International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 2 | Issue – 4

NewSociRank: Recognizing and Ranking Frequent News Topics Using Social Media Factors

Harshitha H, Dr. Mohammed Rafi[#]

[#]Professor

Department of Studies in CSE, University B.D.T College of Engineering
(A Constituent college of VTU, Belagavi), Davanagere, Karnataka, India

ABSTRACT

Mass media sources such as news media used to inform us of daily events before. Now a day, unlike news media, social media services like Twitter provide a huge amount of user-generated data, which contain informative news-related content. For these resources to be useful, we need to find a way to filter the noise and capture only the content based on its similarity to the news media. However, even after noise is removed, information overload may still exist in the remaining data-hence, it is convenient to prioritize it for consumption. To achieve prioritization, the information must be ranked in order of estimated importance considering three factors. First, the media focus(MF) of a topic, the temporal prevalence of a particular topic in the news media. Second, user attention (UA), the temporal prevalence of the topic in social media. Last, the interaction between the social media users who mention this topic indicates the strength of the community discussing it, and can be regarded as the user interaction (UI) toward the topic. We propose an unsupervised framework—NewSociRank—which recognizes the news topics prevalent(common) in both social media and the news media, and then ranks them by relevance(popularity) using their degrees of MF, UA, and UI.

1. 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.

A straightforward approach for recognizing 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—NewSociRank—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, NewSociRank 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[17].

2. BACKGROUND AND RELATED WORK

2.1. Keyphrase Extraction

Many methods have been proposed for keyphrase extraction. Most of them are based on machine learning techniques.

Turney[11] proposed viewing keyphrase extraction as classification. In this approach, phrases are extracted from documents and are labeled as keyphrases or non-keyphrases. The documents and labeled phrases are then used as training data for creating a keyphrase classifier. Two learning methods are applied: the decision tree learning algorithm of C4.5 [8] and a genetic algorithm called GenEx. Features such as phrase frequency, position in document are util ized in the classifiers. The classifiers are then used to categorize phrases of a new document as keyphrases or non-keyphra ses. Experimental results show that GenEx can achieve better performance than C4.5.

Kea is a tool for keyphrase extraction based on Naive Bayes [2,14]. In one version of Kea [14], only two features are used: TF-IDF (term frequency-inverse document frequency), and position of the first occurrence. The numerical values of the features are discretized and used to build the Naive Bayes model. In extraction, candidate phrases are ranked according to their probabilities of being keyphrases, and top-ranked phrases are treated as keyphrases. Experimental results show that Kea can achieve a performance comparable to GenEx. Frank et al. [2] extended the Kea model by adding another feature called keyphrase-frequency, which is the frequency of a phrase's being keyphrase in all the documents in the corpus. This feature is effective in domain-specific keyphrase extraction. Turney [12] further improved the Kea model by replacing this domain-specific feature with a number of new features based on co-occurrence measures.

Hulth [5,6] tried three approaches to candidate phrase identification and employed a rule induction system to classify the candidate phrases. Wang et al. [13] used neural network and the back propagation algorithm in keyphrase extraction. For other related work, see [1, 3, 16, 9, 15]

To evaluate the accuracy of the keyphrase extraction methods, measures such as precision and recall are used. Human annotated keyphrases are usually used as positive examples. Turney[10] and Jones and Paynter[7] also proposed ways of human evaluation on the keyphrases extracted by GenEx and Kea.

2.2. Learning to Rank

Ranking is the central problem for many information retrieval applications, such as document retrieval and collaborative filtering. Recently a new research area is emerging in machine learning, which is called learning to rank. Learning to rank aims at automatically creating a model (function) that can perform ranking on instances, using training data and machine learning techniques. Many learning to rank methods have been developed and applied to information retrieval.

Wang et al. [3] proposed 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 Chen et al. [4] to create, grow, and destroy a topic. 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.

Research has also been carried out in topic discovery and ranking from other domains. Shubhankar et al. [5] developed 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 PageRank [6] algorithm to rank them. Their work, however, does not integrate or collaborate with other data sources, as accomplished by NewSociRank.

3. NewSociRank Framework

The goal of our method—NewSociRank—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.

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

II. **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**.**

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

IV. **Content Selection and Ranking:** The TCs from the graph are selected and ranked using the three relevance factors (MF, UA, and UI).

