SCALABLE LEARNING FOR IDENTIFYING AND RANKINGPREVALANT NEWS TOPICS USING SOCIAL MEDIA FACTORS

Sadesh.S[1], Kousalya.P[2], Kaviya.C[3], Anton joe. A[4]

Department of Computer Science and Engineering,

Velalar College of Engineering and Technology, Erode - 638012.

ABSTRACT

Social platforms, such as Twitter, reveal much about the tastes of the public. We propose a topic model called twitter hierarchical latent Dirichlet allocation (thLDA). Based on hierarchical latent Dirichlet allocation, thLDA aims to automatically mine the hierarchical dimension of tweets' topics, which can be further employed for text OLAP on the tweets. The experimental results demonstrate that it outperforms other current topic models in mining and constructing the hierarchical dimension of tweeter's topics. We introduce a novel hierarchical model called LDA model to construct a dimension hierarchy of tweets' topics, incorporating social relationships and semantic relationships into the modeling process. We have proposed a system to implement sentimental analysis and how to connect it to Twitter and run sentimental analysis queries for TV Shows. Focus on how to discover the underlying topics of tweets from tweeters' social behaviors and their published tweets. This Project is to improve the mining of the topics and conduct extensive experiments on our model with large quantities of Twitter data and find that the results demonstrate its effectiveness. Also, Social networks (TV Shows) are the main resources to gather information about people's opinions and sentiments towards different topics as they spend hours daily on social media and share their opinion. In a proposed system, to realize that the neutral sentiments are significantly high which TV shows there is a need to improve Twitter sentiment analysis.

KEYWORDS: thLDA, OLAP, Social and Semantic relationships, Sentimentalanalysis.

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 tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

- Enriching existing knowledge bases with new entities and new attribute values becomesmore and more important.
- We focus on enriching geographical location attribute values for entities. Twitter is oneof the most popular micro-blogging platforms around the world.

- On Twitter, users can post and share text messages of up to 280 characters named tweetsabout topics ranging from daily life to new events and any other interests.
- With more than 500 million tweets posted per day, Twitter has become a very importantsource of information.

II. LITERATURE SURVEY

Lei Tang and Huan Liu, Arizona State that collective behaviour refers to how individuals behave when they are exposed in a social network environment. In the paper, they examined how they could predict online behaviours of users in a network, given the behaviourinformation of some actors in the network.

Parag Singla and Matthew Richardson et al stated that characterizing the relationship that exists between a person's social group and personal behaviour 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.

Lynn Smith-Lovin and James M Cook et al 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, behavioural, and intrapersonal characteristics.

III. RESEARCH METHODOLOGY

PROPOSED METHODOLOGY

The project includes twitter hotspot detection and forecast using text mining approaches. This is developed in two stages: emotional polarity computation and integrated sentiment analysis based on hierarchal topic clustering. More related words in the specific topic graph helps for better and finer analysis of contents posted Not only forums are clustered basedon 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.

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. The set of terms from the news data source consists of keywords extracted from all the queried articles. For the tweets data source, the set of terms are not the tweets' keywords, but all unique and relevant terms.

KEY TERM GRAPH CONSTRUCTION AND SIMILARITY ESTIMATION

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. These are done by following the two steps: Term Document Frequency and Relevant Key Term Identification. 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.

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.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**.**

CONTENT SELECTION AND RANKING

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 ofunique 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.

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 theposts.

IV. THEORY AND CALCULATION

Girvan Newman method

It is a hierarchical method used to detect communities in complex system. It detects communities by progressively removing edges from the original network. The connected components of the remaining network are the communities. Instead of trying to construct a measure that tells us which edges are the most central to communities, the Girvan–Newman algorithm focuses on edges that are most likely"between" communities.

Louvain method

It is a greedy optimization method to extract communities from large network. The optimization is performed in two steps. First, the method looks for "small" communities by optimizing modularity locally. Second, it aggregates nodes belonging to the same community and builds a new network whose nodes are the communities. These steps are repeated iteratively until a maximum of modularity is attained and a hierarchy of communities is produced.

