RECOGNIZE AND CATEGORIZE PREVALENT NEWS TOPICS USING SOCIAL MEDIA FACTORS

[1]A EMMANUEL RAJU , M.Tech ASSISTANT PROFESSOR
[2]H ATEEQ AHMED , M. Tech, ASSISTANT PROFESSOR
[3]SHAIK ABDUL JABBAR, M.Tech, ASSISTANT PROFESSOR
DEPARTMENT OF CSE
Dr. K V SUBBA REDDY INSTITUTE OF TECHNOLOGY, DUPADU, KURNOOL

ABSTRACT

To predict interactions between social media and traditional news streams is becoming increasingly relevant for a variety of applications, including: understanding the underlying factors that drive the evolution of data sources, tracking the triggers behind events, and discovering emerging trends. Researchers have developed such interactions by examining volume changes or information diffusions, however, most of them ignore the semantical and topical relationships between news and social media data. Our work is the first attempt to study how news influences social media, or inversely, based on topical knowledge. We introduce a hierarchical Bayesian model that jointly models the news and social media topics and their interactions. We show that our proposed model can capture distinct topics for individual datasets as well as discover the topic influences among multiple datasets. By applying our model to large sets of news and tweets, we demonstrate its significant improvement over baseline methods and explore its power in the discovery of interesting patterns for real world cases.

KEYWORDS: Information filtering, Social Computing, Social network analysis, Topic identification, Topic ranking

I. INTRODUCTION

Today, online social media such as Twitter have served as tools for organizing and tracking social events. Understanding the triggers and shifts in opinion driven mass social media data can provide useful insights for various applications in academia, industry, and however, there remains a general lack of finding of what causes the hot spots in social media. Typically, the reasons behind the rapid spread of information can be summarized in terms of two categories: exogenous and endogenous factors. Growing factors are the results of information diffusion inside the social network itself, namely, users obtain information primarily from their online social network. In contrast, exogenous factors mean that users get information from outside sources first, for example, traditional news media, and then bring it into their social network.

Although previous works have explored both the social media and external news data datasets, few researchers have looked at the endogenous and exogenous factors based on semantical or topical knowledge. They have either sought to identify relevant tweets based on news articles or simply correlated the two data sources through similar patterns in the changing data volume. Still within the same data source, there could be various factors that drive the evolution of information over time. Exogenous factors across multiple datasets make analyzing the evolution and relationship among multiple data streams more difficult. Watching social media and outside news data streams in a united frame can be a practical way of solving this problem. In this paper, we propose a novel topic model, News and Twitter Interaction Topic model (NTIT), that jointly learns social media topics and news topics and subtly capture the influences between topics. The intuition behind this approach is that before a user posts a message, he/she may be influenced either by opinions from his/her online friends or by articles from news agencies. In our new framework, a word in a tweet can be responsive to the topical influences coming either from endogenous factors (tweets) or from exogenous factors (news)

A straightforward approach for identifying topics from different social and news media sources is the application of topic modeling. Many methods have been proposed in this area, such as latent Dirichlet allocation (LDA) and probabilistic latent semantic analysis (PLSA). Topic modeling is, in essence, the discovery of ―topics‖ in text corpora by clustering together frequently co-occurring words. This approach, however, misses out in the temporal component of prevalent topic detection, that is, it does not take into account how topics change with time. Furthermore, topic modeling and other topic detection techniques do not rank topics according to their popularity by taking into account their prevalence in both news media and social media.

We introduce 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.

II. SYSTEM DESIGN

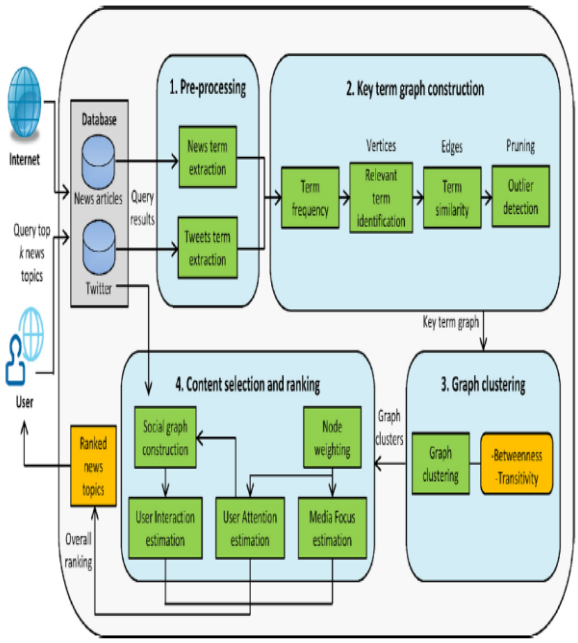

Fig1: Soci Rank Frame work

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 d1 (start) and date d2 (end).

