Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 7, Issue. 5, May 2018, pg.9 – 17

STaR System: Detecting Legitimate Information by Identifying and Prioritizing Prevalent News

Rithesh Pakkala P.[1], Ahana Veeksha Rai B[2], Anusha Shetty[3], Deeksha K Rai[4], Namratha[5]

[1]Assistant Professor, Department of Information Science & Engineering

[2, 3 ,4 ,5]Students , Department of Information Science & Engineering

Sahyadri College of Engineering & Management, Mangaluru, Karnataka, India

[1] rpakkala01@gmail.com, [2] ahanaveeksha@gmail.com, [3] anushashetty405@gmail.com, [4] rai.deeksha1996@gmail.com, [5] namratha.gh@gmail.com

***Abstract*— Mass media such as News media gives us day to day updates about the events happening around the world. Social media is one of the most popular means of communication on the internet. Twitter is one among them which provides large amount of user generated data, which contain information related to news. In this work, a new technique is proposed to filter noise and to obtain the data which is similar in the news media, which is considered to be valuable. However, some noise may still be present in the remaining data, so it is necessary to prioritize it for consumption. Prioritization can be achieved by ranking the information using three factors, firstly, media focus which determines whether the same news has come in the news media or not. Secondly, user attention which tells how many people have shown interest to that particular news. And finally, user interaction where the interests and comments of the user in social media are considered. Hence the proposed work identifies and ranks news topic.**

***Keywords*— News media, Twitter, Prioritization, Media focus, User attention, User interaction, Social media.**

I. INTRODUCTION

Data Mining is the process of automatically identifying the useful information which is taken from large data repositories. The main aim of data mining is to extract information from large datasets and transform them into useful information. This technique can be used to find useful patterns from databases and also provide ways to predict the outcome of observations which might be carried out in the future. The data mining tasks are broadly divided mainly into predictive tasks and descriptive tasks. The aim of predictive task is to predict a value of an attribute based on attribute values. The predicted attribute is called the target variable and the attribute used for prediction is called explanatory variable. The descriptive task aims on deriving patterns which summarizes data relationships. These are often explanatory and they require post processing techniques for validation. The core data mining tasks are predictive modelling, cluster analysis, association analysis and anomaly detection. In

predictive modelling, we build a model for target variable which is a function of explanatory variable. This is further categorized as classification and regression. In cluster analysis, the aim is to find set of related observations which are of the same cluster and are similar to one another when compared to observations of other clusters. Clustering technique can be used to group related objects. Association analysis is used for discovering patterns that define features that are strongly associated with data. The objective is efficient extraction of interesting patterns. In anomaly detection, the task is to identify observations that are diff erent from the characteristics of rest of the data. These observations are the anomalies. The main aim is to detect real anomalies. The data mining applications include Outliers Identification and Detecting Fraud, Customer Profiling, Prediction and Description, Relationship Marketing, Customer Segmentation, Website Design and Promotion. It is used to answer queries, the customer or user asks for. It is also used for the analysis of customer profiles and improvement of marketing plans. The unusual expenses that are claimed by the staff are identified, also the fraud in credit cards are determined. Using web mining, the user behaviour is tracked based on navigation while using the web site. The main source of information for all sites is either from Social Media or News Media. Social media is a place where any user can share any type of information but the information need not be trusted whereas News media is the one where only verified content is published by Journalists. Number of users in Social media is comparatively high when compared to News media, so more data is found in Social media. As a result, the coverage would be more. Also, all data found in Social media cannot be trusted. The news published in News media undergoes some stages where first the origin of the news is found, then the data is verified and finally it is published. The aim of the project is to identify the prevalent message and rank the message or the news from collected data. In social media, the current news may be obtained much faster compared to news media. The news media factor is used for the determination of prevalent data. Prevalent data is obtained by comparing the news common in both social media and social media based on the keywords.

II. RELATED WORK

Literature Survey is a base for research through which knowledge can be gained and also describes the various works done in the relevant field.

Gerardo Figueroa and Yi-Shin Chen [2], introduced a method of automatic key phrase extraction which is based either on supervised approach or unsupervised approach. In supervised approach, the key phrase is extracted using a training document set which acquires knowledge from a collection of texts whereas in unsupervised approach the extraction of key phrase is done by determining the relevance from a single document, without prior learning. For short articles, a hybrid key phrase extraction method called Hybrid Rank is used which includes benefits of both approaches. The system implemented here is the modified version of the supervised and unsupervised methods.

