THROUGH ASPECTS OF SOCIAL MEDIA SOCIAL STATUS RECOGNIZATION AND STATUS PREVALENT NEWS

K. LAKSHMI SRAVANTHI [1], K. SUBBARAO [2]

[1] M.Tech Research Scholar, St. Ann's College of Engineering & Technology, Chirala.
[2] Assoc. Professor & Research Supervisor, St. Ann's College of Engineering & Technology, Chirala
lsravanthi.kanuganti@gmail.com, subbukatte@gmail.com

ABSTRACT

Broad communications sources, specifically the news media, have generally educated us of every day occasions. In present day times, online networking administrations, for example, twitter give a gigantic measure of client produced information, which can possibly contain educational news-related substance. For these assets to be helpful, we should find an approach to filter clamor and just catch the substance that, in light of its comparability to the news media, is viewed as important. In any case, even after clamor is evacuated, information over-burden may at present exist in the rest of the information—thus, it is advantageous to organize it for utilization. To accomplish prioritization, data must be positioned arranged by evaluated significance considering three variables. To begin with, the transient prevalence of a specific point in the news media is a factor of significance, and can be viewed as the media focus of a subject. Second, the worldly pervasiveness of the point in online networking demonstrates its User Attention. Last, the collaboration between the onlinenetworking clients who specify this theme indicates the quality of the group talking about it, and can be viewed as the User Interaction around the subject. We propose an unsupervised structure SociRank—which identifies news subjects predominant in both web-based social networking and the news media, and afterward positions them by pertinence utilizing their degrees of Media Focus, User Attention, and User Interactions. Our trials demonstrate that SociRank enhances the quality and assortment of consequently identified news themes.

I. INTRODUCTION

Verifiable information that notifies the overall population of day by day occasions has been given by broad communications sources, specifically the news media. Huge numbers of these news media sources have either relinquished their printed copy distributions or moved to the World Wide Web, or now create both printed version and Internet forms simultaneously. These news media sources are viewed as dependable since they are distributed by proficient columnists, who are considered responsible for their substance.

1.1 SOCIAL MEDIA

The user generated data exchange through internet, this phenomena is called as social media. The Internet, being a free and open gathering for data trade, has as of late observed a captivating wonder known as online networking. In online networking, consistent, non-writer clients can distribute unverified substance and express their enthusiasm for specific occasions. Microblogs have turned out to be a standout amongst the most famous online networking outlets. One microblogging administration specifically, Twitter, is utilized by a large number of individuals around the globe, star viding tremendous measures of client created information. One may accept that this source possibly contains data with equivalent or more noteworthy incentive than the news media, however one should likewise expect that due to the unverified idea of the source, quite a bit of this substance is futile.

For web-based social networking information to be of any utilization for point identification, we should find an approach to filter uninformative data and catch just data which, in light of its substance comparability to the news media,might be viewed as helpful or important. The news media introduces professionally verified events or occasions, while online networking presents the interests of the group of onlookers in these territories, and may in this way give understanding into their fame. Online networking administrations like Twitter can likewise give extra or supporting data to a specific news media point. In synopsis, genuinelysignificant data might be thought of as the territory in which these two media sources topically converge. Sadly, even after the expulsion of irrelevant substance, there is still data over-burden in the rest of the news-related information, which must be organized for utilization.

II. EXISTING SYSTEM

Two traditional methods for detecting topics are LDA and PLSA. LDA is a generative probabilistic model that can be applied to different tasks, including topic identification. PLSA, similarly, is a statistical technique, which can also be applied to topic modeling. In these approaches, however, temporal information is lost, which is paramount in identifying prevalent topics and is an important characteristic of social media data.

Matsuo *et al*. employed a different approach to achieve the clustering of co-occurrence graphs. They used Newman clustering to efficiently identify word clusters. The core idea behind

Newman clustering is the concept of edge betweenness. The betweenness measure 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 intercluster edges, then all shortest paths between different clusters must go along one of these edges. Consequently, the edges connecting different clusters will have high edge betweenness, and removing them iteratively will yield well-defined clusters.

2.1 DISADVANTAGES OF EXISTING SYSTEM

- Even after the removal of unimportant content, there is still information overload in the remaining news-related data, which must be prioritized for consumption.
- LDA and PLSA only discover topics from text corpora; they do not rank based on popularity or prevalence.
- The main disadvantage of the algorithm was its high computational demand.
- The existing work, however, only considers the personal interests of users, and not prevalent topics at a global scale.

