www.ijecs.in
*International Journal Of Engineering And Computer Science*
*Volume 8 Issue 7 July 2019, Page No. 24683-24789*
*ISSN: 2319-7242 DOI: 10.18535/ijecs/v8i07.4343*

# A new approach of ranking prevalent news topics from social media using unsupervised techniques

**Dr. D. Murali[1], Vinutha B.A[2]**

*1 Professor and Head, Department of Computer science and Engineering, Vemu Institute Of Technology, P. Kothakota, Chittoor, Andrapradesh, India.
2 PG Scholar Department of computer science and Engineering, Vemu Institute Of Technology, P. Kothakota, Chittoor Andrapradesh, India.

**Abstract**
The precious data from online origin has developed into a extended research. The mass media and news media provides the daily events to the common people. Huge amount of information is been achieved by an online social media suchlike Twitter, which contains more information about news-associated content. It is necessary to find a way to filter noise, for these resources to be useful and grab the content that is depend on the similarity to news media. Despite after the noise is eliminated the excessive data still remain in the data so it is essential to prioritize it for utilization. We are introducing three factors for prioritization. The unsupervised technique finds the news topics that are common in the pair of social media and news media, and then ranks them by the applicability factors such as MF, UA and UI. Initially the temporal prevalence of the appropriate topic in news media focus (MF). Secondary the temporal prevalence of the appropriate topic in social media illustrates the user attention (UA). Finally the interconnection among the social media users who specify this topic demonstrates the power of the society who is discussing; it is termed as the user interaction (UI).

***Index Terms*—Information filtering, social computing, social network analysis, topic identification, topic ranking.**

## I. Introduction

In recent years the extraction of useful data through online sources has grown into a remarkable research field in information technology. The daily fact is provided by the mass media source, especially the news media. The news media sources have eliminated paper copy publishing and replaced to the World Wide Web, and produce both paper copy as well as Internet variant. The news media sources are presented by the experienced journalists and hence they are considered trustworthy. Social media is the attractive aspect for information transaction. The most famous social media outlet is the micro blogs. Twitter is considered as the micro blogging service which is used by millions of people across the world, and provides excessive volume of user generated data. The information obtained from social media are unverified and hence much of the content will be useless. These information to be useful and valuable the information gathered should be filtered and collects only information related to the news media.

The professionally verified events are given by news media whereas the social media provides the unverified content, presents interest of the public in these fields. Social media services such as Twitter serve as an additional data to a specific news media event. Even after eliminating the unwanted content, the information overload still prevail in the news associated data and it requires prioritization for utilization. The news information must be ranked in order of predicted importance to achieve prioritization.

The extensively covered topics from news media sources is referred as MF of the topic .The

information gathered from the social media such as Twitter specify the users interest in the topic and it is regarded as the UA of the topic. The number of user discussing the topic and the inter communication among them indicates as UI. It is possible to rank the news topics by combining these factors. There are several advantages if these news topics are combined, filtered and ranked from both news supplier and individuals. The most important usage is to enhance the quality and analysis of news recommender systems.

It is an unsupervised system which adequately recognizes the news topics that are common in both news media and social media and finally ranks them by the significance degrees of MF, UA and UI. It undergoes a speculative framework including and incorporating various techniques such as graph clustering, measure of similarity, keyword extraction and social network analysis .To acquire this objective, and ranking system uses keywords from news media to discover the overlap with social media content. Finally a graph is constructed whose nodes symbolize. The keywords and the edges illustrate their coexistences in social media, the topic clusters are obtained (TCs) and the topics are ranked according to the degrees of MF, UA, and UI.