Elizabeth Kwan *et al.* [3], explained about social network which has data about real world events. This data is generated by user. It is more ahead and faster than that of news links. The main goal of this is to identify events from social streams. A Keyword Based Evolving Graph Sequences (KEGS) model is used which captures characteristics of propagation of information in social streams. The disadvantage of this work is that the tweets are limited to English language. Another issue is that several keywords do not refer to one event.

An efficient algorithm to detect and rank the topics was introduced by K. Shubhankar *et al.* [4],. It uses the closed frequent keyword set approach which helps in solving a challenging task where it is difficult to detect and rank topics from a large

academic literature. A modified Time independent Page Rank algorithm is used which assigns an authoritative score to each topic. This score is assigned by considering a sub graph. It also uses the Clustering technique where similarity measure is used. The algorithm is further extended for topic evolution. The algorithm is tested on DBLP and shows that algorithm is effective, scalable and fast.

Comparison is done between the traditional media with twitter by W. X. Zhao *et al.* in [5]. Twitter is a type of social media which has useful information but the analysis of contents on twitter is not carried out. In this work, they also compare the contents of a Twitter with that of a news medium. The concept of unsupervised Topic Modelling is used where a Twitter-LDA model is used to discover topic from a sample taken from twitter. This LDA model which is developed for short tweets is an effective model. In order to compare, they use the Text mining technique where considerations like topic categories and topic types are used. The output is used for IR and DM applications.

M. Cataldi *et al.* [6], used the technique of Topic detection where the most popular emergent topics are retrieved. Twitter is a form of social media which allows sharing text messages called as tweets. The extracted set of terms from twitter undergoes novel aging theory to get the emerging terms. The emerging terms are the ones which occur frequently and rare in the past. The page rank algorithm is used. A topic graph is generated in this technique. This graph connects the emerging terms with related keywords. The social relationship between the networks are analysed to determine the authority of users using page rank algorithm.

A new approach was introduced by K. Sarkar *et al.* [7], which is based on neural network which is used to identify key phrases. Key phrases have a wide range of applications which include retrieval engines, text mining etc. In this paper they use key phrases which give a clue of the document to the reader. It uses a neural network approach in order to extract key phrases from scientific articles. This is better than any other key phrase extraction approach like KEA. To identify the key phrases features like position of the phrase's first appearance, length of the phrase, word length in phrase etc. is used.

K. Kireyev [8], proposed a method of computing specificity of term, based on latent semantic analysis. This work describes about the word informativeness measures. The word should carry more sematic content than others. A method is proposed for calculating the term specificity using the latent sematic analysis approach. The performance of this method is analyzed both qualitatively and quantitatively and it demonstrates excellent performance when compared with the methods that are existing with the help of wide range of tests. It is attempted more deeply at relevant characteristics of word specificity.

A method to build a news processing system, TwitterStand which uses twitter tweets was implemented by Jagan Sankaranarayanan *et al.* [9]. The idea formulated here is to obtain tweets corresponding to the latest news. The technique is used to extract news from a noisy medium. The issue faced in this work is dealing with the data that contains noise. Several strategies that are productive are developed to overcome the noise. TwitterStand works based on our ability in understanding the ways in dealing with different data qualities. For mitigating the noise, an algorithm called online clustering is used.

Owen Phelan *et al.* [10] introduced an approach for recommendation of news which makes use of microblogging activity in real time from twitter through which news can be promoted based on feeds from the user. These feeds are ranked and articles are recommended. A prototype system was introduced and when evaluations were carried out in early steps, the results showed that

users were benefitted from recommendations taken from twitter. A buzzer system would provide an opportunity for experimentation and innovation for tests in future.

Keyword extraction is based on neural network approach. This will tell if a phrase is a key phrase or not. The training algorithm used is backpropagation. For the determination of phrases as key phrase frequency analysis is done, considering the features such as the frequency of term, inverted document frequency, whether the phrase appears in the heading or title of a document, frequency of it appearing in paragraph. The evaluation of this approach is done using the standard information retrieval metrics of recall and precision. This is also efficient when the key phrases are unavailable. This method was proposed by J. Wang, H. Peng, and J.-S. Hu in [11].

