

RANKING NEWS BASED ON SOCIAL MEDIA AND NEWS CHANNEL SOURCES USING SOCIAL MEDIA FACTORS

¹Sumit Nihalani, ²Vaishali Malik, ³G.K.Sandhia

SRM Institute of Science & Technology, Kattankulathur, Tamil Nadu, 603203, India

²vaishali_malik@srmuniv.edu.in, ³Sandhia.g@ktr.srmuniv.ac.in,

Abstract: News media channel gives us information about the things and events happening all over the world. In present time, social media services like Twitter give a massive quantity of user-generated information. Now in order to use these resources we need to find some way to filter out the noise so that only the useful content is left which will be based on news media and will be valuable. However, even when we eliminate the noise, info overload should exist within the remaining data—hence, it's a good idea and will be helpful if we rate the data empirically. To realize prioritization, information should be graded so as of calculable importance considering 3 major elements. First, how many days the topic was in news media, and may be thought-about the media focus (MF) of a subject. Second, amount of time the topic was talked about and was active in social media tells about its user attention (UA). Finally, the users of social media who discusses the topic by using tweets, hashtags or replies may be thought to be the user interaction (UI) toward the subject. We came up with an unsupervised framework that identifies news topics dominating in both social media and also the mass media, then ranks them by the degrees of three elements - Media Focus(MF), User Attention(UA) & User Interaction (UI).

1. Introduction

In today's world, extracting and mining valuable info from online sources has become an important zone in IT. Professional journalist publish news in news sites and news resources so those sources can be trusted whereas on the other side internet sites like twitter and other social media resources where people are free to post and give their suggestion is not a very reliable source of information. To identify any topic from social media we need to find a way so that we can filter unwanted information and present only useful information based on the similar contents in news media sources and social media or twitter sources. Based on factors like Media Focus (MF), User Attention(UA) and User Interaction(UI), news topics should be ranked and prioritised. Also, not all news topics can be ranked popular by news sources but are actually popular in news channel. These topics can be

identified and uncovered from social media and can be given coverage in mass media channels. Not only this but the news topic which are filtered out can tell us that how a particular topic is talked about and discussed in different parts to the world or different states of a country which can intern be used as a feedback for government, constitution or any business.

To calculate effectiveness of our proposed system we validate it controlled and uncontrolled system experiments. In order to achieve the goal, keywords are extracted from both the sources, namely mass media and social media from a specific period of time. From the keywords extracted, we form a graph where the nodes denote the extracted keywords and edges denote their occurrences in social media. This graph formed is then clustered to detect unique topics. Post forming Topic Clusters (TCs), the three factors are calculated and finally ranking of the topics is carried out by the overall calculations of these three elements.

2. Literature Review

Topic Detection by Clustering Keywords

Authors: C. Wartena and R. Brussee

It conducts topic identification without the use of any set of category document. The various keywords from the data set are extracted. These keywords are then clustered using the k bisecting algorithm. These clustered keywords match within the article categories from Wikipedia. The paper claims to give best results of finding term distribution by taking co-occurrence of terms.

Probabilistic latent semantic analysis

Authors: T. Hofmann

Probabilistic Latent Semantic Analysis is a applied math technique for the analysis of knowledge, that has applications in info retrieval and filtering, language process, machine learning from text, and in connected areas. Compared to plain Latent linguistics Analysis that stems from algebra and performs a Singular price Decomposition of co-occurrence tables, the projected methodology relies on a mix decomposition derived from a latent category model. This ends up in a lot of

high-principled approach that encompasses a solid foundation in statistics. so as to avoid overfitting, we tend to propose a wide applicable generalisation of most probability model fitting by tempered EM. The approach yields substantial and consistent enhancements over Latent linguistics Analysis in an exceedingly variety of experiments.

Emerging topic detection on Twitter based on temporal and social terms evaluation

Authors: M. Cataldi, L. Di Caro, and C. Schifanella

Twitter is a huge source of data which is generated by users in form of small text messages which are called tweets. It's network is spread across the world among users of all the age group and social background. This paper highlights the role of Twitter in keeping trending topics live and extracts these topics within a time frame through a unique technique. It firstly selects the content and models the life cycle of the words in the content to find out the emerging terms. It uses the page rank algorithm to find out the user authority and their relationship socially. At the end, a graph is constructed which connects emerging terms to find emerging topics.