III. IMPLEMENTATION

Admin

In this module, the Admin has to login by using valid user name and password. After login successful he can perform some operations such as Authorizing users, Login ,View all users and authorize, give click option to view all users locations in GMap using Multiple Markers ,View all Friend Request and Response ,View all users time line tweet details with Soci rank, rating and give tweet ,View all tweets by clustering based on tweet name and show tweeted details, Soci_Rank ,rating and View all Relevant Term Identification on all tweets and group together(similar tweeted details for each and every created tweet) ,View all users outlier detection tweet with its tweeted details, Soci_Rank, rating and View all term frequency on all tweets count(Display the tweets which is getting tweet regularly) based on tweet name, View all tweet news Socirank in chart and View all tweet term frequency count in chart based on date and time, View all tweets tweeted socirank in chart

Friend Request & Response

In this module, the admin can view all the friend requests and responses. Here all the requests and responses will be displayed with their tags such as Id, requested user photo, requested user name, user name request to, status and time & date. If the user accepts the request then the status will be changed to accepted or else the status will remains as waiting.

- **User**

In this module, there are n numbers of users are present. User should register before performing any operations. Once user registers, their details will be stored to the database. After registration successful, he has to login by using authorized user name and password. Once Login is successful user can perform some operations like Register with Location with lat and login using GMap and Login, View Your Profile with location ,Search Friend and Find Friend Request, View all Your Friends Details and Location Route path from Your Location, View all your time line tweets with Soci rank, rating and give tweet, Create tweet for News like Tweet name, tweet uses, Tweet desc(enc),tweet image and View all your tweet with re tweet details,Socirank,rating,Search tweet and list all Tweets and view its details and give re tweet, give rank by hyper link and View all your friends Tweets and give Tweet

Searching Users to make friends

In this module, the user searches for users in Same Site and in the Sites and sends friend requests to them. The user can search for users in other sites to make friends only if they have permission.

IV. ANALYZING TWITTER DATA DURING REAL-WORLD EVENTS

The posts and activity on Twitter, impacts and plays a vital role in various real world events. Role of Twitter has been analyzed by computer scientists, psychologists and sociologists for impact in the real-world. Twitter has progressed from being merely a medium to share users' opinions; to an information sharing and dissemination agent; to propagation and coordination of relief and response efforts. Some of the popular case studies analyzed by computer scientists have been, Twitter activities during elections, natural disasters (like hurricanes, wildfires, floods, etc.), political and social uprisings (like Libya and Egypt crisis) and terrorist attacks (like Mumbai triple bomb blasts). Content and user activity patterns of Twitter during events have been analyzed for both positive and negative aspects. Some of the problems studied that result in bad quality of data, presence of spam and phishing posts, content spreading rumors / fake news, privacy breach of users via the content shared by them and use of Twitter for propagation and instigation of hate among people. Researchers have used machine learning, information retrieval, social network analysis and image and video analysis for the purpose of analyzing and characterizing Twitter usage during real-world events.

We introduce some of the research work done in applying user modeling techniques to analyze behavior of users on social networks. Yin et al. modeled user behavior using two factors: the topics related to users' intrinsic interests and the topics related to temporal context . They

created a latent class statistical mixture model, called Dynamic Temporal Context-Aware Mixture model (DTCAM). They evaluated their system on four large-scale social media datasets. The authors demonstrated how user modeling techniques can be effectively used to improve the performance of recommender systems for social networks. Xu et al. introduced a mixed latent topic model to combine various factors to model users' posting behavior on Twitter. The authors assumed that a user's behavior is influenced by three factors: breaking news, posts from social friends and user's interest. They developed and showed that their model outperforms other user models in handling the perplexity of held-out content and the quality of generated latent topics. Abel et al. developed a user modeling framework for news recommendations on Twitter using more than 2 million tweets. The authors proposed different strategies for creating hash tag-based, entity based or topic-based user profiles using semantic enrichment and temporal factors. Their results showed that consideration of temporal profile patterns can improve recommendation quality.

