ANALYZING AND EXTRACTING SOCIALMINING TRENDS THROUGH WEB OPINION DEVELOPMENTS VIA DENSITY BASED CLUSTERING

Jeswin Roy Dcouth, MohanRaj.T

***Abstract*—The advancement of Web technologieslead to a large volume of Web opinions,which is available throughout on social media sites such as twitters,Facebook and LinkedIn. These technologies provide a platform for Internet users around the world to communicate with each other and express their various opinions. Analysis of developing Web opinions is potentially valuable for discovering ongoing topics of interests of the public such as current trends and crime detection. Unlike regular documents, Web opinions are short and sparse text messages with noisy content. Typical document clustering techniques with the goal of clustering all documents applied to Web opinions produce degradable performance. In this paper, we investigated the density-based clustering algorithm and proposed the scalable distance-based clustering technique for Web opinion clustering. This Web opinion clustering technique enables the identification of themes within discussions in web social networks and their development, as well as the interactions ofactive participants. We also developed interactive visualization tools, which make use of the identified topic clusters to display social network development, the network topology similarity between topics, and the similarity values between participants.**

***Index Terms*— Web opinion mining, density-based clustering, information visualization, social media analytics, social network analysis.**

I. INTRODUCTION

The Internet facilitates communication between individuals not restricted to geographical boundaries.For example, users interact with each other in a Web forum when they have a common interest. An online forum could be a virtual platform for expressing personal and communal opinions, comments, experiences, thoughts, and sentiments in discussion threads.There, net users areable to share their personal details to a circle of friends, amplify their voices and sentiment, establish an on-line communication in an exceedingly topic ofinterest, and promote an ideology.

The continual user interaction on a web forum becomes a virtual community for members to share thoughts on subjects of their interest while not face-to-face contact with one another. The messages in an exceedinglynet forum generallydon't have strong factual content. However, the factual content is usually hidden inside user's perspicacity in opinions. Additionally, there are factual connections that replicate the focuses of discussions among the forum members of a thread. Forum members express their opinions just about on all types of topics such as political and social problems, technology, hobby, movies, music, health, sports and religion. For example, to associate extreme, the grey forums within the recent years has centered on topics which may potentially state and encourage biased, offensive, or disruptive behaviors and would possibly disturb the society, or threaten the general publicor perhaps national safety.

By analyzing the content development and visualizing the social interactions in forums, we want to identify the focuses of public attention and their sentiments similarly as their interaction patterns within the virtual communities expeditiously and effectively. Such knowledge are valuable for detection, understanding, and pursuit the social responses to popular and sensitive problems. In this paper, wepresent Web opinion clustering and informationvisualization techniques, which are components ofan ongoing project of Web opinion analysis and understanding.

Manuscript received January 20, 2013.

Jeswin Roy Dcouth, *Department of Computer Science and Engineering, Karpagam University, Coimbatore, India.*

MohanRaj.T, *Department of Computer Science and Engineering, Karpagam University, Coimbatore, India.*

II.FRAMEWORK

The framework of the overall project isdepicted in Fig. 1with three major components. In the first component, i.e. web forum discover and collection, a monitoring agent monitors a forum, and a crawler fetches messages in the forum according to the hyperlink structure. The collected three dimensions: member identity, timestamp of messages, and structure of threads. In the second component, i.e., Web forum content and link analysis, we utilize machine learning and social network analysis techniques to extract useful knowledge. In the third component, i.e., user interface and interactive information visualization, we provide a user interface for users to submit their queries and present results through interactive visualization techniques for users to explore the forum social networks and content. Unlike Web or regular documents, web opinions are usually less organized, short, and sparse text messages. Thus, traditional ways of clustering Web opinions become very challenging. The special properties of Web opinions that do not exist in regular documents include the following:

1) the content of messages is less focused; 2) the messages are usually short in length ranging from a few words to a couple paragraphs;3) the terms used in the messages are sparse because different users may use different terms to discuss the same topic; 4) the messages contain many unknown terms or slang that do not exist in typical dictionary or ontology, e.g., iPhone and Xbox;5) there are many noises like unrelated text or typographical error so that many web opinions do not fall into any categories;6) the volume of Webopinion messages is huge and ever expanding in an enormous rate; and 7) the topics in these messages keep evolving.

