Journal of Computer Science 8 (8): 1219-1224, 2012
ISSN 1549-3636

Online Forums Hotspot Prediction Based on Sentiment Analysis

[1]K. Nirmala Devi and [2]V. Murali Bhaskarn
[1]Department of CSE, Kongu Engineering College, Perundurai, Erode, Tamil Nadu,India
[2]Department of CSE, Paavai College of Engineering, Pachal, Namakkal, Tamil Nadu, India

Abstract: Problem statement: Online forums hotspot prediction is one of the significant research areas in web mining, which can help people make proper decision in daily life. Online forums, news reports and blogs, are containing large volume of public opinion information. Rapid growth of network arouses much attention on public opinion, it is important to analyse the public opinion in time and understands the trends of their opinion correctly. **Approach:** The sentiment analysis and text mining are important key elements for forecasting the hotspots in online forums. Most of the traditional text mining work on static data sets, while the online hotspot forecasts works on the web information dynamically and timely. The earlier work on text information processing focuses in the factual domain rather than opinion domain. Due to the semi structured or unstructured characteristics of online public opinion, we introduce traditional Vector Space Model (VSM) to express them and then use K-means to perform hotspot detection, then we use J48 classifier to perform hotspot forecast. **Results:** The experimentation is conducted by Rapid Miner tool and performance of proposed method J48 is compared with other method, such as Naive Bayes. The consistency between K-means and J48 is validated using three metrics. They are accuracy, sensitivity and specificity. **Conclusion:** The experiment helps to identify that K-means and J48 together to predict forums hotspot. The results that have been obtained using J48 present a noticeable consistency with the results achieved by K-means clustering.

Key words: Hotspot, J48, K-means, sentiment analysis, text mining, Vector Space Model (VSM), noticeable consistency, hotspot detection, sentiment analysis, consistency between

INTRODUCTION

Opinion mining is an important sub discipline within data mining and Natural Language Processing (NLP), which automatically extracts, classifies and understands the opinion generated by various users. These techniques also help to enhance the value of existing information resources that can be integrated with new products and systems as they are brought on-line.

Rapid progress of the web and information age, online data grows too fast in an exponential form. The most of the online data is semi structured or unstructured format and that is difficult to decipher automatically. The growth of large volume of heterogeneous online information from various forums has made very difficult for the customers to acquire information that are useful to them. Therefore, online forums hotspot prediction has becoming promising research field in web mining.

An automation of online forums hotspot prediction can be beneficial to customers in many ways. For instance, the company could collect comments to their new products or the marketing department understands the timely requirements of the customers regarding products and services. So, this has motivated on the detection of hotspot as well as prediction of hotspot forums (Li and Wu 2010) where useful information are made available quickly for those customers which might make them benefit in decision making process.

The large volumes of online data are efficiently processed with help of statistical (Thongwan *et al*., 2011) and machine learning techniques. An emergent technique called Emotional polarity computation also known as sentiment analysis (Li *et al*., 2010) can also be performed during online text mining. The purpose of text sentiments is determining the attitude of a speaker or person with respect to some specific topic. However, in opinion classification, topic-related words are not very important. But, opinion words that indicate positive or negative opinions are important, e.g., great, excellent, amazing, horrible, bad, worst. Most of the methodologies for opinion mining apply some forms of machine learning techniques for classification.

Customized-algorithms specifically for opinion classification have also been developed, which exploit opinion words and phrases together with some scoring functions. In this study we detect the hotspot forums by computing text sentiment analysis. This method quantifies the user attention on any forum with which hotspot forums can be identified. The proposed work uses an integration approach of text mining with sentiment analysis.

Corresponding Author: K. Nirmala Devi Department of CSE, Kongu Engineering College, Perundurai, Erode, Tamil Nadu, India

Literature review: The various streams of related work are review mining, sentiment classification, predicting hotspots using machine learning techniques.

Review Mining: Mining of online reviews has become a flourishing frontier in today's environment as it can provide a solid basis for predicting future events. The online reviews (Chaovalit and Zhou, 2005) became more useful and influence the sales as it provides important information about the product to potential consumers.

A multi-knowledge based approach is proposed where WordNet, statistical analysis and movie knowledge are integrated. The experimental results have shown the effectiveness of the approach in movie review mining and summarizing.