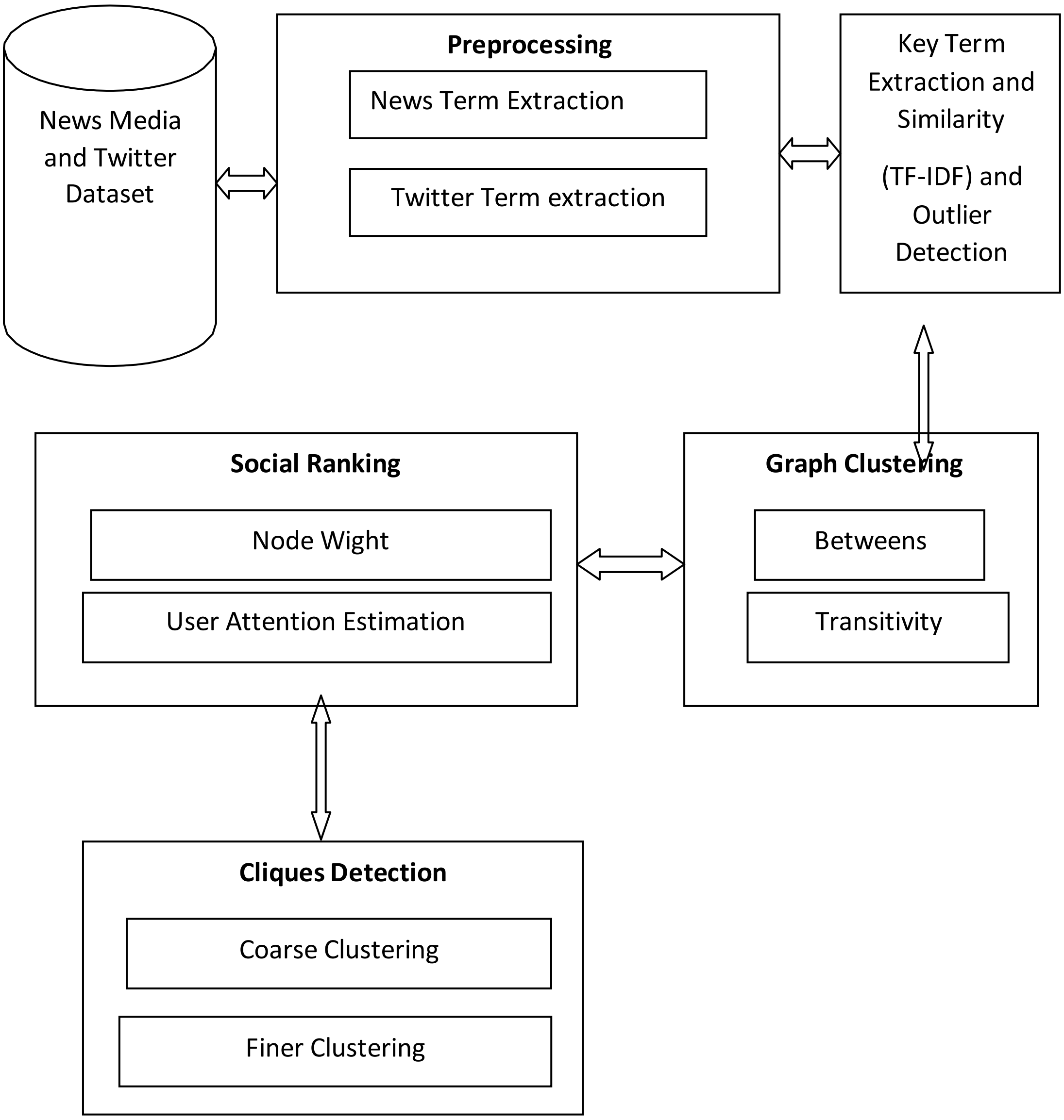
$$\Delta Q=\left[\frac{\Sigma_{in}+k_{i,in}}{2m}-\left(\frac{\Sigma_{tot}+k_i}{2m}\right)^2\right]-\left[\frac{\Sigma_{in}}{2m}-\left(\frac{\Sigma_{tot}}{2m}\right)^2-\left(\frac{k_i}{2m}\right)^2\right]$$

