Volume 8, Issue 4
August 2019

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

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

[1]R. Logeswari and [2] Dr. A. Vinayagam

[1] PG and Research Department of Computer Science, M.Phil., Scholar, Govt. Arts College, Karur.

[2]PG and Research Department of Computer Science, Assistant Professor,

Govt. Arts College, Karur

E-mail: [1]lavan130495@gmail.com, [2] a_vinay@rediffmail.com

ABSTRACT

In this paper to achieve prioritization and information ranking, the temporal prevalence of a particular topic in the news media is a factor of importance and can be considered the media focus (MF) of a topic. The temporal prevalence of the topic in social media indicates its user attention (UA). And the interaction between the social media users who mention this topic indicates the strength of the community discussing it and can be regarded as the user interaction (UI) toward the topic. This project studies how networks in social media can help predict some human behaviors and individual preferences. In particular, given the behavior of some individuals in a network, how can infer the behavior of other individuals in the same social network is analyzed. This could help better understand behavioral patterns of users in social media for applications like social advertising and recommendation. 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.

Keywords: *Media focus, User interaction, Edge centric clustering, User Attention.*

1. INTRODUCTION

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.

2.RELATED WORKS

Lei Tang and Huan Liu, Arizona [1] 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.

Many social media tasks can be connected to the problem of collective behavior prediction. Since connections is a social network representing various kinds of relations, a social-learning framework based on social dimensions. This framework suggests extracting social dimensions that represent the latent affiliations associated with actors, and then applying supervised learning to determine which dimensions are informative for behavior prediction.

It demonstrates many advantages, especially suitable for large-scale networks, paving the way for the study of collective behavior in many real-world applications. Social media such as Facebook, MySpace, Twitter, BlogCatalog, Digg, YouTube and Flickr, facilitate people of all walks of life to express their thoughts, voice their opinions, and connect to each other anytime and anywhere. For instance, a popular content-sharing site like Delicious, Flickr, and YouTube allows users to upload, tag and comment different types of contents (e.g., bookmarks, photos, videos).

M. E. J. Newman[2] 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. Previous work indicates that a robust approach to this problem is the maximization of the benefit function known as "modularity" over possible divisions of a network. Here the author showed that this maximization process can be written in terms of the eigenspectrum of a matrix they called the modularity matrix, which plays a role in community detection similar to that played by the graph Laplacian in graph partitioning calculations.

The result leads us to a number of possible algorithms for detecting community structure, as well as several other results, including a spectral measure of bipartite structure in networks and a new centrality measure that identifies those vertices that occupy central positions within the communities to which they belong. The algorithms and measures proposed are illustrated with applications to a variety of real-world complex networks. Networks have attracted considerable recent attention in physics and other fields as a foundation for the mathematical representation of a variety of complex systems, including biological and social systems, the Internet, the World Wide Web, and many others.

Parag Singla and Matthew Richardson[3] 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.

The analysis reveals that people who chat with each other (using instant messaging) are more likely to share interests (their Web searches are the same or topically similar). The more time they spend talking, the stronger their relationship. People who chat with each other are also more likely to share other personal characteristics, such as their age and location and, they are likely to be of opposite gender. Similar findings hold for people who do not necessarily talk to each other but do have a friend in common.

Miller McPherson, Lynn Smith-Lovin and James M Cook[4] 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.

They argued for more research on: (a) The basic ecological processes that link organizations, associations, cultural communities, social movements, and many other social forms. (b) The impact of multiplex ties on the patterns of homophily and (c) The dynamics of network change over time through which networks and other social entities co-evolve. People with different characteristics-genders, races, ethnicities, ages, class backgrounds, educational attainment, etc. appear to have very different qualities.

Guoshuai Zhao, Xueming Qian[5] proposes 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.

This paper, focus on user ratings confidence to discriminate ratings to conduct service objective evaluation. Shown as the left service can learn user ratings confidence from training set. Additionally, explore user ratings confidence with combining spatial-temporal features of ratings to deep understand social users. Proposed approach can learn the confidence value of a rating within specific spatial-temporal context. Specifically, conduct service objective evaluation by deep understanding social users with exploring user ratings confidence.

Jan Zahálka, Stevan Rudinac, and Marcel Worring[6] proposes a City Melange, an interactive and multimodal content-based venue explorer. 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.

This paper, presents City Melange, an interactive multimedia content-based venue explorer. The first step involves collecting a cross-platform multimedia dataset of venues and social media users. In the second step, this dataset is used to construct a number of semantic topics for each venue and social media user by clustering on state-of-the-art visual (ConvNet) and text (LDA) features. These topics are then used in the third step: the interactive city exploration session. City Melange allows the interacting user to iteratively build her user preference profile and get highly personalized recommendations regardless of previous user activity, with each interactive step taking seconds at most.

Quan Fang, Jitao Sang, Changsheng Xu, M. Shamim Hossain[7] investigates 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.

This paper, proposed a Relational LSVM model to address the problem of relational user attribute inference on user-generated multimedia information in social media. The extensive experiments have justified our motivation that exploring the dependency relations between attributes can help achieve better user attribute inference performance. The effectiveness of the whole framework is verified by combining the inferred attribute and mined attribute relation into the structured attribute based user retrieval application.