4.1 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 (subgraphs) in the graph. Before explaining the graph clustering algorithm, the concepts of betweenness and transitivity must first be understood. 1) Betweenness: Matsuo et al. [38] proposed an efficient approach to achieve the clustering of co-occurrence graphs. They use a graph clustering algorithm called Newman clustering [39] 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 intercluster 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. As outlined by Brandes [45], the betweenness measure of an edge e is calculated as follows:

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

where V is the set of vertices, σ (i, j) is the number of shortest paths between vertex i and vertex j, and σ (i, j|e) is the number of those paths that pass through edge e. By convention, 0/0 = 0.

4.2 **Transitivity:**

Iwasaka and Tanaka-Ishii [37] developed a method that accomplishes the clustering of a co-occurrence graph based on a concept known as transitivity. Transitivity is a property in a relation between three elements such that if the relation holds between the first and second elements, and between the second and third elements, then it also holds between the first and third elements. The authors indicate that each output cluster is expected to have no ambiguity, and that this is only achieved when a graph's edges—representing cooccurrence relations—are transitive. Thus, a graph with higher global transitivity is considered to have better cluster quality than one with lower global transitivity. The transitivity of a graph G is defined as

$$\text{transitivity}(G) = \frac{\#\text{triangles}}{\#\text{triads}}$$

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 $e \in G$) **do**
6: Calculate betweenness(e) and append to B
7: **end for**
8: **if** first iteration of loop **then**
9: $b_{avg} = \text{avg}(B)$
10: **end if**
11: $b_{max} = \max(B)$
12: $trans_0 = \text{transitivity}(G)$ ▷ previous transitivity
13: Remove edge with b_{max} from G
14: $trans_1 = \text{transitivity}(G)$ ▷ posterior transitivity
15: Clear set B
16: **until** ($trans_1 < trans_0$ or $b_{max} < b_{avg}$)
17: Add edge with b_{max} to G

4.3 Graph Clustering Algorithm:

We apply the concepts of betweenness and transitivity in our graph clustering algorithm, which disambiguates potential topics. The process is outlined in Algorithm 1. First, the betweenness values of all edges in graph G are calculated in lines 5–7. Then, the initial average betweenness of graph G is calculated in line 9; we wish for all edges to approach this betweenness. To achieve this, edges with high betweenness values are iteratively removed in order to separate clusters in the graph (line 13). It is worth pointing out that set B, which keeps track of all betweenness values in the graph, is emptied at the end of each iteration. The edge-removing process is stopped when removing additional edges yields no benefit to the clustering quality of the graph. This occurs: 1) when the transitivity of G after removing an edge is lower than it was before or 2) when the maximum edge betweenness is less than the average—the latter indicating that the maximum betweenness is considered normal when compared to the average. Once the process has been stopped, the last-removed edge is added back to G.

4.4Time Complexity Analysis of Graph Clustering:

In this section, we provide a brief time complexity analysis of the graph clustering algorithm described above. Our initial assumption is that, after performing the outlier detection step (Section III-B4) to obtain a reduced set of edges, graph G may be considered a sparse graph. In other words, the number of edges in G is approximately the same as the number of vertices, or |E|≈|V|. First, we analyze the time complexity for calculating the betweenness of all edges in G (10). Given that G is sparse, Johnson's algorithm [46] may be used for calculating the shortest paths of all vertex pairs. This algorithm runs in O(|V| 2 log |V|+|V||E|) time, performing faster than the Floyd–Warshall algorithm [47], which solves the problem in O(|V| 3). The shortest paths of all vertex pairs are first computed and stored in a data structure in order to avoid calculating them for every edge (lines 5–7).

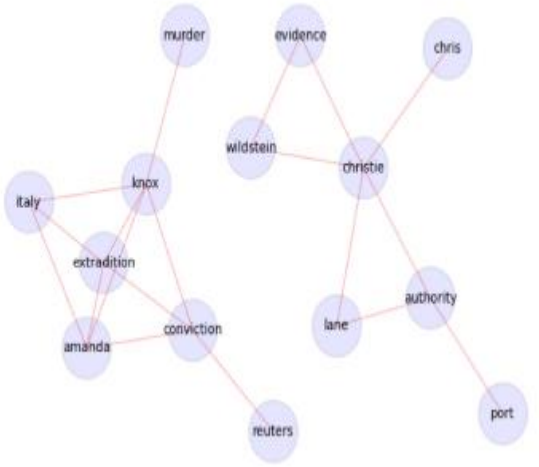

Fig2: Unambiguous TCs in graph G after running the cluster-quality improvement algorithm

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.