These methods, however, only use data from microblogs and do not attempt to integrate them with real news. Additionally, the detected topics are not ranked by popularity or prevalence.

III. PROPOSED SYSTEM

We propose 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.

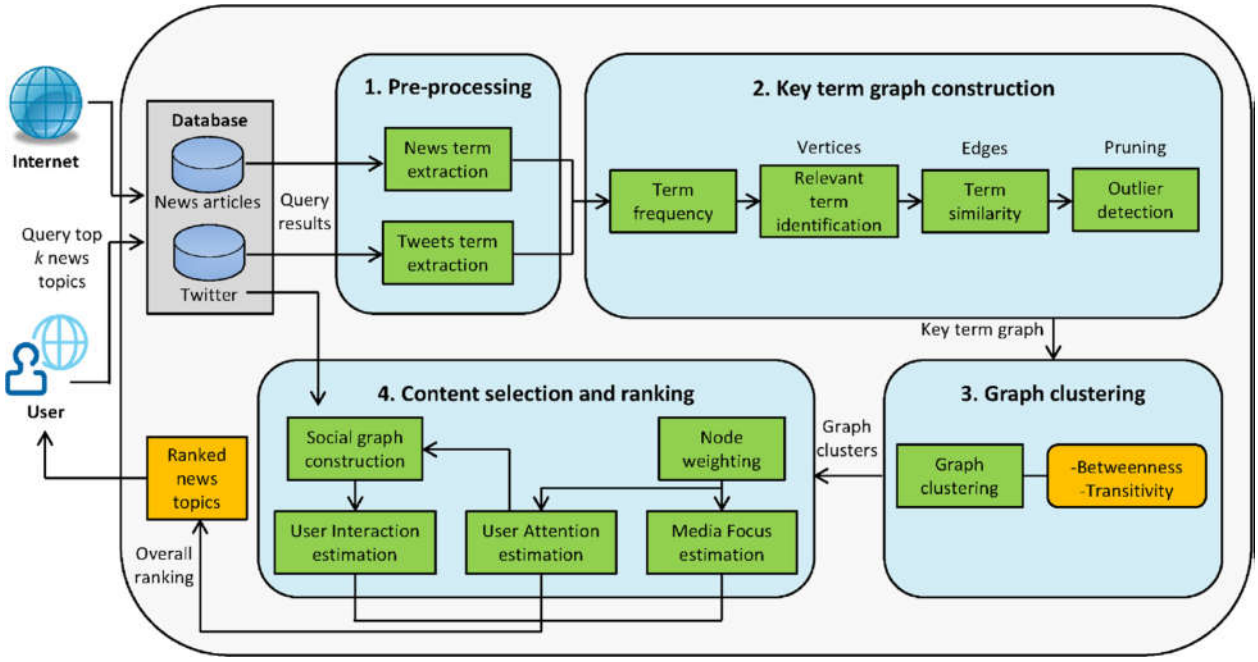
To achieve its goal, SociRank uses keywords from news media sources (for a specified period of time) to identify the overlap with social media from that same period.We then build a graph whose nodes represent these keywords and whose edges depict their co-occurrences in social media. The graph is then clustered to clearly identify distinct topics. After obtaining well-separated topic clusters (TCs), the factors that signify their importance are calculated. Finally, the topics are ranked.

3.1 ADVANTAGES OF PROPOSED SYSTEM

- 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.

IV. SYSTEM ARCHITECTURE



The goal of our method—SociRank—is to identify, consolidate and rank the most prevalent topics discussed in both news media and social media during a specific period of time. The system framework can be visualized in Fig. 1. To achieve its goal, the system must undergo four main stages.

1) *Preprocessing:* Key terms are extracted and filtered from news and social data corresponding to a particular period of time.

2) *Key Term Graph Construction:* A graph is constructed from the previously extracted key term set, whose vertices represent the key terms and edges represent the co-occurrence similarity between them. The graph, after processing and pruning, contains slightly joint clusters of topics popular in both news media and social media.

3) *Graph Clustering:* The graph is clustered in order to obtain well-defined and disjoint TCs.

