paper "BIRDS OF A FEATHER: Homophily in Social Networks" [4] the authors Miller McPherson, Lynn Smith-Lovin and James M Cook stated that "Similarity breeds connection". This principle the homophily principle-structures network ties of every type, including marriage, friendship, work, advice, support, information transfer, exchange, co-membership, and other types of relationship. The result is that people's personal networks are homogeneous with regard to many socio demographic, behavioral, and intrapersonal characteristics. Homophily limits people's social world in a way that has powerful implications for the information they receive, the attitudes, and the interactions they experience.

In this paper,[5] propose a model to solve service objective evaluation by deep understanding social users. As known, users' tastes and habits are drifting over time. Thus, focus on exploring user ratings confidence, which denotes the trustworthiness of user ratings in service objective evaluation. utilize entropy to calculate user ratings confidence. In contrast, mine the spatial and temporal features of user ratings to constrain confidence. Recently people receive more and more digitized information from Internet. The volume of information is larger than any other point in time, reaching a point of information overload.

In this paper, proposed City Melange, an interactive and multimodal content-based venue explorer[6]. Our framework matches the interacting user to the users of social media platforms exhibiting similar taste. The data collection integrates location-based social networks such as Foursquare with general multimedia sharing platforms such as Flickr or Picasa. In City Melange, the user interacts with a set of images and thus implicitly with the underlying semantics. The semantic information is captured through convolutional deep net features in the visual domain and latent topics extracted using Latent Dirichlet allocation in the text domain

In this paper, [7] investigate the problem of relational user attribute inference by exploiting the rich user-generated multimedia information and exploring attribute relations in social media network sites. Specially, study six types of user attributes: gender, age, relationship, occupation, interest, and emotional orientation. Each type of attribute has multiple values. In this paper, [8] aim to study the semantics of point-of-interest (POI) by exploiting the abundant heterogeneous user generated content (UGC) from different social networks. Our idea is to explore the text descriptions, photos, user check-in patterns, and venue context for location semantic similarity measurement. Recommender systems have become an invaluable asset to online services with the ever-growing number of items and users.

Most systems focused on recommendation accuracy, predicting likable items for each user. Such methods tend to generate popular and safe recommendations, but fail to introduce users to potentially risky, yet novel items that could help in increasing the variety of items consumed by the users. This is known as popularity bias, which is predominant in methods that adopt collaborative filtering. However, recommenders have started to improve their methods to generate lists that encompass diverse items that are both accurate and novel through specific novelty driven algorithms or hybrid recommender systems. In this paper, propose a recommender system that uses the concepts of Experts to find both novel and relevant recommendations.

I. III. SOCIAL RANK MODEL

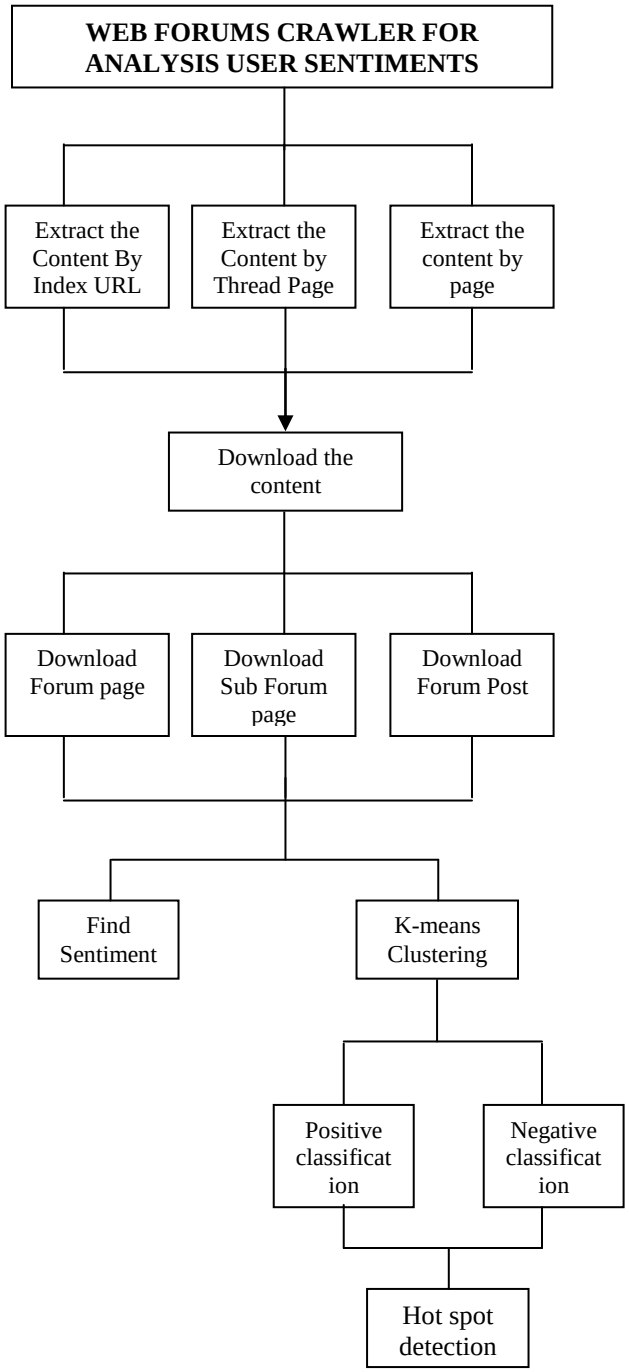
Online social networks play an important role in everyday life for many people. Social media has reshaped the way in

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 scale of these networks entails scalable learning of models for collective behavior prediction.