Xiangyu Wang, Yi-Liang Zhao, Liqiang Nie, Yue Gao[8] aims 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.

In this paper, the authors argued that the venue semantics play an important role in user check-in behavior and modeled it using the heterogeneous user generated content. To the best of our knowledge, this is the first work that targets venue semantics using UGC. Different from the traditional geographical location representation, it represents the semantic information related to the locations.

Kibeom Lee, Kyogu Lee[9] proposes a Dynamically-Promoted Expert (DPE)-based recommender system that was based on collaborative filtering and used the concepts of Experts to provide recommendations to Novices. These recommendations were aimed to be both novel and relevant to the user. The recommender worked by creating clusters of similar items, which would become areas that users could be Experts in. Each user was analyzed to see if they met the requirements to be considered an Expert and if so, on which cluster of items to be an Expert on. Experts were defined as users who are listening behavior were concentrated on certain song clusters, which the requirements of being an Expert in the algorithm tried to formulate.

3. METHODOLOGY

In this paper proposed system used an unsupervised system SociRank which effectively identifies news topics that are prevalent in both social media and the news media, and then ranks them by relevance using their degrees of MF, UA, and UI. Even though this paper focuses on news topics, it can be easily adapted to a wide variety of fields, from science and technology to culture and sports. To the best of our knowledge, no other work attempts to employ the use of either the social media interests of users or their social relationships to aid in the ranking of topics. Moreover, SociRank undergoes an empirical framework, comprising and integrating several techniques, such as keyword extraction, measures of similarity, graph clustering, and social network analysis. The effectiveness of our system is validated by extensive controlled and uncontrolled experiments.

In this proposed model to achieve its goal, SociRank uses keywords from news media sources (for a specified period of time) to identify the overlap with social media from that same period. We then build 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 signify their importance are calculated: MF, UA, and UI. Finally, the topics are ranked by an overall measure that combines these three factors.

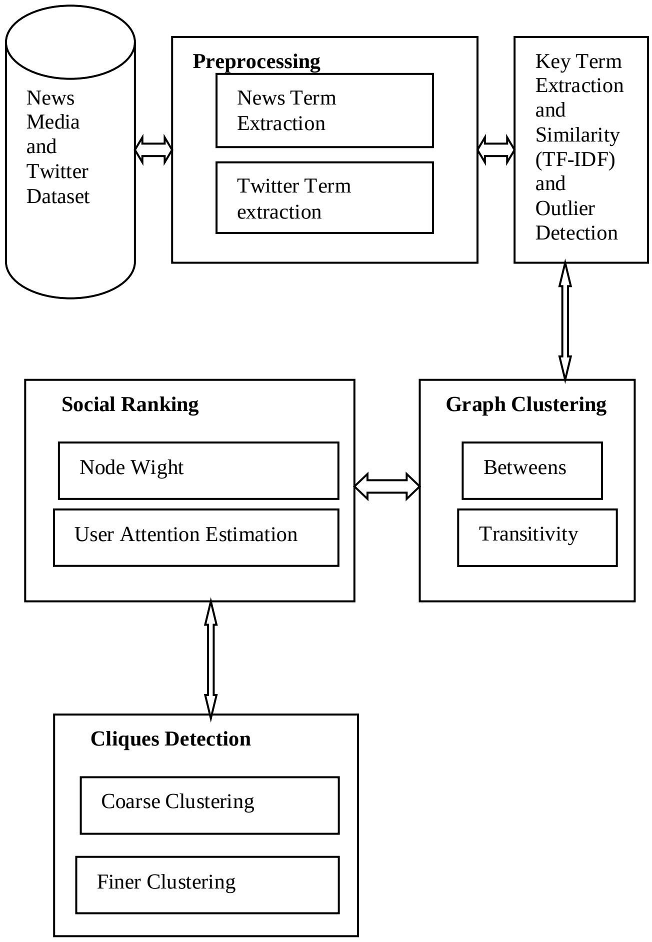

Fig. 3.1 Architecture Diagram

3.1 Preprocessing

In the preprocessing stage, the system first queries all news articles and tweets from the database that fall within date d1 and date d2. Additionally, two sets of terms are created: one for the news articles and one for the tweets, as explained below.

1) News Term Extraction: The set of terms from the news data source consists of keywords extracted from all the queried articles. Due to its simple implementation and effectiveness, we prepare document term matrix to extract the top k keywords from each news article. Then all unique terms are added to set N. It is worth pointing out that, since N is a set, it does not contain duplicate terms.

2) Tweets Term Extraction: For the tweets data source, the set of terms are not the tweets' keywords, but all unique and relevant terms. First, the language of each queried tweet is identified, disregarding any tweet that is not in English. From the remaining tweets, all terms that appear in a stop word list or that are less than three characters in length are eliminated. To eliminate terms that are not relevant, Unicode characters and punctuators are removed. The terms are then added to set T. Then N and T are intersected and set as I.

