International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:9, No:4, 2015

Online Forums Hotspot Detection and Analysis Using Aging Theory

K. Nirmala Devi, V. Murali Bhaskaran

Abstract**—**The exponential growth of social media arouses much attention on public opinion information. The online forums, blogs, micro blogs are proving to be extremely valuable resources and are having bulk volume of information. However, most of the social media data is unstructured and semi structured form. So that it is more difficult to decipher automatically. Therefore, it is very much essential to understand and analyze those data for making a right decision. The online forums hotspot detection is a promising research field in the web mining and it guides to motivate the user to take right decision in right time. The proposed system consist of a novel approach to detect a hotspot forum for any given time period. It uses aging theory to find the hot terms and E-K-means for detecting the hotspot forum. Experimental results demonstrate that the proposed approach outperforms k-means for detecting the hotspot forums with the improved accuracy.

Keywords**—**Hotspot forums, Micro blog, Blog, Sentiment Analysis, Opinion Mining, Social media**,** Twitter, Web mining.

I. Introduction

SENTIMENT analysis is one of the emerging research area within data mining and Natural Language Processing (NLP), which automatically extracts, classifies and understands the opinion generated by users. The existing information resources are easily integrated with the sentiment analysis techniques for enhancing the information. Rapid growth of web and information age arouses much attention on public opinion and most of these are in unstructured and semi structured format. It is difficult to decipher automatically as well as it is very much difficult for the customers to acquire information that are useful to them. This has motivated on the online forums hotspot detection, where the useful information are quickly made available to the customers which might make them benefit in decision making process.

Nowadays, the tremendous growth of information is available in social media sites such a twitter [1], forums [2], [13] face book, blogs and news reports etc., are having large volume of public opinion information. It is essential to extract sentiment information from these social networks for making predictions in time and understand the trends of the opinion correctly. Exploiting social media sentiment textual information in addition to numeric historical time series stock data increases the prediction accuracy with quality of the data. The social media are platforms of open, honest and real sharing of news to clarify investment behavior in the stock market.

Although timely access to information [3] is becoming increasingly important in today's knowledge-based economy, gaining such access is no longer a problem because of the widespread availability of broadband in both homes and business. However, traditional topic detection methods are not much effective for detecting the hotspot forums application. Therefore, a specific approach is needed to detect hotspot in specific time slot.

The primary aim of this research paper is to investigate the online forums hotspot detection based on enhanced k-means and aging theory. Experimental results show that there is some causal relationship between public sentiment in the forum and hotspot detection. The accuracy of hotspot forum detection can be significantly improved by the incorporation of aging theory for identifying the relevant terms.

The rest of the paper is structured as follows: Section II briefly discusses the review of related works. The details of the proposed system are presented in Section III. The experimental setup and empirical results drawn from the proposed system is compared with the others is discussed in Section IV and Section V concludes the paper.

II. Review of Related Works

The hotspot forum is defined as a forum that appears frequently over a time period and has bulk threads as well as posts. The hotness of a forum depends upon how often it covers hot terms and many discussions that contains those terms. Moreover, no forum can remain hotspot indefinitely, (ie) for any hot topic or emerging area goes through a lifecycle of birth, growth, maturity and death. Since, the hotspot of each forum or topic evolves over a given period of time.

A hotspot forum 'f' has the following characteristics:

- It appears in many discussions
- It has many threads
- It has more comments / posts / replies
- It has many viewers
- It varies in popularity over a time

Mining of online reviews has become a flourishing frontier in today's environment as it can provide a solid basis for predicting future events [4]. Therefore, internet public opinion monitoring and analyzing [5] have become a hot issue in recent years. Although news certainly influences the stock market prediction and play a significant role in financial decisions. The sentiments expressed in twitter can predict the box office receipts. It is therefore reasonable assumption that

K.Nirmala Devi is with the Kongu Engineering College, Perundurai, Erode, Tamil Nadu, India (e-mail: sunsys19@yahoo.com).

Dr. V. Murali Bhaskaran is with the Dhirajlal Gandhi College of Technology, Salem, Tamil Nadu, India (e-mail: murali66@gmail.com).

the public sentiments in social media can predict the stock price directional movements as up/down.

Yet, major works done with sentiment analysis mainly focuses on the product review, movie review and blogs [6]. On the other hand, the lexicon developed for one domain misclassifies information in another domain. Since, they are domain dependent.

A term weighting approach plays major role to identify significant or representative information from large collection of documents. There are many ways to evaluate the significance of terms present in the documents. TFIDF is a most common method for identifying the representative terms. It is not suitable for finding the hot topic because it treats all the words equally and linearly follow the distribution. In another approach TFPDF assigns higher weights for frequent words those appearing in multiple documents. Even though it identifies the hot term efficiently but it suffers to omit the popularity of the term.

