Ranking & Filtering Prevalent News using Media Factors

Srishank Jaiswal[1] Kumar Vibhaw[2] Harish Kumar.T[3] Chethan.K.[4] VijayaLaxmi Mekali[5]
[1,2,3,4]B.E. Student [5]Assistant Professor
[1,2,3,4,5]Department of Computer Science &Engineering
[1,2,3,4,5]K. S. Institute of Technology, Visvesvaraya Technology University, Belagavi, Karnataka, India

Abstract—*The interactions between social media services and traditional news media services are becoming increasingly important for various applications, like: finding the prevalent news, tracking the triggers behind events, and finding the emerging trends. Researchers have researched 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 survey is the first attempt to study how news influences social media services or inversely, based on techniques and knowledge. We introduce a hierarchical Bayesian model that jointly models the news media services and social media services and we show that our proposed model can capture different topics for individual datasets as well as finds the topic influences among multiple datasets. By proposing our model to large sets of news and tweets or social media data, we will show 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 services such as Twitter is serving as a platform for organizing and updating social events. Finding the triggers and the shifts driven in mass media services and social media services can get us useful information for various applications in academic, industry, and however, there remains a general lack of information of what reason causes the hot spots in social media. The reasons of the fast and enormous spread of information can be concluded in two categories: exogenous and endogenous factors. Growing factors are the results of information diffusion inside the social network itself, namely, users get information first from their online social media services. As exogenous mean that users get information from outside sources first, for example, traditional news media, and then bring it into their social network. Although past researches have found both the social media and outside news data datasets, few researchers have looked at the endogenous and exogenous factors based on semantical or topical knowledge. They have either found 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 a variety of factors that drive the evolution of information over time. Exogenous factors across multiple datasets make analyzing the evolution and relationship among multiple datasets more difficult. Watching social media and outside news data streams in one frame can be a practical way of solving this problem. In this survey paper, we propose a topic model, mass media 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 process 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 method for finding topics from different social media services and news media services is the application of topic modeling. Many methods have been proposed in this area, such as latent Dirichlet allocation (LDA) which is used for identifying topics 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 specification, which means, it does not predict how the topics changes with time. Further, 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 specify 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 applied to various 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. Furthermore, SociRank undergoes an empirical framework, consisting and integrating several processes and methods, 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.

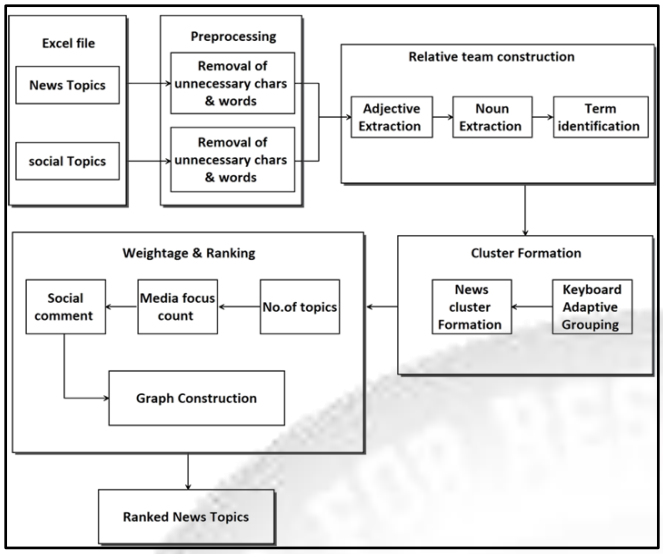

Fig. 1: System Architecture

II. SCOPE

The web application can be used by anyone using a supported computer. Contents are expected in English language. News are supposed to be clearly specified. The process will then be performed accordingly. This web application is not meant to be a talking digital assistant, so user can't have a chat with it, nor can they use this to look up information.

III. PLATFORM

Since smartphones have nowadays become an integral part of our life and are used for work purposes, it is clear that user will be doing many tasks on it. So, it becomes obvious to help out the user, we need this web application to be present on both their computer as well as on their phone. And in order to extend its functionality, both of these components can interact with each other.

A. Computer

Computers are the device which most of us use for majority of tasks. But there can be a chance that user wants to go through either with news media services or social media services and as well as perform their tasks that. It could be something like starting a particular program at startup, shutdown at a particular time or when a particular program completes its task. Thus, the users can save their time by this web application and perform their other necessary tasks. Thus, user becomes free to focus on more important task and don't have to be bothered about the prevalent, relevant and the most trending news.