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Key Term Graph Construction

In this component, a graph G is constructed, whose clustered nodes represent the most prevalent news topics in both news and social media. The vertices in G are unique terms selected from N and T, and the edges are represented by a relationship between these terms. In the following sections, we define a method for selecting the terms and establish a relationship between them. After the terms and relationships are identified, the graph is pruned by filtering out unimportant vertices and edges.

1) Term Document Frequency: First, the document frequency of each term in N and T is calculated accordingly. In the case of term set N, the document frequency of each term n is equal to the number of news articles (from dates d1 to d2) in which n has been selected as a keyword; it is represented as df(n). The document frequency of each term t in set T is calculated in a similar fashion. In this case, however, it is the number of tweets in which t appears; it is represented as df(t). For simplification purposes, we will henceforth refer to the document frequency as "occurrence." Thus, df(n) is the occurrence of term n and df(t) is the occurrence of term t. 2) Relevant Key Term Identification:

Let us recall that set N represents the keywords present in the news and set T represents all relevant terms present in the tweets (from dates d1 to d2). We are primarily interested in the important news-related terms, as this signal the presence of a news related topic. Additionally, part of our objective is to extract the topics that are prevalent in both news and social media. To achieve this, a new set I is formed. This intersection of N and T eliminates terms from T that are not relevant to the news and terms from N that are not mentioned in the social media. Then Itop is set which represents the subset of top key terms from date d1 to date d2, taking into account their prevalence in both news and social media.

3.2 Key Term Similarity Estimation

Next, a relationship is identified between the previously selected key terms in order to add the graph edges. The relationship used is the term co-occurrence in the tweet term set T. The intuition behind the co-occurrence is that terms that co-occur frequently are related to the same topic and may be used to summarize and represent it when grouped.

Several similarity measures were tested in the experiments. They are

a) Dice_QS(i,j) is found out based on above equation where dftop(i) is the number of tweets that contain term i ∈ Itop, dftop(j) is the number of tweets that contain term j ∈ Itop, and co (i, j) is the number of tweets in which terms i and j co-occur in Itop. ϑ is a threshold used to discard quotients of similarity that fall below it. Given the scale of and noise in social media data, it is possible that a pair of terms co-occurs purely by chance. In order to reduce the adverse effects of these co-occurrences, the quotient of similarity QS of two terms is set to zero if their co-occurrence value is less than ϑ. In the experiments, we set ϑ to 5, though this can be adjusted as needed.

$$\text{dice\_QS}(i,j) = \begin{cases} 0 & \text{if } \text{co}(i,j) \le \vartheta \\ \frac{2\times \text{co}(i,j)}{\text{df}_{\text{top}}(i)+\text{df}_{\text{top}}(j)} & \text{otherwise} \end{cases}$$

b) Jacc_QS(i,j) is found out based on the equation.

$$\text{jacc\_QS}(i,j) = \begin{cases} 0 & \text{if } \text{co}(i,j) \le \vartheta \\ \frac{\text{co}(i,j)}{\text{df}_{\text{top}}(i)+\text{df}_{\text{top}}(j)-\text{co}(i,j)} & \text{otherwise.} \end{cases}$$

c) Cosine_QS(i, j) is found out based on the equation.

$$\text{cosine\_QS}(i,j) = \begin{cases} 0 & \text{if } \text{co}(i,j) \le \vartheta \\ \frac{\text{co}(i,j)}{\sqrt{\text{df}_{\text{top}}(i) \times \text{df}_{\text{top}}(j)}} & \text{otherwise.} \end{cases}$$

Note: QS – Quotient of Similarity.

3.3 Graph Clustering

Once graph G has been constructed and its most significant terms (vertices) and term-pair co-occurrence values (edges) have been selected, the next goal is to identify and separate well-defined TCs (sub graphs) in the graph.

a) Betweenness: an efficient approach to achieve the clustering of co-occurrence graphs is finding betweenness. They use a graph clustering algorithm called Newman clustering to efficiently identify word clusters. The core idea behind Newman clustering is the concept of edge betweenness.

The betweenness value of an edge is the number of shortest paths between pairs of nodes that run along it. If a network contains clusters that are loosely connected by a few inter cluster edges, then all shortest paths between the different clusters must go along these edges. Consequently, the edges connecting the clusters will have high edge betweenness. Removing these edges iteratively should thus yield well-defined clusters.

3.4 Content Selection and Ranking

Now that the prevalent news-TCs that fall within dates d1 and d2 have been identified, relevant content from the two media sources that is related to these topics must be selected and finally ranked. Related items from the news media will represent the MF of the topic. Similarly, related items from social media (Twitter) will represent the UA—more specifically, the number of unique Twitter users related to the selected tweets. Selecting the appropriate items (i.e., tweets and news articles) related to the key terms of a topic is not an easy task, as many other items unrelated to the desired topic also contain similar key terms.

3.5 Finding Cliques

In addition, cliques are found out in the graph with given 'n' nodes, the words which are co-related more times are found out. So the main area of the topic can also be identified. If the graph is big, then using the cliques, the words with more density can be found out i.e., more co-related and frequently occurred in the posts.

4. CONCLUSION

UA 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.

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. 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

[3] 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.

[4] 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.