III. SYSTEM DESIGN and METHODOLOGY

The Architecture diagram illustrates the overall structure and components associated with the system. The architecture diagram of the proposed system is shown in Figure 1. The system has modules namely Pre-processing, String matching, and Content Selection and Ranking.

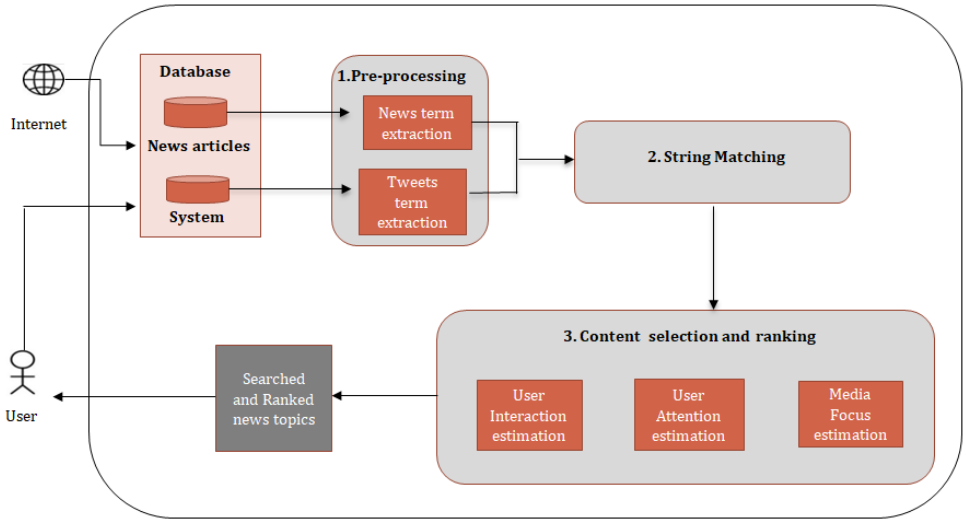

Figure 1: Architecture Diagram

Data from Internet and social media like twitter is taken and collected together. In the Preprocessing module, we collect the data which is stored as terms that contains news term extraction and tweets term extraction. The extracted terms are sent to the next module called keyterm graph construction. In this module, we have term frequency, relevant term identification, term similarity and outliers detection and a graph is created. This graph consists of terms as vertices and relationship between terms form an edge. The keyterm graph thus obtained is sent to the next module called graph clustering. In this module, the graph is clustered to obtain well defined and disjoint topic clusters. In the next module, Content Selection and Ranking, the topic clusters are selected and then ranked based on three factors: Media Focus (MF), User Attention (UA) and User Interaction(UI). The top ranked news topic is available to the user whenever requested.

Figure 2, represents a number of actions called activities and control flow in the system.

