Analyzing Content Development and Visualizing Social Interactions in Web Forum

Christopher C. Yang[1,2] and Tobun D. Ng[2]
1: College of Information Science and Technology
Drexel University, Philadelphia
2: Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Hong Kong SAR

Abstract— **Web forums provide platforms for any Internet users around the world to communicate with each other and express their opinions. In many of the discussions in Web Forums, it involves issues related to terrorism and crime. Some participants are even using the platform to propagandize their ideology or recruit members to commit crime. In this work, we propose a Web forum analysis system to analyze the content development and visualize the social interactions in Web forum.**

***Index Terms*— link analysis, content analysis, Web forum analysis, information visualization, social interactions**

I. Introduction

Internet facilitates the communication between people without geographical boundary. Users interact with each other in a Web forum when they have a common interest. A Web forum is a virtual platform for expressing personal and communal opinions, comments, experiences, thoughts, and sentiments [7]. The messages in a Web forum do not have strong factual content as other Web sites such as CNN or BBC. The factual content is usually hidden in the user subjective opinions. On the other hand, there are factual connections that reflect the focus of discussion among the forum members.

A Web forum is a virtual community that builds on top of the Internet technologies for their members to share information on the subjects of public interest without face-to-face contact with others [7]. In a Web forum, we can find forum members expressing their opinions about political issues including terrorism and government policy. By observing the content development in these discussions, we can identify the public attention and their sentiments. In some specific forums, we also see criminals and terrorists using the virtual communities as a medium to recruit members and identify victims. For example, there are cases that the gangsters use the virtual community to call for members to participate in group raping or drug parties. Some may upload obscene pictures in the Web forum or intrude the privacy of others. Terrorists are also active to propagandize their ideology, recruit members, and even provide videos for instruction of bomb making. By analyzing the content development and visualizing the social interactions in Web forum, we want to identify the focus of interest and their interaction patterns in the virtual community efficiently and effectively. Such knowledge will be valuable for understanding the social response to sensitive issues and crime investigation.

II. Related Work

A number of research works have been conducted on Gray Web Forum in the recent years [4], [8], [9]. The Gray Web Forum is defined as the virtual community formed through Internet forums, which focused on topics that might potentially encourage biased, offensive, or disruptive behaviors and may disturb the society or threaten the public safety. These forums are usually available at specific sites with focused themes such as extremist group, gambling, pirated software, etc. In this work, we investigate the popular forums in a rather general domain such as politics and international policies that cover terrorism, national strategies, and crime related issues.

III. Web Forum Social Network

Figure 1 presents the framework of our Web Forum system. There are three major components, namely Web forum discovery collection, Web forum content and link analysis, and user interface and interactive information visualization. In the Web forum discover and collection component, a module is monitoring the forum and a crawler is fetching the messages in a forum according to the hyperlink structure. In the collected data, we record the three dimensions, member identity, timestamp of messages, and structure of threads. In the Web forum content and link analysis, we utilize machine learning and social network analysis techniques to extract useful knowledge. In the user interface and interactive information visualization component, we provide 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.

IV. Content Clustering

The value of performing content clustering on forum's interactive discussions has two folds. The first is to identify

ISI 2008, June 17-20, 2008, Taipei, Taiwan
1-4244-2415-3/08/$20.00 ©2008 IEEE

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 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 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.