These different properties do not exist in typical documents. In addition to the aforementioned special properties of Web opinions, the traditional clustering characteristics like assigning all documents into clusters or having predefined set of clusters may not be applicable to Web opinion content analysis. Given a collection of documents D document clustering techniques identify a set of clusters C and assign a Boolean value to each pair (di, cj), where D = {d1, d2, . . . , d|D|} and C ={c1, c2, . . . , c|C|}. The Boolean value assigned to each (di, cj) determines whether di is assigned to cj. However, the set of clusters is not predefined in the setting of Web opinions because the topics of discussion are always evolving and usually not known in advance. Therefore, the cluster analysis in this paper employs the unsupervised learning approach in which the set of clusters is not predefined and samples of documents for each cluster are not available. Specifically, we propose a scalable distance-based algorithm for clustering Web opinions.

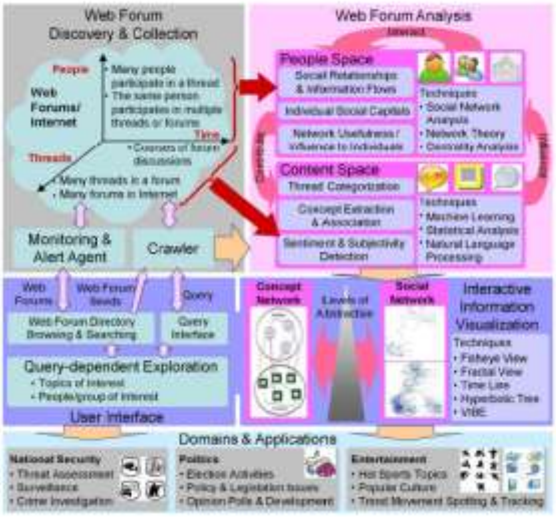


Figure 1 Framework of the Web opinion analysis

A. Content Clustering

The value of performing contentclustering on a forum’s interactive discussions is twofold. The first is to identify and group similar threads together and, hence, to abstract the topics or themes from all clusters. The overall clustering result is to provide a high level content summarization of the underlying threads in forums. It is a typical content clustering value to all document sets. The second value is to unveil the ideological topic similarity between forum participants who may or may not have direct interaction. The value of discovering semantic linkage between participants is unique to the content analysis in online virtual communities. From the perspective of forum participants, it may be useful for them to identify other participants whom they have never interacted with but share with similar ideologies.

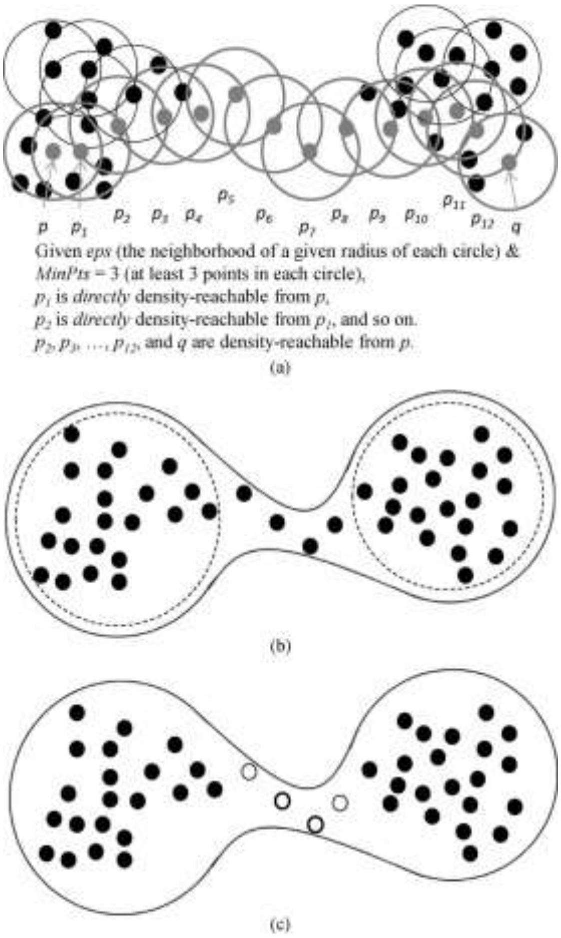
From the perspective of online community analysts, it may be useful to examine the possibility of some participants bearing multiple screen names and participating in multiple threads across different forums. The objective of content clustering in forum discussions is to cluster similar threads without any predefined cardinality and at the same time without forcing any rare topic or noisy threads to be clustered. This process is somewhat different from hierarchical or partition-based document clustering, where each document is assigned into at least one cluster. For example, the partitionbased differential evolution algorithm determines the optimal number of partitions to cluster all data in a data set .event episodes from document sequences DBSCAN is a density-based cluster algorithm that can discover the clusters and filter the noise into a spatial database DBSCAN stands for Density-Based Spatial Clustering of Applications with Noise. The intuition behind DBSCAN clustering comes from how we recognize clusters. A typical density of points within each cluster is considerably higher than outside of the cluster.