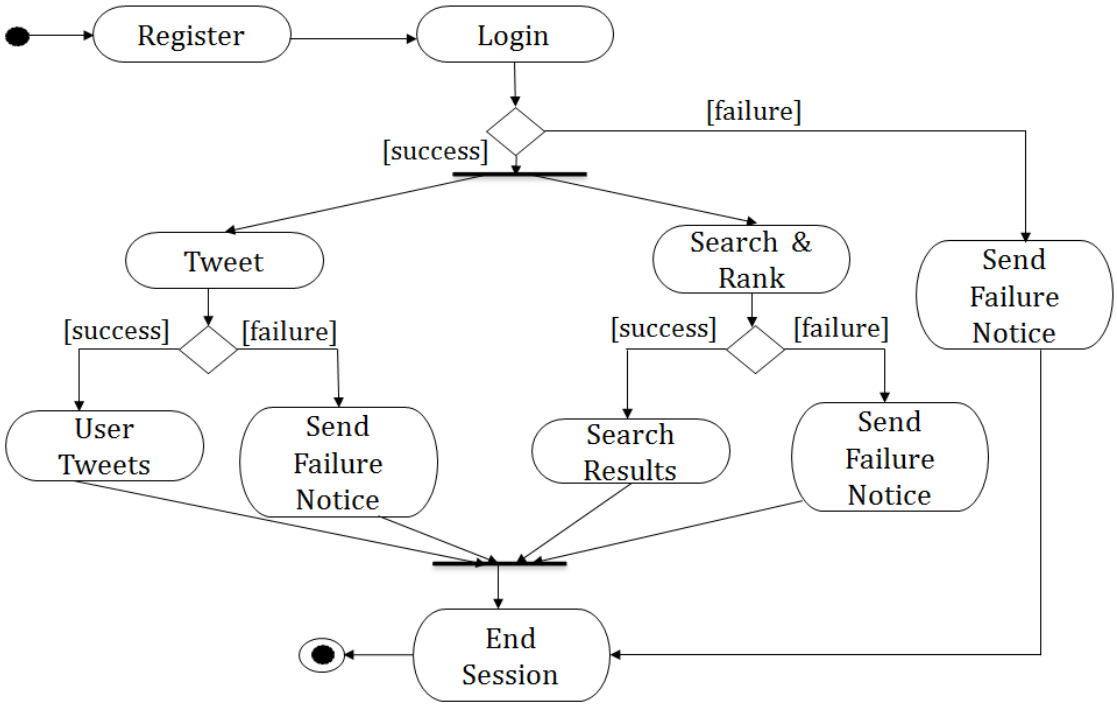

Figure 2: Flow of activities in the system.

The registered user logs in to the system using his username and password. After a successful login the user can perform the concurrent activities, i.e., Search for prevalent news, tweet into twitter and view ranked news. User has to specify a keyword in order to search for news. If the keyword is found then the search results will be displayed else a failure notice will be sent to the user. Similarly, the user can specify tweets which will get posted into twitter if successful else failure notice will be sent to the user. The rank option gives top ranked news to the user. If the login fails, then a failure notice will be sent to the user and the session will be ended.

First, the user registers into the system where he has to provide his name, email ID, and the four API keys i.e., consumer key, consumer secret key, access token and access secret token. Data from Internet and social media like Twitter is taken and stored in the database. There are modules namely Preprocessing, Keyterm graph construction, and Content selection and Ranking. In the first module, Preprocessing terms are stored which contains two set: News term extraction and Tweets term Extraction. The extracted terms are sent to the next module, Keyterm graph construction. In this module, we have term frequency, relevant term identification, term similarity and outliers detection and a graph is created. This graph consists of terms as vertices and relationship between terms form an edge. The keyterm graph thus obtained is sent to the next module, Graph clustering. In the next module, Content Selection and Ranking the topic clusters are selected and then ranked based on three factors: Media Focus (MF), User Attention (UA) and User Interaction (UI). The top ranked news topic is available to the user whenever requested.

Psuedocode for search operation:

```
Step 1      Generate API keys from twitter.
Step 2:     Register using API keys.
Step 3:     Give the login credentials.
Step 4:     Initialize the Twitter DF to 0 and Media DF to 0.
Step 5:     Read the search string.
Step 6:     Compare the search string with all the user tweets.
            If found
            {
              Twitter DF++
            }
            Go to step 7
Step 7:     Compare the search string with news media.
            For i=0 until i<newssize
             {
              If(found)
               {
                Media DF++
               }
              Display genuine news.
             }
            Else
             {
              Go to step 8
             }
Step 8:     If Media DF==0
             Display Fake news.
Step 9:     If search key is not found in user tweets and news media
             Display news not found.
```