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 SocioDim framework demonstrates promising results toward predicting collective behavior. However, many challenges require further research. This dynamic nature of networks entails efficient update of the model for collective behavior prediction. It is also intriguing to consider temporal fluctuation into the problem of collective behavior prediction.In discriminative approaches, one directly models the mapping from inputs to outputs (either as a conditional distribution or simply as a prediction function) parameters are estimated by optimizing objectives related to various loss functions. Discriminative approaches have shown better performance given enough data, as they are better tailored to the prediction task and appear more robust to model misspecification.
Despite the strong empirical success of discriminative methods in 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 which social dimension correlates with the targeted behavior and then assign 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.

A huge number of actors, extracted dense social dimensions cannot even be held in memory, causing a serious computational 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. However, the networks in social media are normally of colossal size, involving hundreds of thousands of actors. The scale of these networks entails scalable learning of models for collective behavior prediction. To address the scalability issue, an edge-centric clustering scheme is required to extract sparse social dimensions. With sparse social dimensions, it can efficiently handle networks of millions of actors while demonstrating a comparable prediction performance to other non-scalable methods. In the existing system proposed an unsupervised system SociRank which identifies news topics that are prevalent in both social media and the news media, and then ranks them by relevance using their degrees of 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.

IV. SOCIAL RANK ANALYSIS

A. 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.

B. 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.

C.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.

In this section, 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.

This algorithm also maximizes where k is the number of clusters, S = {S1, S2,...,Sk} is the set of clusters, and μi is the centroid of cluster Si. It keeps only a vector of MaxSim to represent the maximum similarity between one data instance and a centroid. For each iterations, first identify the instances relevant to a centroid, and then compute similarities of these instances with the centroid. This avoids the iteration over each instance and each centroid, which will cost O(mk) otherwise. Note that the centroid contains one feature (node), if and only if any edge of that node is assigned to the cluster.

D.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.
- Construct social dimensions based on edge partition. A node belongs to one community as long as any of its neighboring edges is in that community.
- Apply regularization to social dimensions.
- Construct classifier based on social dimensions of labeled nodes.
- 6. Use the classifier to predict labels of unlabeled ones based on their social dimensions.

V. SENTIMENT ANALYSIS

A. FORUM TOPIC DOWNLOAD

In this section, the source web page is keyed in (default: http://www.fourms.digitalpoint.com) and the content is being downloaded. The HTML content is displayed in a rich text box control. An Internet forum or message board is an online discussion site where people can hold conversations in the form of posted messages. Depending on the access level of a user or the forum set-up, a posted message might need to be approved by a moderator before it becomes visible. Forums have a specific set of jargon associated with them.

B. PARSE FORUM TOPIC TEXT AND URLS

In this section, the downloaded source page web content is parsed and checked for forum links. The links are extracted and displayed in a list box control. Also the link text are extracted and displayed in another list box control. By taking a reading for a chosen time interval and at the end of the reading the acquired data to a text file, this repeats and the text file grows with each new reading interval. Text File: For each interval it places a time and date stamp on line, the second line is the title of that reading which changes and can include text and numbers, the third line reports the recording

and recording length, the remaining lines in each section report the counts recorded in each of levels.

C. FORUM SUB TOPIC DOWNLOAD

In this section, all the forum link pages in the source web page are downloaded. The HTML content is displayed in a rich text box control during each page download. All the topic links are fetched and retrieved from the given link of the forum page. It includes the entire hyperlinks of all the sub topics posted in a forum page. To keep the forum organized, the main topics, the ones on the forum home page are created by the site moderator. This module can able to read sub-topics within them. Before reading a sub-topic, it is read the main topic of a similar one does exist already. Based on the main topics, or a sub-topic, under which contents are read and downloaded through this module.

D. PARSE FORUM SUB TOPIC TEXT AND URLS

In this section, the downloaded forum pages web content are parsed and checked for sub forum links. The links are extracted and displayed in a list box control. Also the link text are extracted and displayed in another list box control. Through this module, the sub topic of the forum text is parsed from the given url of a forum page. The process is by means of reading with regular interval and at the end of the reading the acquired data to a forum page file, this repeats and the text file grows with each new reading interval. At the interval it places a time and date stamp on line, the second line is the title of that sub topic which reads and can include text and numbers, the third line reports the recording and recording length, the remaining lines in each section report the counts recorded in each of levels of the topic is extracted and parsed in this process.

VI. CONCLUSION ANDFUTURE 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 LiveJournal case study," IEEE Internet Computing, vol. 14, pp. 15–23, 2010.