Identifying variations in the distribution of key terms over a time period is an important task for hotspot detection. Therefore, it is very much essential to identify the life cycle stage of a key term for detecting the hotspot. Key terms have the features of representativeness, conciseness, timelines, high degree of association, mass information and are helpful for detecting the hotspot of the forums. The keywords are used in blogs, news, internet public opinion, web applications to detect the hotness in their respective sources.

In [7] Back Propagation Neural network based classification algorithm was used to evaluate the hotness of the topic through its popularity, quality and message distribution. In another work [8] semi automatic approach was proposed to find the hot events. In [9] hot terms were extracted based on the TFPDF with lifecycle to detect the hot topics. The hot topic [10] detection during a time interval was detected based on burstiness. The statistics and correlation of popular words in network traffic content to extract popular topics on the internet. The aging theory was used in detecting the hot topic in BBS [11] also.

The above discussed approaches could analyze text opinion and detect hot topic efficiently. However, online forum has some specific structures and those approaches were not completely suitable for it. The proposed system consists of a novel approach to find the hotness of the key terms for online forums hotspot detection.

III. The Proposed System

The tremendous growth of social media arouses much attention on public opinion and it provides a lot of opportunities to analyze as well as for predicting the hotspot. Moreover, hot topic detection is flourishing more and more popular, many approaches were proposed to predict the popularity. Although the conventional approaches used in information retrieval such as TFIDF and TFPDF alone might not give good results, since the document features are sparse in forum data. Therefore, in this paper a novel approach is used to identify the key terms based on aging theory for hotspot forum detection. The enhanced k-means algorithm is also used for improving the hotspot detection accuracy. The schematic process of proposed system is shown in Fig. 1.

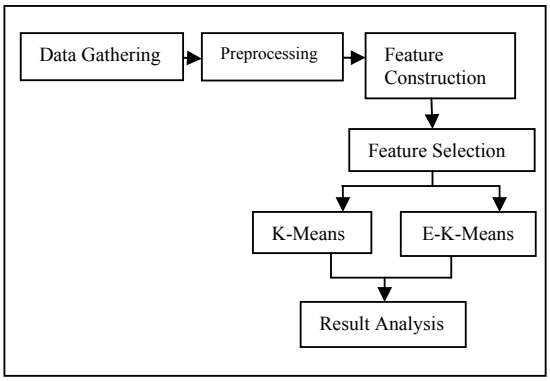

Fig. 1 The Proposed System Process

A. Data Gathering

The data set used in this experimental research is acquired from forums.digitalpoint.com [14] and after data cleaning they are formatted to 39 different forums and 1933 threads. The data collection is initiated by crawling all the URL links of 50 forums and its links are stored in the data base. Then all the topic posts and the comment posts contained in the corresponding web pages and their links are parsed and they are stored in the data base.

B. Preprocessing

The raw data collected from the forum is preprocessed by eliminating the noise data, non-opinion information, removal of stop words and stemming. Noise data include forums with picture postings that are not clearly shown online.

Irrelevant data are from forums where the posting contents are not related to the forum threads at all. The threads that have no replies and the forums that have no threads across the time window are also removed. Finally after cleaning, 39 forums are narrowed down within the time span from January to December and each time window is a half month length (i.e., Fifteen days duration) over the year 2011. The data before cleaning and after cleaning are listed in Table I.

TABLE I
DATA VIEW BEFORE CLEANING AND AFTER CLEANING

Particulars	Before Cleaning	After Cleaning
No. of forums	50	39
No. of threads	2,916	1,933
No. of replies	39,239	21,245
No. of views	84,321	53,218

C. Feature Construction

The preprocessing work is followed by feature extraction. There are four kinds of features are constructed from the forum data for hotspot detection is as follows in Table II.

1. Historic based features (F-His)
2. Statistical based features (F-Sta)
3. Semantic based features (F-Sem)
4. Sentiment based features (F-Sen)

TABLE II
FEATURE TYPES

Historic based	Statistic based	Semantic based	Sentiment based
No. of threads	TFIDF	TSFISF	No. of positive replies
No. of replies	TFPDF		No. of negative replies
No. of views			No. of opinionated replies
Avg. threads			Avg. opinionated replies
Avg. replies			Proportion of positive replies
Avg. views			Proportion of negative replies

D. Feature Selection

The feature selection uses the life time support of the key terms for detecting the hotspot forums. After constructing the features to identify the most significant features among TFIDF, TFPDF and TSFISF the aging theory is used to select for hotspot detection. Simply considering the term frequency is not enough for detecting the hotspot forum. Thus, the variant usage of the term is also taking into account with time analysis.