## II Related Work

Much research has been carried out in the field of topic identification referred to more formally as topic modeling. Two traditional methods for detecting topics are LDA and PLSA. Blei et al[1] have reported ,Latent Dirichlet Allocation, LDA is a generative probabilistic model that can be applied to different tasks, including topic identification. T. Hofmann[2] have reported ,Probabilistic Latent Semantic Analysis ,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. Furthermore, LDA and PLSA only discover topics from text corpora; they do not rank based on popularity or prevalence

Wartena et al[3] have reported, Topic detection by clustering keywords, implemented a method to detect topics by clustering keywords. Their method entails the clustering of keywords based on different similarity measures using the induced k-bisecting clustering algorithm [4].Archetti et al[4] have reported ,A hierarchical document clustering environment based on the induced bisecting k-means, they do not employ the use of graphs, they do observe that a distance measure based on the Jensen Shannon divergence (or information radius) of probability distributions performs well. More recently, research has been conducted in identifying topics and events from social media data, taking into account temporal information. Cataldi et al[5] have reported, Emerging topic detection on Twitter based on temporal and social terms evaluation, a topic detection technique that retrieves real-time emerging topics from Twitter. Their method uses the set of terms from tweets and model their life cycle according to a novel aging theory. Additionally, they take into account social relationships more specifically, the authority of the users in the network to determine the importance of the topics.

Zhao et al[6] have reported, Comparing Twitter and traditional media using topic models, carried out similar work by developing a Twitter-LDA model designed to identify topics in tweets. Their work, however, only considers the personal interests of users, and not prevalent topics at a global scale. Another trending area of related research is the detection of bursty topics (i.e., topics or events that occur in short, sudden episodes).Diao et al[7] have reported ,Finding bursty topics from microblogs, a method that uses a state machine to detect bursty topics in microblogs. Their method also determines whether user posts are personal or refer to a particular trending topic.

Yin et al[8] have reported, A unified model for stable and temporal topic detection from social media data, also developed a model that detects topics from social media data, distinguishing between temporal and stable topics. 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. Wang et al[9] have reported, Automatic online news topic ranking using media focus and user attention based on aging theory, a method that takes into account the users' interest in a topic by estimating the amount of times they read stories related to that particular topic. They refer to this factor as the UA. They also used an aging theory developed by Chenet al.[10] to create, grow, and destroy a topic.

Chen et al[10] have reported, Life cycle modeling of news events using aging theory,". The life cycles of the topics are tracked by using an energy function. The energy of a topic increases when it becomes popular and it diminishes over time unless it remains popular. We employ variants of the concepts of MF and UA to meet our needs, as these concepts are both logical and effective. Other works have made use of Twitter to discover news-related content that might be considered important. Sankaranarayanan et al[11] have reported, TwitterStand: News in tweets, a system which identifies that correspond to breaking news. They accomplish this by utilizing a clustering approach for tweet mining.

Phelan et al[12] have reported ,Using Twitter to recommend real-time topical news, a recommendation system that generates a ranked list of news stories. News is ranked based on the co-occurrence of popular terms within the users RSS and Twitter feeds. Both of these systems aim to identify emerging topics, but give no insight into their popularity over time. Moreover, the work by Phelan et al. [12] only produces a personalized ranking (i.e., news articles tailored specifically to the content of a single user), rather than providing an overall ranking based on a sample of all users. Nevertheless, these works provide us with a basis for extending the premise of UA. Research has also been carried out in topic discovery and ranking from other domains.

Shubhankar et al[13] have reported, An efficient algorithm for topic ranking and modeling topic evolution, an algorithm that detects and ranks topics in a corpus of research papers. They used closed frequent keyword-sets to form topics and a modification of the Page Rank [14] algorithm to rank them. Brin et al[14] have reported ,Reprint of the anatomy of a large-scale hyper textual web search engine, their work however, does not integrate or collaborate with other data sources.

## III. Proposed System

The objective of this system ranking social media is to determine, consolidate and rank the most popular revalent topics discussed in both media news and media social news during a particular period of time. To determine its agenda, the method must process four main steps

**A) Preprocessing:** The terms which are important are with drawed and filtrate from data news and social media matching to the specific period of time. Pre-processing is the mining technique that involves exchanging raw data into an acceptable format.

**B) Key Term Graph Construction:** The graph is built from the previous retrieved key term set, whose vertices considers the terms key and edges considers the co-exist the likeliness between them. The graph, after processing step and pruning, contains little joint clusters of concepts admired in both media news and media social.