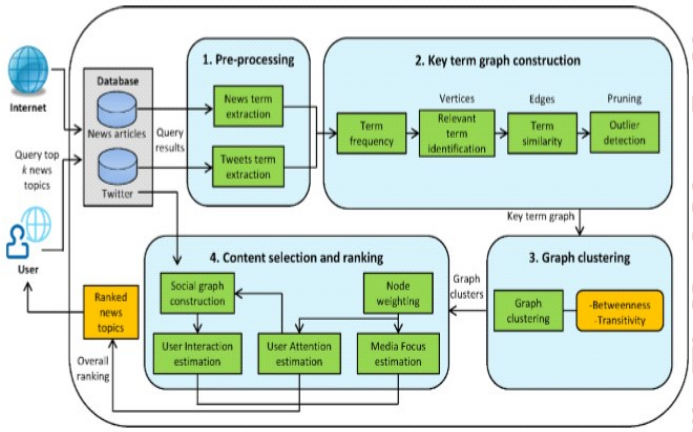

Fig.1. NewSociRank Framework

CONCLUSION

In this paper, we proposed an unsupervised method—NewSociRank—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 topics.

REFERENCES

1. K. Barker and N. Cornacchia. Using noun phrase heads to extract document keyphrases. In Proceedings of the 13th Biennial Conference of the Canadian Society on Computational Studies of Intelligence, pages 40–52, 2000.
2. E.Frank,G.W.Paynter,I.H.Witten,C.Gutwin,andC. G.Nevill-Manning. Domain-specifickeyphraseextraction. In Proceedings of 16th international joint conference on Artificial Intelligence, pages 668–673, 1999.
3. Y. HaCohen-Kerner, Z. Gross, and A. Masa. Automatic extraction and learning of keyphrases from scientific articles. Lecture Notes in Computer Science, 3406:657–669, 2005.
4. J. Han, T. Kim, and J. Choi. Web document clustering by using automatic keyphrase extraction. In Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, pages 56–59, 2007.
5. A. Hulth. Improved automatic keyword extraction given more linguistic knowledge. In Proceedings of the 2003 conference on Empirical Methods in Matural Language Processing, pages 216–223, 2003.
6. A. Hulth. Combining Machine Learning and Natural Language Processing for Automatic Keyword Extraction. PhD thesis, Department of Computer and Systems Sciences, Stockholm University, Sweden, 2004.
7. S. Jones and G. W. Paynter. Automatic extraction of document keyphrases for use in digital libraries: evaluation and applications. Journal of the American Society for Information Science and Technology, 53(8):653–677, 2002.
8. J. R. Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.
9. M. Song, I.-Y. Song, and X. Hu. Kpspotter: a flexible information gain-based keyphrase extraction system. In Proceedings of the 5th ACM international workshop on Web Information and Data Management, pages 50–53, 2003.
10. P. D. Turney. Learning algorithms for keyphrase extraction. Information Retrieval, 2(4):303–336, 2000.

11. P. D. Turney. Learning to extract keyphrases from text. Technical Report ERB-1057, National Research Council, Institute for Information Technology, 2000.

12. P. D. Turney. Mining the web for lexical knowledge to improve keyphrase extraction: Learning from labeled and unlabeled data. Technical Report ERB-1096, National Research Council, Institute for Information Technology, 2002.

13. J.-B. Wang, H. Peng, and J.-S. Hu. Automatic keyphrases extraction from document using backpropagation. In Proceedings of 2005 international conference on Machine Learning and Cybernetics, pages 3770–3774, 2005.

14. I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning. Kea: practical automatic keyphrase extraction. In Proceedings of Digital Libraries 99: The 4th ACM conference on Digital Libraries, pages 254–255, 1999.

15. Y.-F. B. Wu, Q. Li, R. S. Bot, and X. Chen. Finding nuggets in documents: a machine learning approach. Journal of the American Society for Information Science and Technology, 67(6), 2006.

16. R. Mihalcea and P. Tarau. Textrank: Bringing order into texts. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, pages 404–411, 2004.

17. T.Jeya and V.Madhu Bala: ”Socirank: Identifying And Ranking Prevalent Newstopics Using Social Media Factors”, 2018.