B. Core Thread Concept Selection for Clustering

The words found in forum messages are relativelynoisy because the content usually consists of nonedited and conversation-like material. In order to deal with this noisy content, we have defined three criteria for selecting top-N core concepts to represent each thread for the purpose of document clustering. In this paper, top 20 terms in a thread are selected for forming a document vector.1) The first criterion is to select a certain number of topranked terms based on term frequency–inverse document frequency (TFIDF) computation to form a document vector for each thread. The rationale is to exclude some words that are commonly used in conversation or casual online discussions and, at the same time, to use the most important set of terms to represent each thread for similarity comparison in the clustering process.

After tokenizing a document, commonly used terms or stop words are first removed from the term set of each document. Then, the statistics of term frequency *tf* and document frequency *df*for all terms in the document set is computed. 2) The second criterion is to exclude terms that do notcontribute to the comparison process, which computes the similarity score between a pair ofdocument vectors. That is, terms that appear in only a document in a data set are excluded for participating in document vectors. This criterion drops some of the “good” terms in the top N selected terms under the first criterion and replaces them with some other terms havingcertain level of comparison value.3) The third criterion is to use bigrams or two-word terms as part of the document vectors. Natural language processing is an ideal tool to identify noun and verb phrases, which carry higher specificity than single words or monograms do. Nonetheless, the no edited nature and conversational style found in forum messages do not facilitate the natural language processing to perform well.

In this paper, we employ a mechanism to form bigrams by joining two adjacent words without any punctuation or stop word between them. From our empirical observation, a particular bigram has a higher probability of being found in multiple documents than a particular trigram or term with more number of words clusters.



Given *eps* (the neighborhood of a given radius of each circle) & *MinPts* = 3 (at least 3 points in each circle),
p_1 is *directly* density-reachable from p,
p_2 is *directly* density-reachable from p_1, and so on.
$p_2, p_3, \ldots, p_{12}$, and q are density-reachable from p.
(a)

(b)

(c)

Figure 2 Illustrations of DBSCAN and SNN algorithms. (a) Clustering weakly (b) Two high-density sub clusters identified by the dotted circles in the same DBSCAN-identified cluster. (c) SNN: Two clusters of (solid)

In Fig. 2 the similarities between the connecting points (noncore points) and the core points from both clusters are higher than a threshold because they have many nearest neighbors that are the same as the nearest neighbors of the core points from both clusters.As a result, the two clusters are finally joined together.

C. SDC Algorithm

In this paper, we propose a distancebasedalgorithm that ensures that a required density must be reached in the initial clusters and uses scalable distances to expand the initial clusters. Similar to DBSCAN, this approach does not require a predefined number of clusters. It is also able to filter noise. As illustrated in Fig. 3, the solid circles are the initial clusters that meet the requirement of initial density. By scaling up the size of cluster iteration by the iteration shown by dotted circles, the cluster grows until it cannot be further enlarged.

The final clusters are bounded by the outermost circles in Fig. 3. In this case, the radius of cluster is gradually increased to include the closest points to the existing clusters. Points that are directly density-reachable are not necessarily included to the clusters because they are still further away from the expanded clusters. By using the scalable distance, we ensure that the points within a cluster are close to one another with a reasonable distance but not to a few points that are directly density-reachable. The complexity of SDC is *O*(N), which is the same as the complexity of DBSCAN. In our experiment, as presented in Section IV, we select DBSCAN as the benchmarking algorithm.

The proposed SDC algorithm is initialized by identifying small clusters with very high density as initial clusters. Rather than expanding the clusters. My Space social network in political subjects. Density-reachable points, SDC increases the radius of the identified clusters iteratively until it cannot further expand. The notion of densityreach ability in DBSCAN is maintained in SDC.