**C) Graph Clustering:** The graph is combined in order to obtain precise and disconnect TCs. Graph clustering is the

Work of collecting the vertices of the graph into gathering and taking into inspection of the edge construction of the graph.

**D) Content Selection and Ranking:** The TCs from the graph are elected and ranked using the three popular

factors (MF, UA, and UI). Initially, news media and tweets from twitter data are selected from the cyberspace and saved in the database. News

articles are obtained from specific news websites via their RSS feeds and tweets are crawled from the Twitter public timeline. A user then requests an output of the top k ranked news topics for a specified period of time between date d1 (start) and date d2 (end).

We can find a way to filter noise and only capture the news.

- We can filter the news based on topic
- Main use potential to improve the quality and coverage of news recommender system.

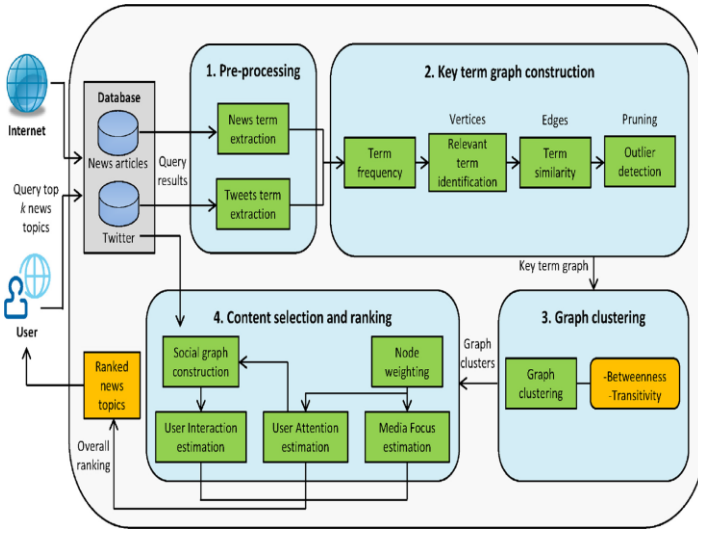


Fig1 : Ranking framework

**E) Stanford POS (Parts Of Speech) Tagger:** Using Stanford POS tagger, a category to which a word is assigned in accordance with its syntactic functions. In English the main parts of speech are noun, pronoun, adjective, determiner, verb, adverb, preposition, conjunction, and interjection. This is helpful to extract the noun from the sentences, which we are considering noun as terms.

**F) N-gram technique:** This methodology is used to find the co-occurrence of the words in the sentences of tweets as well as media news and the Outlier detection.

We are implementing two gram and three gram techniques.

Here is the sentence

Bangalore is the Silicon City of India.

Here key words are Bangalore, Silicon, City, India

No of keywords are 4 let us take it as N.

For two gram, number of loops is N-1

Bangalore-Silicon

Silicon-City

City-India

For three gram, number of loops is N-2

Bangalore-Silicon-City

Silicon-City-India

**G) Cosine Similarity:** This methodology is used to find the similarity between the sentences. If the cosine value of two sentences is 1 means, those are 100% similar, if it is 0.98 means 98% similar, this is useful to find that where the sentences related to the same terms. Here are two very short texts to compare:

Julie loves me more than Linda loves me

Jane likes me more than Julie loves me

We want to know how similar these texts are, purely in terms of word counts (and ignoring word order). We begin by making a list of the words from both texts:

me Julie loves Linda than more likes Jane

Now we count the number of times each of these words appears in each text:

| | | |
|---|---|---|
| me | 2 | 2 |
| Jane | 0 | 1 |
| Julie | 1 | 1 |
| Linda | 1 | 0 |
| likes | 0 | 1 |
| loves | 2 | 1 |
| more | 1 | 1 |

than 1 1

We are not interested in the words themselves though. We are interested only in those two vertical vectors of counts. For instance, there are two instances of 'me' in each text. We are going to decide how close these two texts are to each other by calculating one function of those two vectors, namely the cosine of the angle between them. The two vectors are, again:

a: [2, 1, 0, 2, 0, 1, 1, 1]

b: [2, 1, 1, 1, 1, 0, 1, 1]