IV. RESULTS

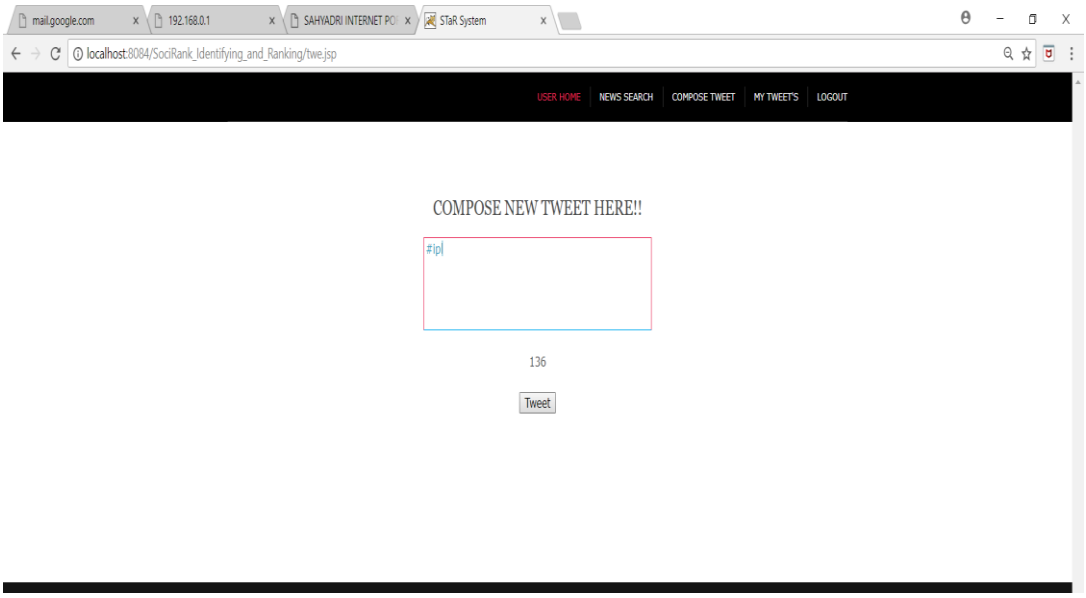

Figure 4: Snapshot of Posting Tweets.

Figure 4 Illustrates how tweets can be directly posted into the twitter account of the user logged into the STaR System. Once the user types the tweet in the text area he can click on the tweet option so that the tweet will get posted into his account.

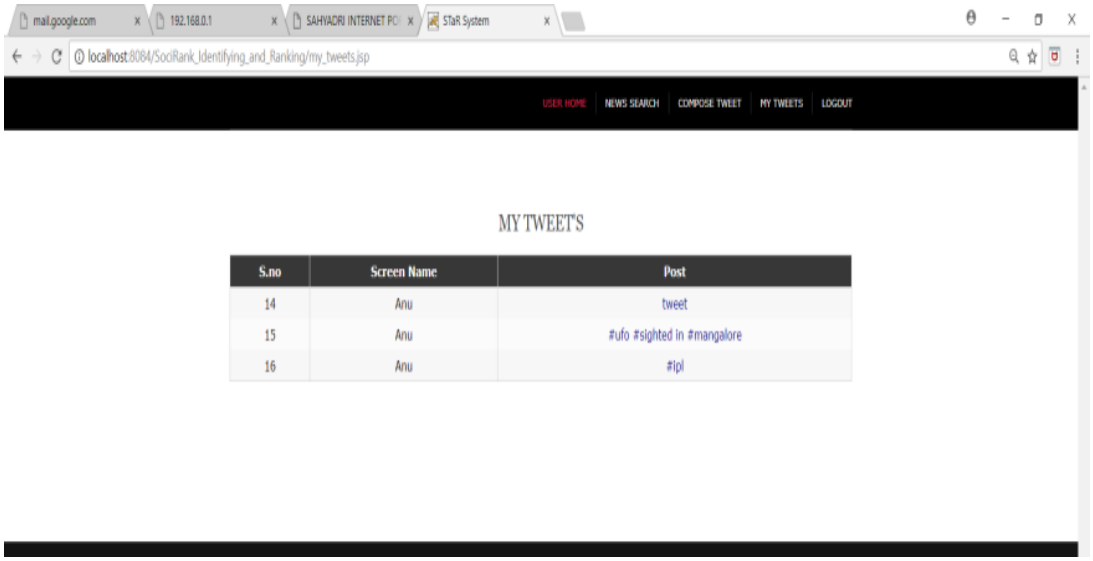

Figure 5: Snapshot of Posted Tweets

Figure 5 illustrates the previously posted tweets which will be stored in a table present in "my tweets". When the user searches for any news, the keyterm will be compared with news present in the database and the contents present in table shown in the above figure.