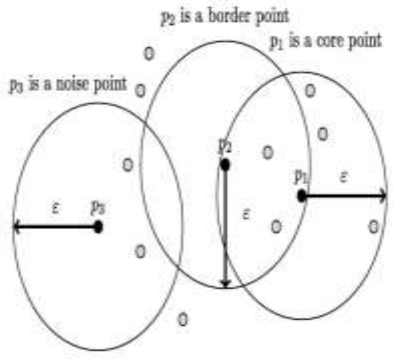


Fig 3.Shows the density-reachable from a point *p*

The eps-neighborhood (*eps*) of a point, *p*, is the set of points such that the distance between each point and *p* is higher than or equal to *eps*. *eps*is measured by the cosine similarity of two threads as presented. As a result, the range of *eps* is [0, 1]. The higher the similarity of the two threads, the higher the value of *eps*. In SDC, the value of *eps*decreases iteration by iteration to include less similar threads in a cluster. To ensure that the initial clusters are dense enough, it is required that *MinPts* points are within the *eps* of the seed point of an initial cluster. *MinPts* is a constant. For each identified initial cluster, SDC iterates to include more points to the cluster until no other points are found. A point is included if its distance from the centroid is larger than *eps*. In each iteration, *eps* is adjusted by Δ*eps*. Δ*eps* is a constant, which is 1/10 of the initial *eps*. For instance, if the initial *eps* is 0.2, Δ*eps* will be 0.02.

Hence the SDC algorithm is initialized by identifying small clusters with very high density as initial clusters. Rather than expanding the clusters by including other techniques.The main steps followed by clustering algorithm are as follows

- Scalability to larger datasets
- Capable of handling high dimensional datasets
- Have the ability to find clusters of any shape
- Have optimal time/space complexities Eliminate any data order dependencies

By using SDC, Web opinions that have similar content are clustered and identified as a theme of discussions. eb opinions that are not similar to others are considered noise because they do not have sufficient participants to contribute their opinions in particular topics. An important theme usually draws attention from many participants, and many Web opinions on this theme will be created. SDC provides a good content analysis to extract the major themes.

III.INTERACTIVE INFORMATION VISUALIZATION

Social network visualization is helpful inexploring the communication between participants in a Web forum. Our interactive visualization tool provides an effective exploration through selecting focus nodes as well as applying fisheye view to explore the area of interest and fractal view to abstract the network so that interesting pattern can be extracted efficiently.The interactive interface allows users to select forum participants as focus nodes by sorting their in-degrees and out degrees, adjust the parameters of fisheye view and fractal view to explore the neighborhood of focus nodes, and filter less relevant nodes, as well as select the topics extracted by the proposed clustering algorithm. The topics extracted by the clustering algorithm are “Al Qaeda,” “Authorizing EavesdroppingProgram,” “Climate Change,”“Congress Approves Iraqi WarBudget,”“Immigration Policy,” “Memorial Day,” “New World Bank Chief,” “Nuclear Weapon,”“Oil,”“President Election,” “Raymond Ronald Kaczynski’s Articles,” “Tainted Food from China,”and “Venezuela Shuts down RCTV.”

It shows alist of forum participants’ screen names, which can be sorted by in-degree or out-degree. Users can also select one or more screen names to pin them as focus nodes for visual navigation. It shows a panel that allows users to select fisheyeview or fractal view and change corresponding parameters tofacilitate visual examination of network connections. Display a list of extracted discussion topics that users can selectone or multiple discussion topics.By selecting a topic in our user interface, we can visualize thesocial network of a particular topic.

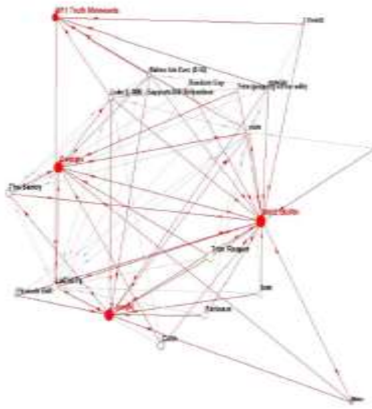


Figure 4. Social network of "Al Qaeda" cluster.

Figure 5. Control Panel. (a) List of forum participants sorted by in-degree, where users can select forum participants by clicking on them. (b) Fisheye view and fractal view parameters. (c) List of extracted discussion topics, where users can select one or multiple discussion topics.