Software	Platform	Version	Services
Java	Windows	Jdk 1.7	Java development kit
Servlet	Windows	2.2 and Above	capabilities of a server
JSP	Windows	3.0	create dynamically, generated web pages
Tomcat	Windows	6.0	Is used to deploy to your java servlets and JSPs
My-SQL	Windows	5.0	Database management system.
SQLyog	Windows	12.5.1	Is a GUI tool for the RDBMS MySQL.

Table 1: PC Software

B. Web Application

Web application have captured our life as much as mobile application nowadays. They are in a way small computer itself. They can perform productive tasks as well as keep us updated all the time. Some people may be spending equal time on their phone as their computer, or in some case even more. Thus, it becomes an important platform to provide this service. And having this web application be present on both of these platforms and connecting them together ensures that no matter which device the user currently is on, he can always be notified of important and prevalent news. For example, sometimes the user might be doing an important work on his computer as well as he might be interested to know the top trending and prevalent news. Rather than going through all the news again and again, he can simply use this web application to notify him on his computer for the top trending and prevalent news of a specific period. Thus, we save him the effort of repeatedly checking his news media services and he can focus on the task at hand.

IV. EXISTING SYSTEM

Historically, knowledge that apprises the general public of daily events has been provided by mass media sources, specifically the news media. The news media presents professionally verified occurrences or events, while social media presents the interests of the audience in these areas, and may thus provide insight into their popularity.

V. PROPOSEDSYSTEM

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 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.

A. Advantages of Proposed System

- We can a find a way to filter noise and only capture the news.
- We can filter the news based on topics.
- Main use potential to improve the quality and coverage of news recommender system.
- The effectiveness of our system is validated by extensive controlled and uncontrolled experiments.

– Moreover, SociRank undergoes an empirical framework, comprising and integrating several techniques, such as keyword extraction, measures of similarity, graph clustering, and social network analysis.

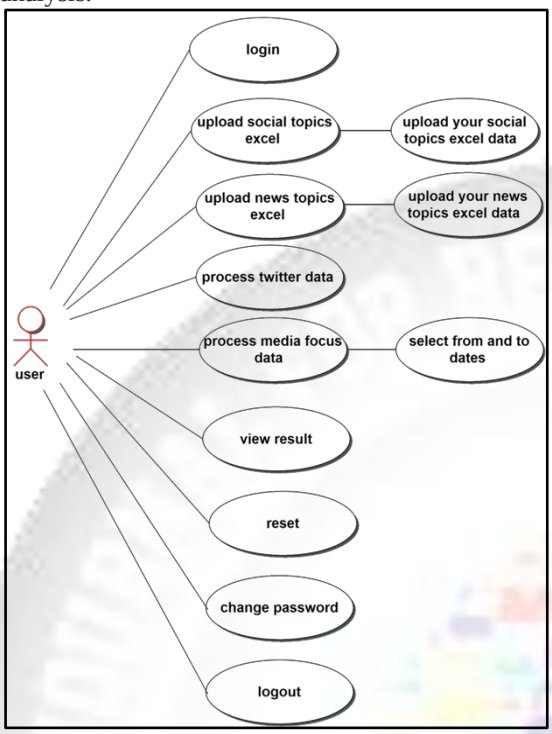

Fig. 2: Use Case Diagram

VI. EMERGENCE OF TWITTER AS A NEWS MEDIA

Computer science research community has analyzed relevance of online social media, in particular Twitter, as news disseminating agent. Kwak et al. showed the prominence of Twitter as a news media, they showed that 85% topics discussed on Twitter are related to news. Their work highlighted the relationship between user specific parameters v/s the tweeting activity patterns, like analysis of the number of followers and follower's v/s the tweeting (re-tweeting) numbers. Zhao et al. in their work, used unsupervised topic modeling to compare the news topic from Twitter versus New York Times (a traditional news dissemination medium). They showed that Twitter users are relatively less interested in world news, still they are active in spreading news of important world events. Lu et al. showed how tweets related to news event on Twitter can be mapped using energy function. The methods proposed act like novel event detection techniques. The study analyzed 900 news events through 2010-2011. Castillo et al. performed qualitative and quantitative analysis on online social media activity about news articles. They concluded that news articles describing breaking news events have more repetitive social media reactions, than in-depth articles.