1. Aging Theory Frame Work

Chen [12] is the first person who models the topic using aging theory. The life cycle of a topic is divided in to four parts namely birth, growth, decay and death. In order to track the life cycle of topic or event, it uses energy function. Whenever there is an increase in energy that shows the popularity of the topic during that time period and diminishes with the time. The relevance of the topic during a specific time slot shows the nutrient and that can be converted to energy.

The timeline is divided into an equal width of intervals. The proposed system time span from January to December and each time slot is a half month length (i.e., Fifteen days duration) over the year 2011. There are twenty four time slots namely S1, S2…S24 and each time slot Si consists of five time intervals. For example in time slot S1 there are I11, I12, I13, I14, I15 intervals. Each interval denotes three consecutive days. The proposed system mainly uses the following four functions for finding the energy of a term in the time slot.

- getEnergy() – compute the nutrition of term in the time interval. The energy accumulated from all intervals of a specific time slot 'Si' for term 't'.

$$E_{t,Si} = tfreq(t) * \sum_{j=1}^{n} B_j(t) \quad (1)$$

$$B_j(t) = \frac{(A+B+C+D)*(AD-BC)^2}{(A+B)(C+D)(A+C)(B+D)} \quad (2)$$

$tfreq(t)$ – statistic based / semantic based features (TFIDF, TFPDF, TSFISF) i.e., $tfreq(t)$ = Wt1 / Wt2 / Wt3 to denote TFIDF / TFPDF / TSFISF respectively.

A. Frequency of term't' in time interval I
B. Frequency of term't' in time other intervals
C. No. of intervals in current time slot Si do not have term't'
D. No. of intervals in other time slots do not have term 't'

- energyFun() – converts a term's nutrition value into an energy value. energyFunction^{-1}() coverts an energy value into a term's nutrition value.

$$lifesup_{t,Si} = \ln(E_{t,Si}) \quad (3)$$

- energyDec() – decreases the energy of term in each time slot. If the decayed life support value is negative, then it is set zero.

For each time slot Si

$$lifesup_{t,Si}' = lifesup_{t,Si} - \delta \quad (4)$$

δ - is the decay nutrition factor and it is an empirical constant.

- getVar() – computes the variance of the life support value of term .

$$Var_t = \sqrt{\frac{1}{N}\sum(lifesup_{t,Si} - \overline{lifesup})^2} \quad (5)$$

N – number of intervals in given time slot Si; $\overline{lifesup}$ - average life support of term 't'

2. Hot Term Weight Computation**:**

In order to find the hot forum the existing term frequency is not only sufficient to determine. Since, a hot term is typically should exist in many intervals of given time slot and its frequency of usage vary over the time slot. Therefore, the proposed system extracts the hot terms based on frequency, position, topicality and pervasiveness.

Consider a forum text stream arriving in sequence of time slots S1,S2,…Sk from different intervals I11,I12,…I1i, i.e., the set of documents D= { D11, D12, … D1k, D22, … D2k, …., Di1, Di2, … Dik }. Where 'i' denotes numbers of intervals and 'k' denotes number of time slots. The process of computing hot term weighting is shown in the following algorithm in Fig. 2.

E. Hotspot Detection Using K-means Clustering:

The K-means clustering is an efficient and a simple algorithm, it has a wide applications. Although, it has many advantages of being easy to implement, efficient, it has some disadvantages also. The effectiveness of the K-means clustering is depending up on the initial centroids selection. The accuracy of this algorithm may relatively poor due to the random selection of the initial centroids. Since, the quality of the final clusters have mostly depending up on the selection of the initial centroids and we get different cluster centroids for different runs for same data objects. The flow of the K-means algorithm is shown in Fig. 3.

After identifying significance hot terms, clustering can be carried out using K-means algorithm. Each forum may be represented as a data point in a vector space. A vector is used to represent any forum and it is composed of four kinds of elements from Table II. These datasets are given as the input to the k-means clustering where a clustered view of all the forums is obtained. The hotspot and non-hotspot forums

obtained, within each time window are those closest to the theoretical centers of clusters.