IV.EXPERIMENT

We conducted an experiment thatinvestigated the effectiveness of the SDC algorithm in clustering topics in Web forum and analyzed how the parameters of the eps neighborhood of a point (*eps*) and the minimum number of points required for being aneighborhood (*MinPts*) affect the performance.

Both *eps* and *MinPts* are the important parameters determining the density for clustering. The micro accuracy and macro accuracy are used as the metrics to measure the Performance of SDC and benchmark with the performance of DBSCAN. Micro accuracy measures the overall average clustering accuracy, whereas macro accuracy measures the average of the clustering accuracy of all clusters.

$$microaccuracy = \frac{\sum_{i=1}^{|C|} |H_i|}{N}$$

$$macroaccuracy = \frac{\sum_{i=1}^{|C|} |H_i|/|C_i|}{|C|}$$

where |*C*| is the number of clusters created, |*Hi*| is the numberof threads that is correctly classified in the cluster *Ci*, |*Ci*| is thenumber of threads in the cluster *Ci* and |*Ci*| is greater than one,and *N* is the total number of threads.To determine whether a thread is correctly classified, two judges were recruited in the coding process. The two judges coded the results of DBSCAN and SDC independently. Cohen's Kappa, an interpreterreliability test, was measured to determine the degree of agreement between the two judges. A high degree of agreement was obtained. DBSCAN and SDC do not require specifying the number of clusters to be formed. As a result, changing *eps*and *MinPts* will also affect the number of clusters being generated in addition to accuracy.

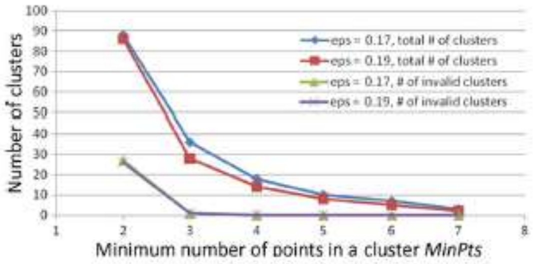


Figure 6 Number of clusters and number of invalid clusters versus MinPts

We first investigate the effect of *MinPts* on micro accuracy and macro accuracy of DBSCAN by setting *eps* as 0.17 and 0.19. The best microaccuracy and macroaccuracy are obtained at *MinPts* = 4when *eps* = 0.19, and the best microaccuracy and macroaccuracy are obtained at *MinPts* = 5 when *eps* = 0.17. Fig. 5 shows the total number of clusters and the number of invalid clusters generated by DBSCAN when *eps* = 0.17 and 0.19. A cluster is considered invalid when a theme cannot be identified from the threads in the cluster. There are some similarities between the threads, but there is not a focus in the discussion between them. A cluster is considered valid if a theme is identified and only some threads are considered noise. For instance, if a cluster has five threads and discussions in these threads are all related to a theme such as Al Qaeda, this cluster will be a valid cluster. When *MinPts* = 4, all clusters with size ofthree or smaller are discarded regardless of the validity ofthe clusters. As a result, as *MinPts* increases from four, itstarts to discard valid clusters of smaller size. The total numberof threads in all clusters also decreases significantly for eachincrement of *MinPts*. As shown in Fig 6, the microaccuracy increases as the totalnumber of clusters decreases until it reaches the optimal at 91%and 87% when *MinPts* = 4 and *MinPts* = 5, respectively.The microaccuracy decreases as the total number of clusterscontinues to decrease. However, when we reach the optimalaccuracy, we are sacrificing the valid clusters of smaller size.

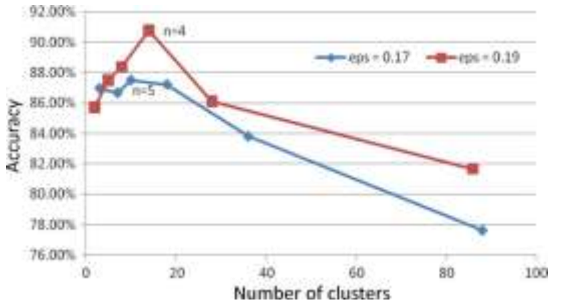


Figure 7 Microaccuracy versus total number of clusters for eps=0.17 and 0.19.