VII. 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.

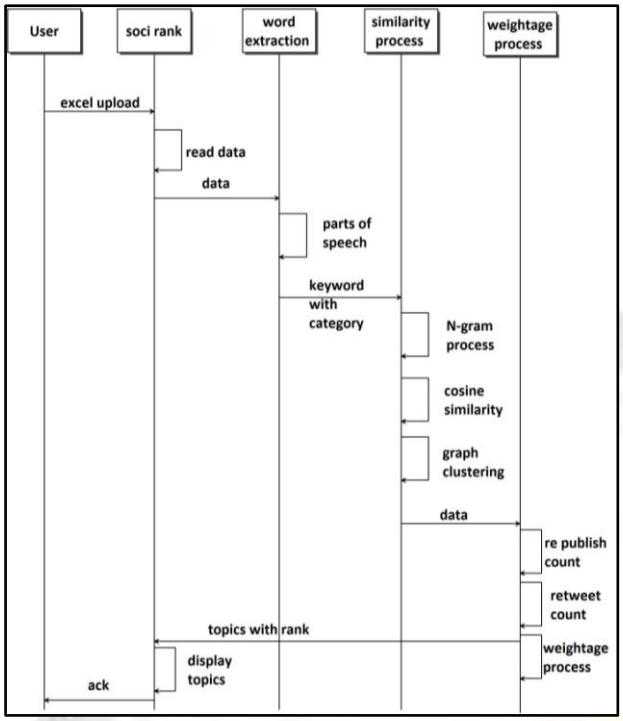

Fig. 3: Sequence Diagram

VIII. CONCLUSION

Our model includes jointly topic modeling on multiple data sources in an asymmetrical frame, which benefits the modeling performance for both long and short texts. We present the results of applying model to two largescale datasets and show its effectiveness over non-trivial baselines. 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.

[13] 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.

[14] O. Phelan, K. McCarthy, and B. Smyth, ―Using Twitter to recommend real-time topical news,‖ inProc. 3rd Conf. Recommender Syst., New York, NY, USA, 2009, pp. 385–388.

[15] 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.

[16] S. Brin and L. Page, ―Reprint of: The anatomy of a large-scale hypertextual web search engine,‖Comput. Network., vol. 56, no. 18, pp. 3825–3833,2012.

[17] E. Kwan, P.-L. Hsu, J.-H. Liang, and Y.-S. Chen, "Event identification for social streams using keyword-based evolving graph sequences," in Proc. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Min., Niagara Falls, ON, Canada, 2013, pp. 450–457.

[18] K. Kireyev, "Semantic-based estimation of term informativeness," in Proc. Human Language Technol. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguist., 2009, pp. 530–538.

[19] G. Salton, C.-S. Yang, and C. T. Yu, "A theory of term importance in automatic text analysis," J. Amer. Soc. Inf. Sci., vol. 26, no. 1, pp. 33–44, 1975.

[20] H. P. Luhn, "A statistical approach to mechanized encoding and searching of literary information," IBM J. Res. Develop., vol. 1, no. 4, pp. 309–317, 1957.

[21] J. D. Cohen, "Highlights: Language- and domain-independent automatic indexing terms for abstracting," J. Amer. Soc. Inf. Sci., vol. 46, no. 3, pp. 162–174, 1995.
[22] Y. Matsuo and M. Ishizuka, "Keyword extraction from a single document using word co-occurrence statistical information," Int. J. Artif. Intell. Tools, vol. 13, no. 1, pp. 157–169, 2004.
[23] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in Proc. EMNLP, vol. 4. Barcelona, Spain, 2004.
[24] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, "KEA: Practical automatic keyphrase extraction," in Proc. 4th ACM Conf. Digit. Libr., Berkeley, CA, USA, 1999, pp. 254–255.
[25] P. D. Turney, "Learning algorithms for keyphrase extraction," Inf. Retrievel, vol. 2, no. 4, pp. 303–336, 2000.
[26] J. Wang, H. Peng, and J.-S. Hu, "Automatic keyphrases extraction from document using neural network," in Advances in Machine Learning and Cybernetics. Heidelberg, Germany: Springer, 2006, pp. 633–641.
[27] T. Jo, M. Lee, and T. M. Gatton, "Keyword extraction from documents using a neural network model," in Proc. Int. Conf. Hybrid Inf. Technol. (ICHIT), vol. 2. 2006, pp. 194–197.
[28] K. Sarkar, M. Nasipuri, and S. Ghose, "A new approach to keyphrase extraction using neural networks," Int. J. Comput. Sci. Issues, vol. 7, no. 3, pp. 16–25, Mar. 2010.
[29] C. Zhang, "Automatic keyword extraction from documents using conditional random fields," J. Comput. Inf. Syst., vol. 4, no. 3, pp. 1169–1180, 2008.
[30] G. Figueroa and Y.-S. Chen, "Collaborative ranking between supervised and unsupervised approaches for keyphrase extraction," in Proc. Conf. Comput. Linguist. Speech Process. (ROCLING), 2014, pp. 110–124.