A generated and semantic orientation labelled list (Hu and Liu, 2004) containing only adjectives are used for analysing. Finally it is observed that machine learning is used to depict the interacting structure of reviews.

Sentiment classification: The documents available on the web can be classified based on various metrics including topics, authors, structures and so forth. Classification based on sentiments has become a new frontier to text mining community.

The task of sentiment classification is to determine the semantic orientations of words, sentences or documents. Most of the early work on this topic used words as the processing unit. An automatic sentiment classification at document level has been done (Pang *et al*., 2002) in which several machine learning approaches are used with common text features to classify movie reviews from IMDB. It has been pointed out that direct marketing is a promotion process which has motivated customers to place orders through various channels (Sindhwani and Mellville, 2008).

In order to work for this, one is needed to have an accurate customer segmentation based on a good understanding of the customers, so that relevant product information can be delivered to different customer segments. Analysing Twitter (Thelwall *et al*., 2011) has given insights into why certain events resonate with the people. It is found that the customers, who are used to having only a limited range of product choices due to physical and/or time constraints, are now facing the problem of information overload.

An effective way of increasing customer satisfaction and consequently customer loyalty has been done that has helped the customers identify products according to their interests. This again has called for the provision of personalized product recommendations (Popescu and Etzioni, 2005; Thelwall *et al*., 2010).

The Latent Class Model (LCM) to circumvent (Hofmann and Puzicha, 1999) the aforementioned problems. Incorporating sentiment information (Paltoglou and Thelwall, 2010) into Vector Space Model (VSM) values using supervised methods was helpful for sentiment analysis.

Predicting hotspots using machine learning techniques: For predicting online hotspot forums two machine learning techniques (Li and Wu, 2010) have been used. It includes K-means and SVM. Unlike other learning methods, SVM's (Preethi *et al*., 2012) performance is related not to the number of features in the system, but to the margin with which it separates the data.

MATERIALS AND METHODS

The proposed work helps in predicting hotspot forums and achieves highly consistent results by applying an efficient optimization algorithm with J48. The proposed work comprises five modules such as data pre-processing, feature extraction, sentiment computation, forum clustering and forum classification. Figure 1 depicts the conceptual diagram of proposed approach.

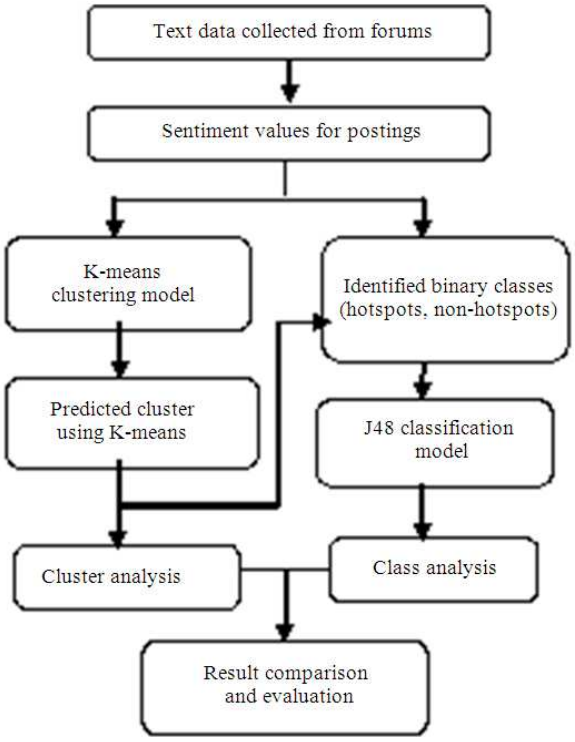

Fig.1: Conceptual diagram of the proposed approach

Table 1: Data view before cleaning and after cleaning

	Before cleaning	After cleaning
Time period	2007 Jan to 2011 Dec	2011 Jan to 2011 Dec
Number of forums	50	39
Number of threads	2916	1933
Number of replies	39239	21245

Pre-processing: The data set used in our experimental research is acquired from forums.digitalpoint.com 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. After crawling process is achieved data cleaning is done where noise data and irrelevant data are removed. 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 1.

Feature extraction: The pre-processing work is followed by feature extraction process. For each forum five features are extracted across each time window such as the number of threads, the average number of replies of threads, the average sentiment value of threads, the fraction of positive threads among all the threads and the fraction of negative threads among all the threads. Sentiment value for each thread can be calculated by computing text sentiment.