$$\boxed{\Delta Q = e_{iC} - \frac{k_i \Sigma_{tot}}{2m}}$$

Clique (CLuster In QUEst) detection

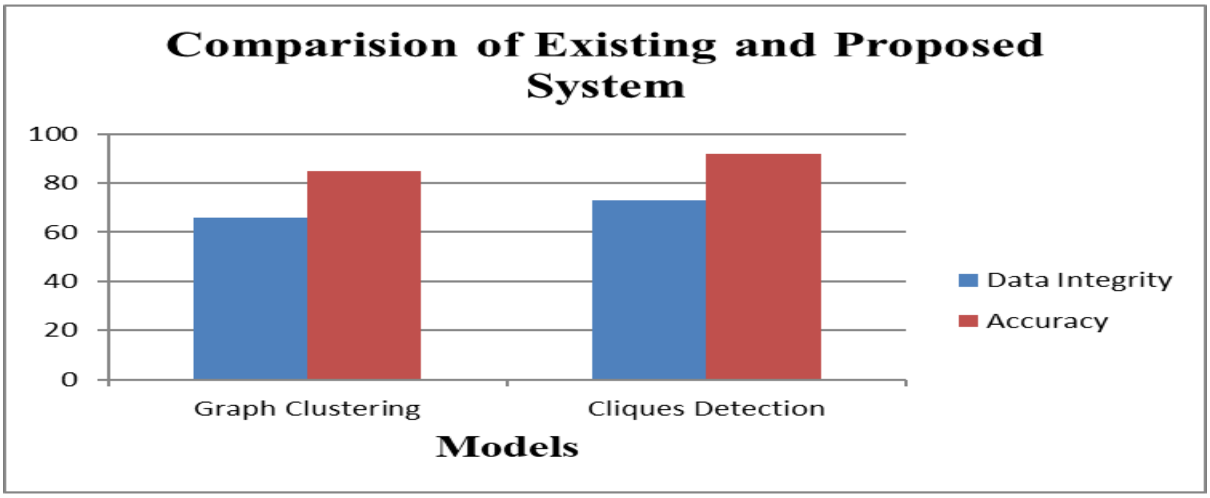
It is a density-based and grid-based clustering algorithm. It is used for the clustering of high dimensional data present in large tables. It identifies the dense units in the subspaces of high dimensional data pace and uses this subspace to provide more efficient clustering.

V. FLOW DIAGRAM



VI. COMPARISON CHART

Models	Graph Clustering	Cliques Detection
Data Integrity	66	73
Accuracy	85	92



ADVANTAGES OF PROPOSED SYSTEM

- It extracts the informative social dimensions for classification.
- Online data is taken for mining.
- More related words in the specific topic graph helps for better and finer analysis of contents posted Not only forums are clustered based on sentiment values, but also posts are clustered to find the number of itemsbelongs to the individual clusters.

VII. RESULTS AND DISCUSSION

The output is designed in such a way that it is attractive, convenient and informative. As the outputs are the most important sources of information to the users, better design should improve the system's relationships with user and also will help in decision-making. Form design elaborates the way output is presented and the layout available for capturing information.

NEWS KEYWORDS LIST

Keywords list with occurrence count are presented as matrix in R Studio console. No. of keywords may be more and so only top 50 words are displayed.

TWITTER KEYWORDS LIST

Keywords list with occurrence count are presented as matrix in R Studio console. No. of keywords may be more and so only top 50 words are displayed

CLIQUES IN GRAPH

The cliques with 6 nodes and their edges are displayed using igraph package in R. Psuedo Clique Finding algorithm is used to find the best nodes in the clique formed.

VIII. CONCLUSION

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

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

[6] R. Guimer`a and L. A. N. Amaral, Functional cartography of complex metabolic

networks.Nature 433, 895–900 (2005).

[7] 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).

[8] 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).

[9] P. Singla and M. Richardson, “Yes, there is a correlation: - from social networks to personal behaviour 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.

[10] 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.

[11] 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.