The cosine of the angle between them is about 0.822

$$\text{dice\_QS}(i,j) = \begin{cases} 0 & \text{if co}(i,j) \le \vartheta \\ 2\times\text{co}(i,j) & \text{otherwise} \end{cases} \quad \text{……(1)}$$

$$\frac{}{\text{dftop}(i)+\text{dftop}(j)}$$

**H) Group Clustering:** This methodology is used to create the Clusters with respect to the terms from the tweets as well as media news. By this methodology we will get the count of tweets and media news which lay in the cluster, by that we can achieve the Media Focus (MF) and User Interaction (UI).

$$\text{betweenness}(e) = \sum_{i,j \in V} \frac{\sigma(i,j|e)}{\sigma(i,j)} \quad \text{……(2)}$$

**Algorithm 1** Improve the Cluster Quality of a Graph

1: **Input:** Graph *G*

2: **Output:** Cluster-quality-improved *G*

3: *B* = {} _ empty set

4: **repeat**

5: **for all** (edge *e* ∈ *G*) **do**

6: Calculate betweenness(*e*) and append to *B*

7: **end for**

8: **if** first iteration of loop **then**

9: *bavg* = avg(*B*)

10: **end if**

11: *bmax* = max(*B*)

12: *trans*0 = transitivity(*G*) _ previous transitivity

13: Remove edge with *bmax* from *G*

14: *trans*1 = transitivity(*G*) _ posterior transitivity

15: Clear set *B*

16: **until** (*trans*1 < *trans*0 or *bmax* < *bavg*)

17: Add edge with *bmax* to *G*

We apply the concepts of betweenness and transitivity in our graph clustering algorithm, which disambiguates potential topics. The process is outlined in Algorithm 1.

## IV. Experiments And Results

The tweets crawled from Twitter public timeline and news articles crawled from popular new websites during the period between November 1, 2013 and February 28, 2014 from the testing dataset. The werecnn.com, bbc.com, cbsnews.com, reuters.com, abcnews.com and usatoday.com are the crawled news websites. A total of 105856 news articles and totally 175 044 074 bilingual tweets were collected over the specified period of time. 71 731 730 tweets were remaining after non English tweets discarded. Dataset is divided into two partition collected over the specified period of time

1) From January and February 2014 data were used as the testing dataset, on which experiments were performed for the overall method evaluation.

2)From November and December 2013 data were used as the control dataset, where experiment were performed to establish adequate thresholds and select measures that presented the best results.

| Time period | # topics | Avg. tweets | Avg. news | Avg. users |
|---|---|---|---|---|
| 2014/01/01–10 | 84 | 2138 | 17 | 430 |
| 2014/01/11–20 | 112 | 1585 | 13 | 788 |
| 2014/01/21–30 | 100 | 2615 | 20 | 1626 |
| 2014/02/01–10 | 78 | 3113 | 17 | 776 |
| 2014/02/11–20 | 106 | 3567 | 12 | 932 |
| 2014/02/21–28 | 79 | 2386 | 16 | 398 |
| Average | 97 | 2567 | 16 | 894 |

The challenging process is evaluation of topic ranking as interpretation of the results is generally subjective. An attempt is made to show that the ranked topics are indeed those that user prefer however, a method for ranking popular news topic must be established.

To retrieve the most popular topics, Google's news aggregation service [48] was utilized, the top 10 news storie displayed on this sites were collected at the end of the day for each day from November1, 2013 to February 28, 2014.

The titles of the top 10 new stories from each day were asked to view the next 20 master's and doctoral student and select the ones they considered relevant. A minimum of two articles per day was required select from each participant and a maximum of all 10.