Sentiment computation on forum text: Feature extraction includes text sentiment analysis which aims at calculating an integer value for each piece of text. It is a semantic orientation based approach where the sentiment values for all keywords are added to achieve the sentiment value for the whole article.

The replies of thread are decomposed into a set of keywords. For each keyword a sentiment value is assigned. The sum of the sentiment values for all the keywords will give the sentiment value for the thread. Suppose for a thread t, its replies are decomposed into a set of key words. For each key word wi (i=1, 2,...,n) let the sentiment value be si. Then the sentiment value St of the thread t can be calculated as using Eq. 1:

$$st = \sum_{i=1}^{n} si \quad (1)$$

Calculation of sentiment value is based on SentiStrength. SentiStrength is an algorithm for text sentiment analysis that helps in estimating the sentiment values for texts.

Forum clustering using K-means: After the features are extracted clustering can be carried out using K-means algorithm. Each forum may be represented as a data point in a vector space. During the feature extraction process a vector is used to represent the emotional polarity of any forum and it is composed of five elements: the number of threads, the average number of replies of threads, the average sentiment value of threads, the fraction of positive threads among all the threads and the fraction of negative threads among all the threads. These data 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 being obtained, within each time window are those closest to the theoretical centres of clusters.

Forum classification using J48: Classification can be carried out using J48 (decision tree) classification algorithm. It is a predictive machine learning model that decides the target value of a new sample based on various attributes of available data. J48 is employed to realize hotspot forecasting. In order to forecast the hotspot forums within the current time window the clustering result obtained by K-means approach from the previous time window is used.

It performs forum classification iteratively and tries to find the optimized solution. For each J48, the input is a forum's representation vector and the optimized output is achieved by classifying each forum as either hotspot forum or non-hotspot forum. The accuracy in predicting hotspot forums is improved with the proposed model and the consistency of the model is validated for its performance.

RESULTS

The data that we have collected from the forum consists of a list of posts in the form of threads and replies have been crawled from January 2007 to December 2011. The data view before and after cleaning is depicted in Table 1. After cleaning the data are narrowed to 39 forums from January 2011 to December 2011 and then the features are extracted that includes computing sentiment values for threads.

The feature extraction is then followed by K-means clustering and classification using J48 among the 39 leaf forums for each time window in 2011. The results that have been obtained using J48 present a noticeable consistency with the results achieved by K-means clustering.

The forums that are most popular among the users based on average number of threads include 'Search Marketing, Publisher Network, adcenter, General Marketing. The forums that are popular based on average number of replies include 'Affiliate Programs-Google, Affiliate Network, Payments, Google-Google+'. The forums that are mostly identified as hotspots by both K-means clustering and J48 over the time window from Jan 2011 to Dec 2011 are shown in Table 2.

Performance evaluation: The consistency between K-means and J48 algorithms is validated using three metrics. They are Accuracy, Sensitivity and Specificity. A set of these three metrics are applied for each time window which are defined as follows Eq. 2-4:

$$\text{Accuracy} = \text{TP} + \text{TN} / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (2)$$

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) \quad (3)$$

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}) \quad (4)$$

where, TP denotes the number of forums that are estimated as hotspots by both K-means and J48.

TN denotes the number of forums that are estimated as non-hotspots by both K-means and J48.

FP denotes the number of forums that are estimated as hotspots by J48 whereas non-hotspots by K-means.

FN denotes the number of forums that are estimated as non-hotspots by J48 whereas hotspots by K-means. Using Eq. 2-4, the performance is evaluated for each time window. The time windows are those that are used in J48 classification process.

Table 2: Forums mostly identified as hotspot by K-means and J48

Forum ID	Forum name
11	Affiliate Network
33	Twitter
44	Google
47	Amazon
10	Google+
34	Social Network Google+
16	Publisher Network
7	Payments
6	Reporting and Stats
49	ClickBank

DISCUSSION

Table 3 shows the accuracy (%) in each time window for different K values from K=2 to K=7 while using J48 classification algorithm. It is clearly indicated that the method helps in achieving a satisfying result of accuracy especially when K has reached certain value.

Table 4 suggests that the proposed J48 classification algorithm gives an optimized accuracy result than that of the Naive Bayes classification algorithm. The average accuracy (%) obtained for different K values is shown in the Table 4.