1. For each term 't' in each time interval Iij in the time slot Si
2. Compute TFIDF, TFPDF, TSFISF weight as Wt1, Wt2, Wt3 respectively.
3. Compute position weight as PWt with the following constraints
   If position of 't' is thread then PWt = 1
   Else PWt=0.5
4. Compute aging theory based energy, life support , energy decay and variance for the term 't' as follows
   $E_{t,Si}$, $lifesup_{t,Si}$, $lifesup_{t,Si}'$, Var_t
5. Loop
6. Compute R1 = Rank of term weight Wt1 / Wt2 / Wt3
7. Compute R2 = Rank of Var_t in Wt1 / Wt2 / Wt3
8. Compute new weight for 't' using following

$$New\ Wt1 = \alpha 1 Wt1 + \alpha 2 PWt + \alpha 3\, Var_t \left(1 + \frac{R1 - R2}{T}\right)$$

$$New\ Wt2 = \alpha 1 Wt2 + \alpha 2 PWt + \alpha 3\, Var_t \left(1 + \frac{R1 - R2}{T}\right)$$

$$New\ Wt3 = \alpha 1 Wt3 + \alpha 2 PWt + \alpha 3\, Var_t \left(1 + \frac{R1 - R2}{T}\right)$$

   R1 to be higher value gets top rank and at the same time lower value of R2 gets top rank for the term't'.

9. End

Fig. 2 Hot Term Computation Algorithm

1. Select K data objects randomly from data set as initial centroids
2. Repeat
   - Assign each data object xi to the closet cluster which has the closet centroid i.e., to minimize the squared error function (SE)
     $SE = \sum_{j=1}^{K} \sum_{i=1}^{n} \left\| x_i^{(j)} - C_j \right\|^2$ and this shows the intra cluster similarity.
   - Compute the new center for each cluster

   Until convergence is met

Fig. 3 K-means Algorithm

1. For each column of the data object, find the Low Value (LV) and High Value (HV) element.
2. Compute range = HV – LV
3. Select the column having maximum range value
4. Sort the data objects according to the selected column in the step 3.
5. Divide the data objects in to 'K' parts as equal size as C_1, C_2,..C_K
6. Compute the median value as the initial centroids
   $C_i = \frac{HV_i - LV_i}{2}$
7. Repeat
   - Assign each data object xi to the closet cluster which has the closet centroid i.e., to minimize the squared error function (SE)
     $SE = \sum_{j=1}^{K} \sum_{i=1}^{n} \left\| x_i^{(j)} - C_j \right\|^2$ and this shows the intra cluster similarity.
   - Compute the new center for each cluster

   Until convergence is met

Fig. 4 E-K-means Algorithm

F. Hotspot Detection Using E-K-means Clustering:

In this section, the proposed initial centriod assignment can improves the performance of K-means algorithm. The flow of the E-K-means algorithm is shown in Fig. 4.

IV. Experimental Results

The experimental results obtained for hotspot forum with the different term weighting is shown in the Table III. The performance measures used for evaluating the proposed system are Accuracy, Precision and Recall is defined as follows in (6)-(8):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$Precision(P) = \frac{TP}{TP+FP} \quad (7)$$

$$Recall(R) = \frac{TN}{TP+FN} \quad (8)$$

where, TP denotes the number of forums that are estimated as hotspots. TN denotes the number of forums that are estimated as non-hotspots .FP denotes the number of forums that are estimated as hotspots but they are non-hotspots. FN denotes the number of forums that are estimated as non-hotspots but they are hotspots.

TABLE III
The Result of Accuracy % for K-means and E-K-means

Measures	K-means	E-K-means
Historic based Features (F-His)	64.1	69.01
Statistics based Features (F-Sta)	67.04	71.81
Semantic based features (F-Sem)	70.28	73.49
Sentiment based features (F-Sen)	72.12	75.46
All(F-His, F-Sta, F-Sem & F-Sen)	74.80	80.88

Table IV represents the precision and recall measures for K-means and E-K-means algorithm during time slots S1, S2, ..., S12.

TABLE IV
The Performance Measures of K-means and E-K-means

Time Slot	K-means		E-K-means	
	Precision	Recall	Precision	Recall
S1	75.00	75.00	83.33	83.33
S2	72.73	61.54	81.82	69.23
S3	60.00	54.55	70.00	63.64
S4	55.56	50.00	66.67	60.00
S5	70.00	50.00	80.00	57.14
S6	66.67	53.33	83.33	62.50
S7	72.73	50.00	90.91	58.82
S8	50.00	33.33	62.50	45.45
S9	57.14	28.57	71.43	35.71
S10	70.00	50.00	80.00	57.14
S11	63.64	53.85	72.73	66.67
S12	55.56	41.67	83.33	83.33

Fig. 5 represents the goodness of the proposed E-K-means performance measures with K-means algorithm. The obtained results show that accuracy of hotspot forum detection can be

significantly improved by the incorporation of sentiments and initial centroid assignment.