The stories were grouped into topics manually for each range, a score was given for each topics and its corresponding time range. For each topic the score is calculated by multiplying the number of times the participants voted for the stories in the topic by the topic by the total number of stories in it. In descending order the topics were ranked finally of their score, resulting in a ranked topic list for each time range. The ranked list of topics is refer hereafter as voted topics. In fig 2 the percentage of topics selected by SociRank and by MF is shown that overlap with the voted topics. It clearly outperforms MF in terms of overlap with the voted topics can be seen in the figure.(i.e., the topics selected by user are the most important). This indicates that better thing at discovering to a method that only utilizes data from the news media the user find interesting.It produces a very different ranked list of the news topics that signify on high-frequency relying on news topics. The MF alone is a substandard considering all the results and estimator of what users find interesting or consider important, and therefore in this way should not be used, on the other hand SociRank, proves the performances to be more capable, and so we conclude that the SociRank provides information can prove vital in commerce-based areas where the users interest is paramount.

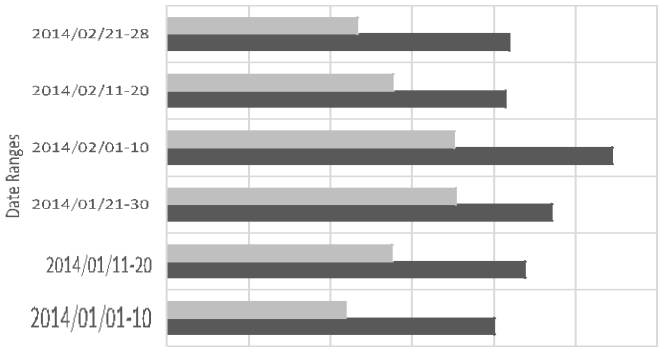


Fig. 2. Percentage of overlap between all voted topics and all topics selected by Ranking and MF.

## V.**Conclusion**

It is an unsupervised method of ranking which identifies news topics prevalent in both social media and the news media, and then ranks them by taking into account their MF, UA, and UI as relevance factors. The temporal prevalence of a particular topic in the news media is considered the MF 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 UA. Finally, the interaction between the social media users who mention the topic indicates the strength of the community discussing it and is considered the UI. 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 topic.

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. Ranking can also be extended and adapted to other topics besides news such as science technology sports and other trends. The extensive experiments to test the performance of ranking, including controlled experiments for its different components. It has been compared to media focus only ranking by utilizing results obtained from a manual voting method as the ground truth.

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 MF only and ranking them by including UA and UI. This distinction provides a basis for the importance of this paper and clearly demonstrates the shortcomings of relying solely on the mass media for topic ranking.

**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 detec- tion on Twitter based on temporal and social terms in *Proc. 10th Int. Workshop Multimedia Data Min. (MDMKDD)*, Washington, DC, USA, 2010, Art. no. 4. [Online]. http://doi.acm.org/10.1145/1814245.1814249

[6]W. X. Zhao *et al.*, "Comparing Twitter and traditional media using topic models," 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] H. Yin, B. Cui, H. Lu, Y. Huang, and J. Yao, "A unified model for stable and temporal topic detection from social media data," in *Proc. IEEE 29th Int. Conf. Data Eng. (ICDE)*, Brisbane, QLD, Australia, 2013, pp. 661–672.

[9] 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, qq. 1033–1042.

[10] C. C. Chen, Y.-T. Chen, Y. Sun, and M. C. Chen, "Life cycle modeling of news events using aging theory," in *Machine Learning: ECML 2003*. Heidelberg, Germany: Springer Berlin Heidelberg, 2003, rr. 47–59.

[11] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "TwitterStand: News in tweets," in *Proc. 17th ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, Seattle, WA, USA, 2009, pp. 42–51.

[12] O. Phelan, K. McCarthy, and B. Smyth, "Using Twitter to recom-mend real-time topical news," in *Proc. 3rd Conf. Recommender Syst.*, New York, NY, USA, 2009, pp. 385–388.

[13] K. Shubhankar, A. P. Singh, and V. Pudi, "An efficient algorithm for topic ranking and modeling topic evolution," in *Database Expert Syst. Appl.*, Toulouse, France, 2011, pp. 320–330.

[14] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertex-tual web search engine," *Comput. Netw.*, vol. 56, no. 18, pp. 3825–3833, 2012.