Similarly the performance can be evaluated using other two metrics and the results can be compared. The sensitivity shows the fraction of forums which are classified by classification algorithm as hotspots among all forums that are labelled by K-means as hotspot. The average sensitivity values obtained for different K values using J48 and Naive Bayes classification is shown in the Table 5.

The next important measurement is the specificity that shows the fraction of forums which are classified by classification algorithm as non-hotspot among all forums that are labelled by K-means as non-hotspots. The sensitivity result is shown in the Table 6.

The accuracy (%) obtained for different K values while using J48 and Naive Bayes is shown in the Fig. 2.

The accuracy (%) obtained in each time for both the classification algorithm J48 and Naive Bayes is shown in the Table 7.

Table 3: Accuracy (%) in each time window while using J48 algorithm

	Accuracy (%) in each time window for different K values					
Time window	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7
2	86.0	87.0	90.0	90.2	89.6	88.2
3	83.0	85.5	88.0	89.0	87.0	86.0
4	86.0	87.2	90.0	91.0	92.0	88.0
5	91.0	91.0	93.0	93.6	94.0	92.0
6	81.0	81.0	84.0	86.0	87.1	81.3
7	90.0	91.0	91.0	92.0	93.0	91.0
8	83.0	83.5	84.9	85.0	87.0	84.0
9	91.0	92.0	92.3	93.0	93.4	92.0
10	79.0	79.0	82.0	81.0	82.0	81.0
11	84.0	85.0	85.0	85.6	86.0	85.0
12	82.0	83.0	86.0	88.0	88.0	84.8
13	82.0	82.0	84.0	84.0	85.0	84.0
14	86.0	89.0	90.0	91.2	92.0	89.4
15	80.0	81.0	81.0	83.0	83.0	81.0
16	86.5	87.0	90.0	90.0	90.0	88.0
17	89.0	91.0	95.0	95.0	96.0	92.0
18	83.0	83.0	84.7	86.0	86.0	84.0
19	80.0	84.0	86.0	88.0	90.0	86.0
20	88.0	90.0	91.0	92.0	93.0	91.0
21	90.0	91.0	93.0	94.0	94.0	93.0
22	91.0	91.6	94.0	94.0	94.0	92.0
23	90.0	91.0	92.0	93.0	93.0	91.0
24	91.0	93.0	94.0	93.5	96.0	93.0

Table 4: Average accuracy (%) for J48 and naive bayes algorithm

	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7
Naive bayes	84.60	85.51	86.16	86.98	87.89	88.96
J48	85.76	86.90	87.73	88.73	89.48	90.05

Table 5: Average sensitivity for j48 and naive bayes algorithm

	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7
Naive bayes	0.17	0.32	0.38	0.39	0.58	0.43
J48	0.19	0.37	0.44	0.41	0.62	0.51

Table 6: Average specificity for j48 and naive bayes algorithm

	K = 2	K = 3	K = 4	K = 5	K = 6	K = 7
Naive bayes	0.82	0.85	0.86	0.87	0.77	0.84
J48	0.91	0.89	0.88	0.88	0.83	0.87

Table 7: Comparison of accuracy using naive bayes with J48

Accuracy (%)		
Time window	Naive bayes	Decision tree
2	88.08	88.50
3	85.95	86.42
4	87.82	89.03
5	91.53	92.43
6	81.42	83.40
7	89.23	91.33
8	82.82	84.57
9	91.38	92.28
10	79.80	80.67
11	82.57	85.10
12	85.32	85.30
13	83.28	83.50
14	90.67	89.60
15	87.17	81.50
16	84.50	88.58
17	82.33	93.00
18	84.13	84.45
19	91.58	85.67
20	90.20	90.83
21	87.80	92.50
22	84.08	92.77
23	91.63	91.67
24	90.33	93.42