4) *Content Selection and Ranking:* The TCs from the graph are selected and ranked using the three relevance factors

(MF, UA, and UI).

Initially, news and tweets data are crawled from the Internet and stored in a database. News articles are obtained from specific news websites via their RSS feeds and tweets are crawled from the Twitter public timeline [41]. A user then requests an output of the top k ranked news topics for a specified period of time between date d_1 (start) and date d_2 (end).

A. Preprocessing

In the preprocessing stage, the system first queries all news articles and tweets from the database that fall within date d_1 and date d_2. 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 implement a variant of the popular TextRank algorithm [23] to extract the top k keywords from each news article.2 The selected keywords are then lemmatized using the WordNet lemmatizer in order to consider different inflected forms of a word as a single item. After lemmatization, 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. The part of speech (POS) of each term in the tweets is then identified using a POS tagger [42]. This POS tagger is especially useful because it can identify Twitter-specific POSs, such as hashtags, mentions, and emoticon symbols. Hashtags are of great interest to us because of their potential to hold the topical focus of a tweet. However, hashtags usually contain several words joined together, which must be segmented in order to be useful. To solve this problem, we make use of the Viterbi segmentation algorithm [43]. The segmented terms are then tagged as "hashtag." To eliminate terms that are not relevant, only terms tagged as hashtag, noun, adjective or verb are selected. The terms are then lemmatized and added to set T, which represents all unique terms that appear in tweets from dates d_1 to d_2.

B. 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:

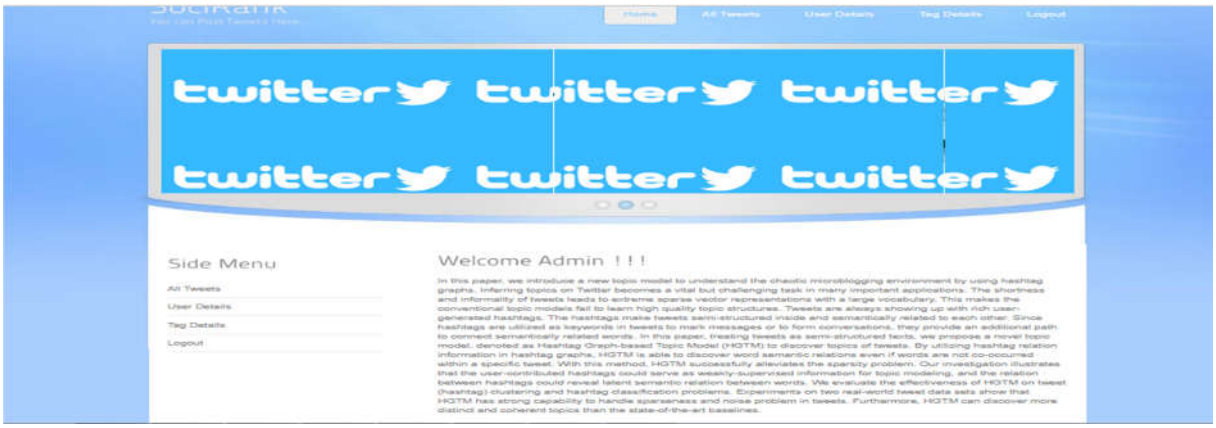
First, the document frequency of each term in *N* and *T* is calculated accordingly. In the case of term set *N*, the document frequency of each term *n* is equal to the number of news articles (from dates d_1 to d_2) in which *n* has been selected as a keyword; it is represented s df *(n)*. The document 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*.