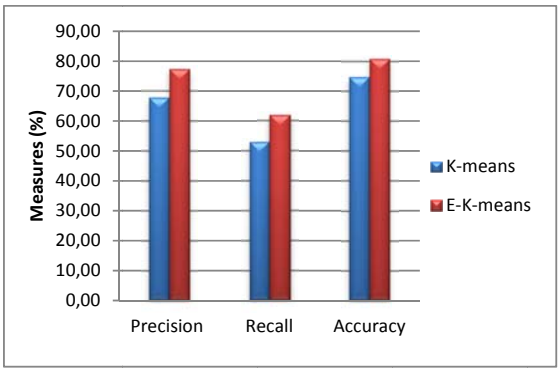

Fig. 5 The Performance measures of K-means and E-K-means

V. Conclusion

The online forums hotspot detection is a promising research field in the web mining and it guides to motivate the user to take right decision in right time. The proposed system consist of a novel approach to detect a hotspot forum for any given time period. It uses aging theory to find the hot terms and E-K-means for detecting the hotspot forum. Tables III and IV suggest that the proposed E-K-means algorithm gives optimized results than that of K-means algorithm. The obtained result shows the improvement of accuracy level of E-K-means from 74.80% to 80.88%. When comparing the accuracy of E-K-means with K-means there is a 6.07% improvement is obtained. The same way when comparing the precision and recall result of K-means with E-K-means there is a 9.54% and 8.96% improvement is obtained in the proposed system. This shows proposed E-K-means approach outperforms k-means for detecting the hotspot forums with the improved accuracy.

References

[1] Johan Bollen, Huina Mao, and Xiao-Jun Zeng, "Twitter mood predicts the stock market*", IEEEComputer*, vol. 44 no.10,pp. 91–94, 2011.

[2] Nirmala Devi, K., Murali Bhaskaran, V.,2012, "A Semantic Enhanced Approach for Online Hotspot Forums Detection", *Proceedings of IEEE Conference - ICRTIT 2012*,pp. 497-501, 2012.

[3] Chen, K., Luesukprasert, L. and Chou, S., "Hot topic extraction based on timeline analysisand multidimensional sentence modeling", *IEEE Transactions on Knowledge and Data Engineering*, vol.19, no.8, pp.1016–1025, 2007.

[4] Peng, W., "Predicting collective sentiment dynamics from time-series social media", *Proceedings of the Conference WISDOM '12*, 2012.

[5] Liu, H., "Internet public opinion hotspot detection and analysis based on Kmeans and SVMalgorithm", *Proceedings of the International Conference of Information Science andManagement Engineering – ISME-2010*, pp.257–261, 2010.

[6] Hu M. and Liu B, "Mining and summarizing customer review", *Proceedings of ACM Transactions on Knowledge and Data Engineering*, pp.168-177, 2004.

[7] Lan You, Yongping Du, Jiayia Ge, Xuanjing Huang and Lide Wu, "BBS based hot topic retrieval using back propagation neural network", *Proceedings of IJCNLP 2004*, Springer – Verlag, pp. 139-148, 2005.

[8] Tingting He, Guozhong Qu, Siwei Li and et.al., "Semi-automatic hot event detection", *Proceedings of the 2nd International Conference on advanced Data Engineering and Applications*, pp.1008-1016, 2006.

[9] Zhang, Z. and Li, Q., "QuestionHolic: hot topic discovery and trend analysis in community question answering systems", *Expert Systems with Applications*, Vol. 38, No. 6, pp.6949–6855, 2011.

[10] Platakis. M., Kotasakos,D and Gunopulos. D, " Discovering Hot topics in the Blogosphere", Proceedings of the 2nd panhellenic Scientific Student Conference on Informatics, Related Technologies and Applications EUREKA 2008, pp. 122-132.

[11] Zheng, D. and Li, F., "Hot topic detection on BBS using aging theory", *Proceedings of the WISM 2009*, LNCS 5854, pp.129–138, 2009.

[12] Chen,C.C., Chen, Y.T., Sun, Y., Chen, M.C., "Life Cycle Modeling of News Events using Aging Theory", LNCS(LNAI), vol.2837, pp.47-59, Springer, Heidelberg, 2003.

[13] Nirmala Devi,K. and Murali Bhaskaran, V. ," Forecasting Indian Stock Market Using Particle Swarm Optimization and Support Vector Machine", *International Journal of Applied Engineering Research*, vol.10, no.1, pp.1891-1900, 2015.

[14] http:// forums.digitalpoint.com