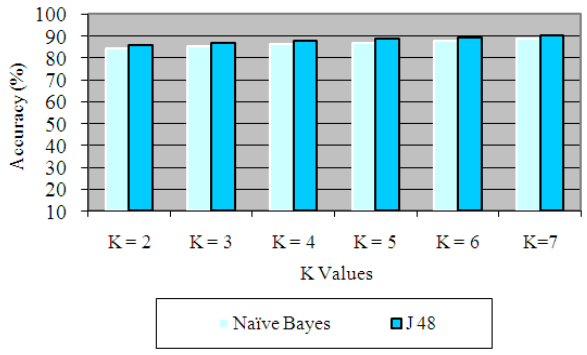

Fig. 2: Accuracy comparison while using J48 and naive bayes algorithm

CONCLUSION

This study proposes a new approach for predicting hotspot forums. In this approach emotional polarity of the text is obtained by computing a value for each part of text. After calculating the sentiment values the method is then integrated with K-means clustering and J48 classification algorithm as well as Naive Bayes classification algorithm for forums hotspot prediction. Computation indicates both K-means and J48 produce consistent grouping results.

The new method that is proposed to helps to achieve a satisfying result for accuracy when K has reached a certain value. The average accuracy of about 88.11% has been obtained for predicting hotspot forums across 20 time window during J48 classification and it is more than that of 86.88% during Naive Bayes classification. Thus the efficient detection of hotspot forums based on sentiment analysis might make internet social network members benefit in the decision making process.

REFERENCES

Chaovalit, P. and L. Zhou, 2005. Movie review mining: A comparison between supervised and unsupervised classification approaches. Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Jan. 03-06, IEEE Xplore Press, pp: 112c-112c. DOI: 10.1109/HICSS.2005.445

Hofmann, T. and J. Puzicha, 1999. Latent class models for collaborative filtering. Proceedings of the 16th International Joint Conference on Artificial Intelligence (AI' 99), San Francisco, CA, USA, pp: 688-693.

Hu, M. and B. Liu, 2004. Mining and summarizing customer reviews. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Aug. 22-25, ACM Press, Seattle, WA, USA., pp: 168-177. DOI: 10.1145/1014052.1014073

Li, F., M. Huang and X. Zhu, 2010. Sentiment analysis with global topics and local dependency. Proceedings of the 24th AAAI Conference on Artificial Intelligence (AI' 10), Association for the Advancement of Artificial, pp: 1371-1376.

Li, N. and D.D. Wu, 2010. Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision Support Syst., 48: 354-368. DOI: 10.1016/j.dss.2009.09.003

Paltoglou, G. and M. Thelwall, 2010. A study of information retrieval weighting schemes for sentiment analysis. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, (ACL' 10), ACM Press, Stroudsburg, PA, USA., pp: 1386-1395.

Pang, B., L. Lee and S. Vaithyanathan, 2002. Thumbs up?: Sentiment classification using machine learning techniques. Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, (EMNLP, 02), ACM Press, Stroudsburg, PA, USA., pp: 79-86. DOI: 10.3115/1118693.1118704

Popescu, A.M. and O. Etzioni, 2005. Extracting product features and opinions from reviewsProceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, (HLTEMNLP' 05), ACM Press, Stroudsburg, PA, USA., pp: 339-346. DOI: 10.3115/1220575.1220618

Preethi, T., K.N. Devi and V.M. Bhaskaran, 2012. A semantic enhanced approach for online hotspot forums detection. Proceedings of the 2nd International Conference on Recent Trends in Information Technology, Apr. 19-21, IEEE Xplor Press, Chennai, Tamil Nadu, pp: 497-501. DOI: 10.1109/ICRTIT.2012.6206785

Sindhwani, V. and P. Mellville, 2008. Document-word co-regularization for semi-supervised sentiment analysis. Proceeding of the 8th IEEE International Conference on Data Mining, Dec. 15-19, IEEE Xplore Press, Pisa, pp: 1025-1030. DOI: 10.1109/ICDM.2008.113

Thelwall, M., B. Kevan, G. Paltoglou, D. Cai and A. Kappas, 2010. Sentiment strength detection in short informal text. J. Am. Soc. Inform. Sci. Technol., 61: 2544-2558. DOI: 10.1002/asi.21416

Thelwall, M., K. Buckley and G. Paltoglo, 2011. Sentiment in twitter events. J. Am. Soc. Inform. Sci. Technol., 62: 406-418. DOI: 10.1002/asi.21462

Thongwan, T., A. Kangrang and S. Homwuttiwong, 2011. An estimation of rainfall using fuzzy set-genetic algorithms model. Am. J. Eng. Applied Sci., 4: 77-81. DOI: 10.3844/ajeassp.2011.77.81