The microaccuracy and macroaccuracycontinue to increase as *eps* increases from 0.1 to 0.22.The value of *eps* controls the minimum similarity between thethreads in a cluster. As we increase the minimum requirement of similarity, the quality of the generated clusters will improve.

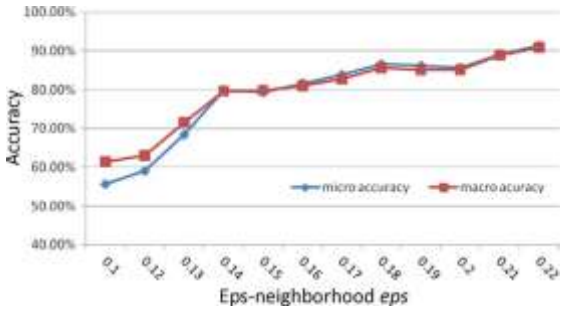


Figure 8 Microaccuracy and macroaccuracy versus eps for MinPts = 3.

Fig.8 compares the macroaccuracy of SDC withthat of DBSCAN. Both show that the microaccuracy and macroaccuracy of SDC are higher than those of DBSCAN. The difference is about 1% to 2% when *eps* is between 0.14 and 0.17. However, the difference is substantial when *eps* is lower than 0.14 or higher than 0.17. The largest difference is as large as 15%, which occurs when *eps* is 0.1. When *eps* is low, DBSCAN tends to merge clusters together and include more noisy threads in the clusters. However, SDC is more capable to retain the clusters without merging multiple clusters through the chain of less relevant threads.

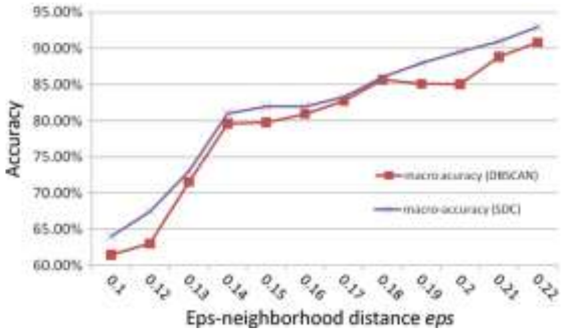


Figure 9 Macroaccuracy achieved by DBSCAN and SDC for different values

Even though SDC achieves good performance in clusteringWebopinions, it has its own limitations. SDC does not require a predefined number of clusters, but it has two parameters *eps*and *MinPts* as inputs. *eps* and *MinPts* are important in identifying the initial clusters. A systematic tuning of these two parameters is needed to achieve optimal performance. These parameters have impacts on micro- and macroaccuracy as well as the number of identified clusters. The tuning can be adjusted to achieve the performance objectives.

V. CONCLUSION

The current Web communities are difficult to organize. Identifying such themes is not trivial. Without a sound Web opinion clustering technique, Web opinions appear as isolated messages spreading along the timeline. In this paper, we have proposed the SDC algorithm for Web opinion analysis. The SDC algorithm overcomes the weakness of DBSCAN algorithm by grouping less number of lessrelevant clusters together when they are density reachable. In our experiment, we have utilized both SDC and DBSCAN algorithms to cluster the major themes in MySpace forum. The result has shown that they are promising to extract clusters of threads with important topics and filter the noise. Moreover,we have shown that SDC performs better than DBSCAN with both microaccuracy and macro accuracy. In addition, using the visualization tools, we have been able to analyze the interaction patterns in each cluster and across clusters. In our future work, we shall further investigate adaptive techniques to make a balance on configuring density-based clustering between these factors to better fit the needs of analysts and users.

ACKNOWLEDGMENT

Thanks to the experts who have contributed for analyzing and extracting social mining trends through web opinion developments via density based clusteringand its simulated solution.

REFERENCES

[1] S. Banerjee, K. Ramanathan, and A. Gupta,"Clustering short texts using wikipedia," in *Proc. ACM SIGIR*, Amsterdam, The Netherlands, 2007,pp. 787–788.

[2] B. Bicici and D. Yuret, "Locally scaled density based clustering," in *Proc.ICANNGA*, 2007, pp. 739–748.

[3] D. Bollegala, Y. Matsuo, and M. Ishizjka, "Measuring semantic similarity between words using Web search engines," in *Proc.Int. WWW Conf.*, 2007, pp. 757–766.