Title: Comparing Twitter and traditional media using topic models in Advances in Information Retrieval.

Authors: W. X. Zhao

Useful information can be obtained from social network sites like twitter. The information is in abundance but proper analysis of data over Twitter has not been well established and researched Specifically, there is always a doubt that twitter tweets as a source can be used as reliable information source like mass media sources or traditional news datasets and RSS feeds. In this paper by using unsupervised topic modelling we can numerically compare the contents of tweets text with a traditional mass media or news source, New York Times. Twitter-LDA model is used to find out topics from a representative sample of entire Twitter. Topics from both the news sources, mass media and social media, are compared using text mining techniques. We also study the relation between the proportions of opinionated tweets and retweets and topic categories and types. The comparisons results were very useful in finding for downstream IR or DM applications.

3. Algorithm

A. Stop Words Removal

We are using a dictionary based approach to remove stopwords from document. We are using a generic stopword list containing seventy five stopwords which

is created using hybrid approach. The algorithm is implemented as below given steps. The text is divided into tokens and stored in a word array. A stop word is extracted from the stop word list one at a time. Now this word array is searched linearly and compared with the stop words and when both the words, stop words and indexed word in the word array are same then the word is delete from the array. We do this for all the elements present in the word array. This algorithm runs linearly until all the stop words are deleted from the text. Final text is devoid of all the stop words. Various statistics like number of stopwords eliminated from final text, number of words left in the final text, number of words in resultant text, number of occurrence of each stop word count.

B. Preprocessing

The preprocessing stage includes querying all the news blogs, mass media RSS feeds and tweets from the given time frame.

News Term Extraction: News term extraction consist of all the keywords which were generated after removing stopwords from all the articles, blogs and feeds which were generated after stop word removal.

Tweets Term Extraction: All the unique and relevant words from the social media data are extracted. If we talk about twitter, it contains all the tweets. Also generally in hashtags, many words are combined so it is very important to separate them in order to use the hashtag text from tweets.

C. Key Term Graph Construction

Term Document Frequency: First of all we calculate the frequency of each term in the document. For term set N the frequency of each term $n = it's$ presence in the no. of news articles, which is equal to $df(n)$. We calculate the document frequency in the same way

Relevant Key Term Identification: $N =$ keywords present in news. $T =$ Terms in the tweets. We are more interested in terms from news as they have more relevance.

Key Term Similarity Estimation: Now that we have all the words extracted from two different sources we need to find out some relationship in-order to join the edges of our clustered graph. This relationship is defined by term co-occurrence.

Outlier Detection: Even after removing a lot of unwanted terms we still have a lot of terms and co-occurrences in the graph.

Graph Clustering: After constructing the graph with the vertices and edges now we need to find relevant sub- graphs. By using the concept of betweenness and

transitivity.

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

4. Conclusion

With this research paper we suggest an unsupervised method which takes out prevalent news topics from both news websites and social websites. The method then ranks them according to MF, UA and UI. To the best of our knowledge no other work considers interest of both the news website users and social website users. Considering the data from both news and social website, has a lot of advantages. One of the most imp. Advantage is we increase the quality and variety of the news which interests the general public.

It can identify the topics which are no more being discussed by news channels but are still of importance to the general population. We have performed a lot of experiments examine the overall efficiency of the model.

We have compared the result obtained by the model to the result obtained by a manual voting, by a group of individuals for a specific time period. The various experiments conducted indicate that our method is robust enough for providing relevantly ranked news.

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] 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. [Online]. Available: http://dx.doi.org/10.1007/11766254_22.
- [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, Art. no. 4. [Online]. Available: <http://doi.acm.org/10.1145/1814245.1814249>
- [8] 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.
- [9] S.V.Manikanthan and T.Padmapriya "Recent Trends In M2m Communications In 4g Networks And Evolution Towards 5g", *International Journal of Pure and Applied Mathematics*, ISSN NO:1314-3395, Vol-115, Issue -8, Sep 2017.
- [10] T.Padmapriya and V.Saminadan, "Utility based Vertical Handoff Decision Model for LTE-A networks", *International Journal of Computer Science and Information Security*, ISSN 1947-5500, vol.14, no.11, November 2016.