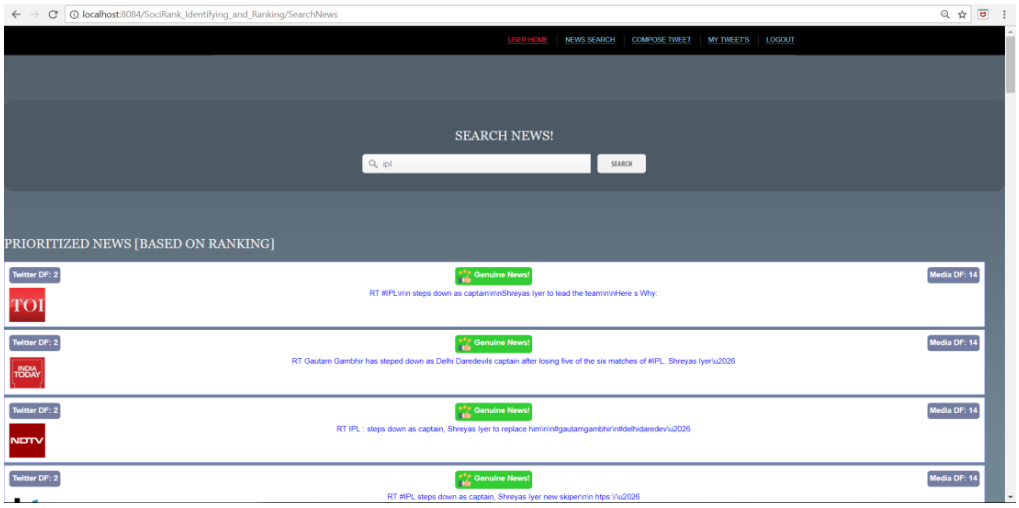

Figure 6: Snapshot of Searched and Prioritized News.

Based on the keyterm given by the user, after comparison the results will be shown in a prioritized order as shown in figure 6.

V. CONCLUSION AND FUTURE ENHANCEMENTS

The system mainly focuses on displaying the prioritized news based on the terms searched by the user. The user can either search the news or compose tweets into their twitter account directly. There is also an option where the user can view all the tweets that he has posted through our system. The searched news are given in an prioritized order. Once the keyterm is given by the user the system will display whether the news related to the keyterm is genuine or not. In future, tweets tweeted through twitter shall also be considered for checking the genuineness of the news. It can also be applied to other social media.

REFERENCES

[1] Derek Davis, Gerardo Figueroa, and Yi-Shin Chen, "*SociRank: Identifying and Ranking Prevalent News Topics Using Social Media Factors*", IEEE Transactions on Systems, MAN, and Cybernetics: Systems.

[2] G. Figueroa and Yi-Shin Chen. "*Collaborative Ranking Between Supervised and Unsupervised Approaches for Keyphrase Extraction.*" in proceedings of the 26th Conference on Computational Linguistics and Speech Processing, 2014.

[3] Elizabeth Kwan, Pei-Ling Hsu, Jheng-He Liang, and Yi-Shin Chen*, "*Event Identification for Social Streams Using Keyword-Based Evolving Graph Sequences*", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining,2013.

[4] K. Shubhankar, A. P. Singh, and V. Pudi, "*An Efficient Algorithm for Topic Ranking and Modeling Topic Evolution*", in Database Expert System Application, Toulouse, France, pp. 320–330, 2011.

[5] W. X. Zhao, Jing JIANG, Jianshu Weng, Jing He, Ee Peng LIM, "*Comparing Twitter and Traditional Media Using Topic Models*", in Advances in Information Retrieval. Heidelberg, Germany: Springer Berlin Heidelberg, 2011.

[6] M. Cataldi, L. Di Caro, and C. Schifanella, "*Emerging Topic Detection on Twitter Based on Temporal and Social Terms Evaluation*", in Proceedings 10th International Workshop Multimedia Data Mining (MDMKDD), Washington, DC, USA, 2010.

[7] K. Sarkar, M. Nasipuri, and S. Ghose, "*A New Approach to Keyphrase Extraction using Neural Networks*", International Journal Computer Science Issues, vol. 7, no. 3, pp. 16–25, Mar. 2010.

[8] K. Kireyev, "*Semantic-Based Estimation of Term Informativeness*", in Proc. Human Language Technologies Annual Conference North America Chapter Association Computer Linguist., pp. 530–538, 2009.

[9] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman and J. Sperling. "*Twitterstand: News in Tweets*", in Proceedings of the 17th ACM Sigspatial International Conference on Advances in Geographic Information Systems, pp. 42-51, ACM, 2009.

[10] Owen Phelan, Kevin McCarthy, and Barry Smyth. "*Using Twitter to Recommend Real-Time Topical News*" Proceedings of the third ACM Conference on Recommender Systems. ACM, 2009.

[11] 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, pp. 633–641, 2006.