[4] S. Das, A. Abraham, and A. Konar, "Automatic clustering using an improved differential evolution algorithm," *IEEETrans. Syst., Man, Cybern.A, Syst., Humans*, vol. 38, no. 1, pp. 218–237, Jan. 2008.

[5] Y. Zhou, E. Reid, J. Qin, G. Lai, and H. Chen, "U.S. domestic extremist groups on the web: Link and content analysis," *IEEE Intell. Syst.*, vol. 20, no. 5, pp. 44–51, Sep./Oct. 2005.

[6] M. Ester, H. Kregel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. Int.Conf. Knowl. Discov. DataMining (KDD)*, 1996, pp. 226–231.

[7] L. Ertoz, M. Steinbach, and V. Kumar, "Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data," in *Proc. 2nd SIAMInt. Conf. DataMining*, San Francisco, CA, 2003, pp. 47–58.

[8] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, "Structure and evolution of blogspace," *Commun. ACM*, vol. 47, no. 12, pp. 35–39, Dec. 2004.

[9] Q. Mei, C. Liu, H. Su, and C. Zhai, "A probabilistic approach to spatiotemporal theme pattern mining on weblogs," in *Proc.Int. WWW Conf.*,Edinburgh, U.K., 2006, pp. 533–542.

[10] B. A. Nardi, D. J. Schiano, M. Gumbrecht, and L. Swartz, "Why we blog,"*Commun ACM*, vol. 47, no. 12, pp. 41–46, Dec. 2004.

[11] X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text &web with hidden topics from large-scale data collections," in *Proc. Int. WWW Conf.*, Beijing, China, 2008, pp. 91–100.

[12] A. Rosenbloom, "The blogosphere," *Commun. ACM*, vol. 47, no. 12,pp. 31–33, Dec. 2004.

[13] M.Sahami andT.Heilman, "Aweb-based kernel function for measuring the similarity of short text snippets," in *Proc. Int. WWWConf.*, 2006, pp. 2–9.

[14] J. Sander, M. Ester, H. Driegel, and X. Xu, "Density-based clustering in spatial databases: The algorithm GDBSCAN and its application," *Data Mining Knowl. Discov.* vol. 2, no. 2, pp. 169–194, Jun. 1998.

[15] J. Wang, T. Fu, H. Lin, and H. Chen, "A framework for exploring gray Web forums: Analysis of forum-based communication in Taiwan," in *Proc. IEEE Int. Conf. Intell.Security Informat.*, San Diego, CA,May 2006, pp. 498–503.

[16] C.-P. Wei and Y.-H. Chang, "Discovering event evolution patterns from document sequences," *IEEE Trans. Syst., Man, Cybern.A, Syst., Humans*, vol. 37, no. 2, pp. 273–283, Mar. 2007.

[17] J. Wen, J. Nie, and H. Zhang, "Query clustering using user logs," *ACMTrans. Inf.Syst.*, vol. 20, no. 1, pp. 59–81, Jan. 2002.

[18] C. C. Yang, N. Liu, and M. Sageman, "Analyzing the terrorist social network with visualization tools," in *Proc. IEEE Int. Conf.Intell. Security Informat.*, San Diego, CA, May 2006, pp. 331–342.

[19] C. C. Yang, T. D. Ng, J. Wang, C. Wei, and H. Chen, "Analyzing and visualizing gray Web forum structure," in *Proc. Pacific AsiaWorkshopIntell. Security Informat.*, Chengdu, China, 2007, pp. 21–33.

[20] C. C. Yang and T. D. Ng, "Terrorism and crime related weblog social networks: Link, content analysis and information visualization," in *Proc. IEEE Int. Conf. Intell.Security Informat.*, 2007, pp. 55–58.

[21] C. C. Yang and T. D. Ng, "Analyzing content development and visualizing social interactions inWeb forum," in *Proc. IEEEInt. Conf. Intell. Security Informat.*, 2008, pp. 25–30.

Jeswin Roy Dcouth received his B.E degree in Computer Science Engineering from Hindusthan Institute of Technology, Coimbatore in 2011 and currently doing his M.E degree in Karpagam University.

MohanRaj.T received his M.E degree in Computer Science Engineering FromKumaraguru College of Technology, Coimbatore and currently working as Assistant Professor, Department of Computer Science in KarpagamUniversity.