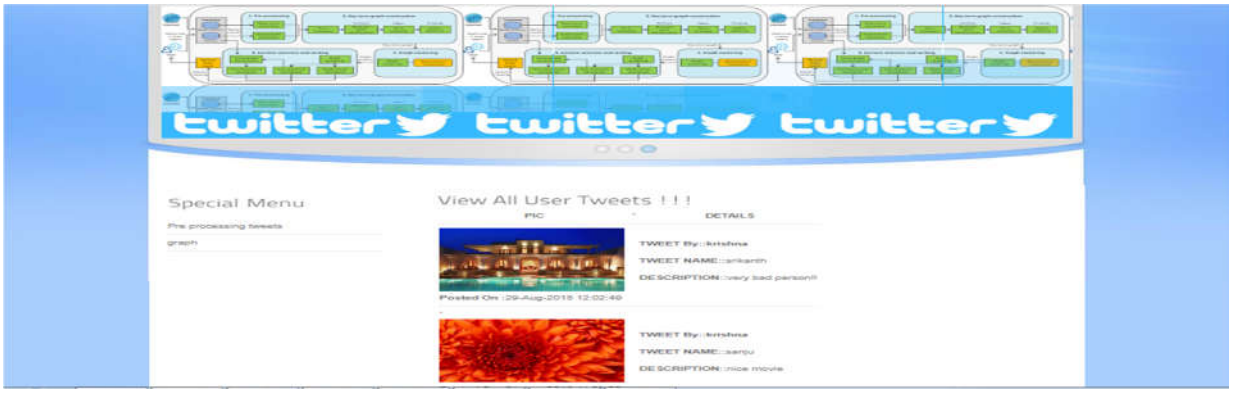
2) Relevant Key Term Identification:

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 d_1 to d_2). We are primarily interested in the important news-related terms, as these signal the presence of a newsrelated 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

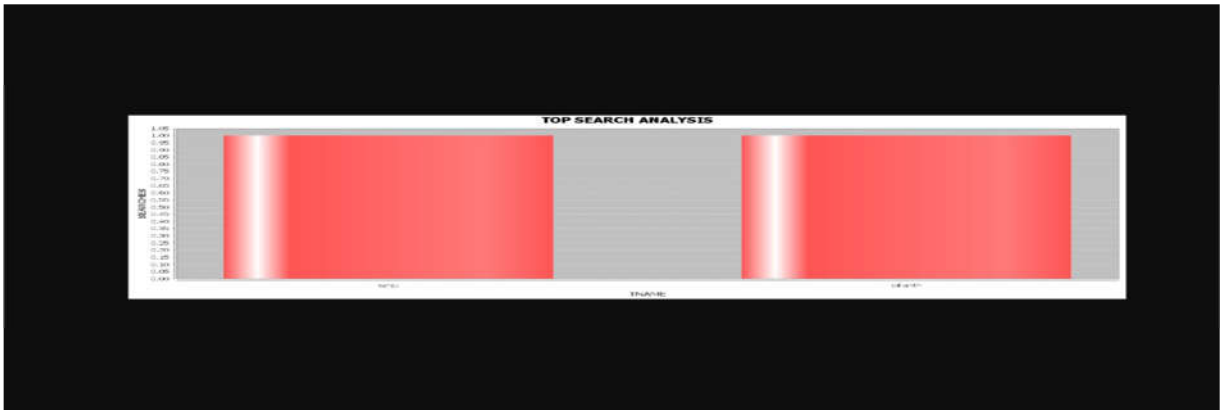
3) Key Term Similarity Estimation: Next, we must identify a relationship between the previously selected key terms in order to add the graph edges. The relationship used is the term co-occurrence in the tweet term set *T*. The intuition behind the co-occurrence is that terms that co-occur frequently are related to the same topic and may be used to summarize and represent it when grouped.

V. SCREEN SHOTOS


Side Menu
All Tweets
User Details
Tag Details
Logout
Welcome Admin ! ! !


Special Menu
Pre processing tweets
graph
View All User Tweets ! ! !
PIC
DETAILS


SociRank
Home
All Tweets
User Details
Tag Details
Logout
DATA MINING
prev
next
Side Menu
All Tweets
User Details
Tag Details
Logout
View ALL Users ! ! !
USER ID
User Image
Name
Email-Id
krishna
U2311
reddy7312@gmail.com
localhost:8091/SociRank/UserDetails.jsp#



VI. CONCLUSION

In this project, An unsupervised method— SociRank—which identifies news topics prevalent in both social media and the news media, and then ranks them by taking into account their Media Focus, User Attention, and User Interaction as relevance factors. The temporal prevalence of a particular topic in the news media is considered the Media Focus of a topic, which gives us insight into its mass media popularity. The temporal prevalence of the topic in social media, specifically Twitter, indicates user interest, and is considered its User Attention. Finally, the interaction between the social media users who mention the topic indicates the strength of the community discussing it, and is considered the User Interaction. To the best of our knowledge, no other work has attempted to employ the use of either the interests of social media users or their social relationships to aid in the ranking of topics. Consolidated, filtered, and ranked news topics from both professional news providers and individuals have several benefits. One of its main uses is increasing the quality and variety of news recommender systems, as well as discovering hidden, popular topics.Our system can aid news providers by providing feedback of topics that have been discontinued by the mass media, but are still being discussed by the general population.