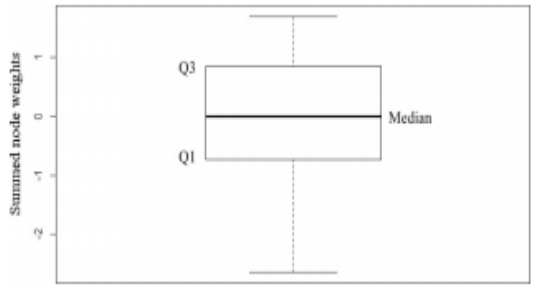

Fig3: Boxplot of the summed node weights of valid combinations of a TC discovered on a particular date.

V. EXPERIMENTS AND RESULTS

The testing dataset consists of tweets crawled from Twitter public timeline and news articles crawled from popular news websites during the period between November 1, 2013 and February 28, 2014. The news websites crawled werecnn.com, bbc.com, cbsnews.com, reuters.com, abcnews.com, and usatoday.com. Over the specified period of time, a total of 105 856 news articles and 175 044 074 bilingual tweets were collected. After non-English tweets were discarded, 71 731 730 tweets remained. The dataset was divided into two partitions. 1) Data from January and February 2014 were used as the testing dataset, on which experiments were performed for the overall method evaluation. 2) Data from November and December 2013 were used as the control dataset, where experiments were performed to establish adequate thresholds and select measures that presented the best results.

5.1 Method Evaluation

The evaluation of topic ranking is quite challenging, as the interpretation of the results is generally subjective. However, in an attempt to show that the ranked topics are indeed those that users would prefer, a method for ranking popular news topics must be established. To retrieve the most popular topics, Google's news aggregation service [48] was utilized. For each day from November 1, 2013 to February 28, 2014, the top 10 news stories displayed on this site were collected at the end of the day.

TABLE I
SOME STATISTICS RELEVANT TO THE TESTING DATASET

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	99	3113	17	1190
2014/02/11–20	106	3567	12	932
2014/02/21–28	79	2386	16	398
Average	97	2567	16	894

Next, 20 master's and doctoral students were asked to view the titles of the top 10 news stories from each day and select the ones they considered relevant. Each participant was required to select a minimum of two articles per day and a maximum of all 10.

The participants' results were then partitioned into 12 date ranges: 1) November 1, 2013–November 10, 2013; 2) November 11, 2013–November 20, 2013; 3) November 21, 2013–November 30, 2013; 4) December 1, 2013– December 10, 2013; 5) December 10, 2013–December 20, 2013; 6) December 20, 2013–December 31, 2013; 7) January 1, 2014–January 10, 2014; 8) January 11, 2014– January 20, 2014; 9) January 21, 2014–January 31, 2014; 10) February 1, 2014–February 10, 2014; 11) February 11, 2014–February 20, 2014; and 12) February 21, 2014–February 28, 2014. As explained earlier, the first six data ranges were utilized for the controlled experiments and the last six for the method evaluation.

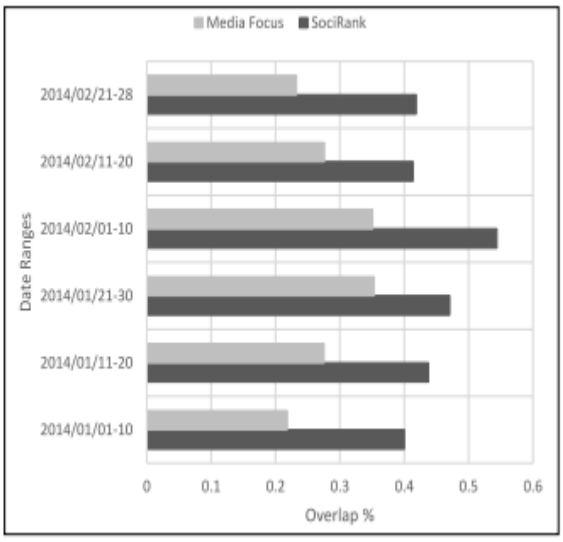

Fig4: Percentage of overlap between all voted topics and all topics selected by SociRank and MF.

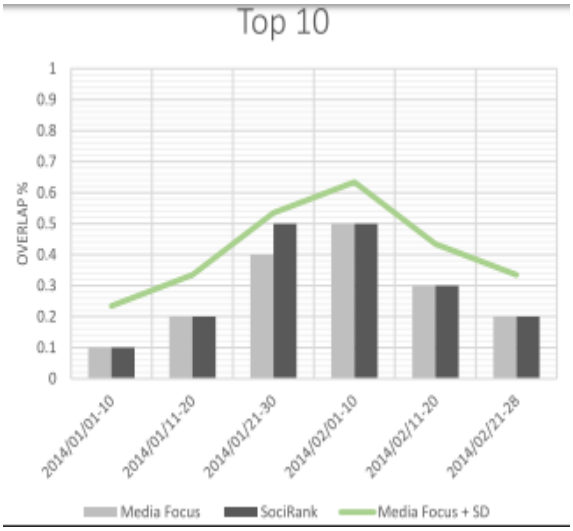

Fig5: Percentage of overlap between top 10 voted topics and top 10 topics selected by SociRank and MF.