SociRank has been compared to mediafocus-only ranking by utilizing results obtained from a manual voting method as the ground truth. In the voting method, 20 individuals were asked to rank topics from specified time periods based on their perceived importance. The evaluation provides evidence that our method is capable of effectively selecting prevalent news topics and ranking them based on the three previously mentioned measures of importance. Our

results present a clear distinction between ranking topics by Media Focus only and ranking them by including User Attention and User Interaction.

VII. FUTURE ENHANCEMENT

In future work, we intend to perform experiments and expand SociRank on different areas and datasets. Furthermore, we plan to include other forms of UA, such as search engine click-through rates, which can also be integrated into our method to provide even more insight into the true interest of users. Additional experiments will also be performed in different stages of the methodology. For example, a fuzzy clustering approach could be employed in order to obtain overlapping TCs. Lastly, we intend to develop a personalized version of SociRank, where topics are presented differently to each individual user.

VIII. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation,"*J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.

[2] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. 15th Conf. Uncertainty Artif. Intell.*, 1999, pp. 289–296.

[3] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Berkeley, CA, USA, 1999, pp. 50–57.

[4] C. Wartena and R. Brussee, "Topic detection by clustering keywords," in *Proc. 19th Int.Workshop Database Expert Syst. Appl. (DEXA)*, Turin, Italy, 2008, pp. 54–58.

[5] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on Twitter based on temporal and social terms evaluation," in *Proc. 10th Int. Workshop Multimedia Data Min. (MDMKDD)*, Washington, DC, USA, 2010, Art. no. 4. [Online]. Available: http://doi.acm.org/10.1145/1814245.1814249.

[6] W. X. Zhao *et al.*, "Comparing Twitter and traditional media using topicmodels," in *Advances in Information Retrieval*. Heidelberg, Germany: Springer Berlin Heidelberg, 2011, pp. 338–349.

[7] Q. Diao, J. Jiang, F. Zhu, and E.-P. Lim, "Finding bursty topics from microblogs," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguist. Long Papers*, vol. 1. 2012, pp. 536–544.

[8] C. Wang, M. Zhang, L. Ru, and S. Ma, "Automatic online news topic ranking using media focus and user attention based on aging theory," in Proc. 17th Conf. Inf. Knowl. Manag., Napa County, CA, USA, 2008, pp. 1033–1042.

[9] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in Proc. EMNLP, vol. 4. Barcelona, Spain, 2004.

[10] H. Iwasaka and K. Tanaka-Ishii, "Clustering co-occurrence graph based on transitivity," Presented at the 5th Workshop Very Large Corpora, 1997, pp. 91–100.

[11] Y. Matsuo, T. Sakaki, K. Uchiyama, and M. Ishizuka, "Graph-based word clustering using a Web search engine," in Proc. Conf. Empir. Methods Nat. Lang. Process., 2006, pp. 542–550.

[12] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," Proc. Nat. Acad. Sci., vol. 99, no. 12, pp. 7821–7826, 2002.

[13] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," Phys. Rev. E, vol. 69, no. 6, 2004, Art. no. 066133.

[14] Twitter. [Online]. Available: http://www.twitter.com, accessed Feb. 2014.

[15] U. Brandes, "On variants of shortest-path betweenness centrality and their generic computation," Soc. Netw., vol. 30, no. 2, pp. 136–145, 2008.

K. LAKSHMI SRAVANTHI

B.Tech (CSE, 2007-2011) – Chirala Engineering college, which is affiliated under JNTU Kakinada. Currently pursuing M.Tech (computer science and engineering) – St.Ann's College of Engineering and Technology which is affiliated under JNTU Kakinada. My area of interests: Development in Programming languages.

K.SUBBARAO M.Tech(CSE)

He is presently working as Associate Professor in Department of Computer Science and Engineering at St.Ann's College of Engineering and Technology. He has 14 years of Teaching Experience. Research Areas of interests: Digital I image Processing, Recommander system, computer networks.