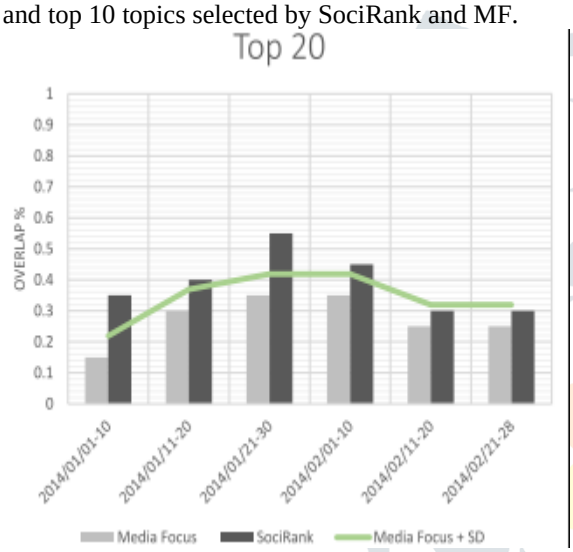

Fig6: Percentage of overlap between top 20 voted topics and top 20 topics selected by SociRank and MF.

5.2 Node Weighting:

The node-weighting formula (13) presented in Section III-D1 takes into account both number of edges connected to a node and their weights. A more naive approach might be to utilize only the number of edges connected to a node as a representation of its weight.

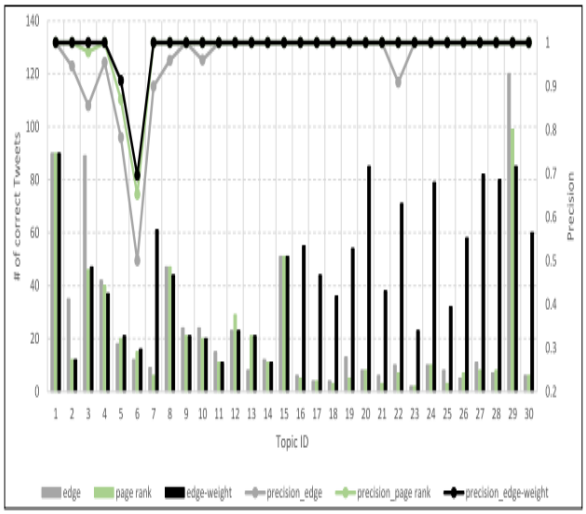

Fig7: Evaluation of different node-weighting approaches

VI. CONCLUSION

System proposed 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 MF, UA, and UI as relevance factors. 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 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 can also be extended and adapted to other topics besides news, such as science, technology, sports, and other trends.

Based on the outputs of model, further efforts are made to understand the complex interaction between news and social media data. Through extensive experiments, we find following factors: 1) even for the same events, focuses of news and Twitter topics could be greatly different; 2) topic usually occurs first in its dominant data source, but occasionally topic first appearing in one data source could be a dominant topic in another dataset; 3) generally, news topics are much more influential than Twitter topics.

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,‖ inProc. 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,‖ inProc. 19th Int. Workshop Database Expert Syst. Appl. (DEXA),Turin, Italy, 2008, pp. 54–58.

[5] F. Archetti, P. Campanelli, E. Fersini, and E. Messina, ―A hierarchical document clustering environment based on the induced bisecting k-means,‖ in Proc. 7th Int. Conf. Flexible Query Answering Syst., Milan, Italy, 2006, pp. 257–269.

[6] C. D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing. Cambridge, MA, USA: MIT Press, 1999.

[7] 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.

[8] W. X. Zhaoet al., ―Comparing Twitter and traditional media using topic models,‖ inAdvances in Information Retrieval. Heidelberg, Germany: Springer Berlin Heidelberg, 2011, pp. 338–349.

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

[10] 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.
[11] 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.
[12] C. C. Chen, Y.-T. Chen, Y. Sun, and M. C. Chen, ―Life cycle modeling of news events using aging theory,‖ inMachine Learning: ECML 2003. Heidelberg, Germany: Springer Berlin Heidelberg, 2003,pp. 47–59.