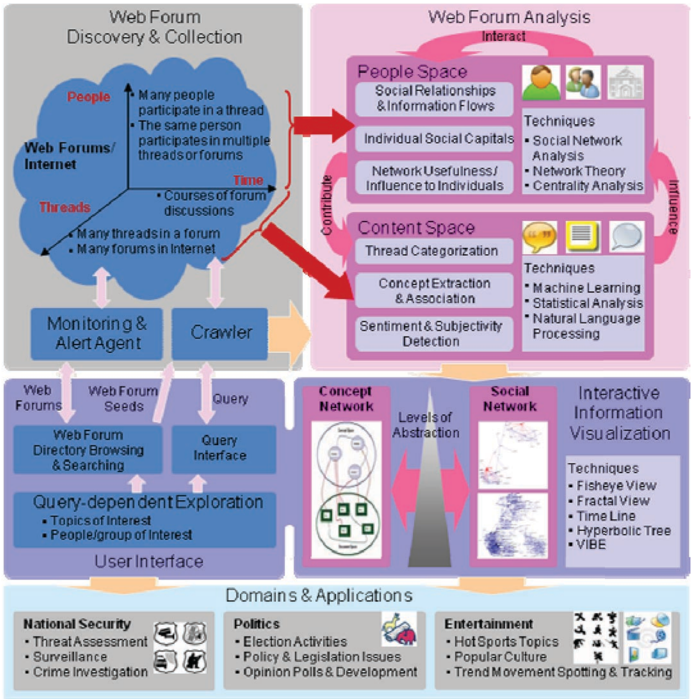


Figure 1. Framework of Web Forum Analysis System

The objective of content clustering in forum discussions is somewhat different from traditional document clustering objective, which is to assign each document into at least one cluster. In forum discussions, any participant is able to post a thread and start the discussion on its topic. Because of this self-interest-oriented posting mechanism, it is possible that the topic of a thread may be unique among all other threads in forums. Hence, in terms of content clustering, a thread with unique or rare topic will not be able to form a cluster or be assigned into any cluster. With this consideration of the forum nature, the objective is to cluster similar threads and simultaneously identify rare-topic or noisy threads from a forum collection.

DBSCAN is a density-based cluster algorithm that is able to discover the clusters and filter the noise in a spatial database [2,3]. DBSCAN stands for Density Based Spatial Clustering of Applications with Noise. The intuition behind DBSCAN clustering is from how we recognize clusters. A typical density of points within each cluster is considerably higher than outside of the cluster. Contrary, the density within the areas of noise is lower than the density in any of the clusters. There are two factors or parameters to quantize this intuitive notion of density of points for clusters and noise in a data set. The first is the boundary, in terms of some chosen distance functions, for any two points to be considered as in a neighborhood. In DBSCAN, this boundary or distance is called an eps-neighborhood of a point. The second is the minimum number of points needs to have in a neighborhood. That is, the neighborhood of a given radius, eps-neighborhood, for each point of a cluster has to contain at least a pre-determined minimum number of points.

Given the global values of eps and MinPts, DBSCAN algorithm makes use of the concept of density-reachability to extend a particular cluster [1,3]. A point, p, is density-reachable from another point, q, with respect to eps and MinPts if there is a chain of points, $p_1, \ldots, p_n, p_1 = q, p_n = p$, such that a particular points, p_{i+1} for i = 1..n-1, is in the eps-neighborhood of a previous point, p_i. To find a cluster, DBSCAN starts with an arbitrary point p and retrieves all points density-reachable from point p with respect to the given eps and MinPts. If no cluster can be formed from point p, DBSCAN will visit the next point in the data set. If a cluster is formed from point p, all other points in this cluster will be used to retrieve all points density-reachable from these points with respect to the given eps and MinPts. The process of forming clusters and finding density-reachable points repeats until all points in the data set are examined. Because the process of cluster expansion or merging clusters in DBSCAN relies on the density-reachability mechanism, some of the resulting clusters may be in nonconvex or elongated shape.

The words found in forum messages are relatively noisy because the content usually consists of non-edited and conversation-like material. In order to deal with this noisy content, we have defined three criteria for selecting core concepts to represent each thread for the purpose of document clustering. The first criterion is to select a certain number of top ranked terms based on 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. In this research work, the number of terms being used for form document vectors is 20.

The second criterion is to exclude terms that do not contribute to the comparison process, which computes the similarity score between a pair of document 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 having certain level of comparison value. The notion of goodness of terms is local to just a particular document. This gain in cohesiveness between document vectors benefits the summarization process. The rationale is to allow all vector elements to contribute in the similarity calculation between a pair of vectors. Non-comparable terms that appear only in a document do not play any role in the comparison process at all.

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 non-edited nature and conversational style found in forum messages do not facilitate the natural language processing to perform well. In this research, 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 does. After extracting bigrams and monograms from a document, we use the following modified TFIDF formula to score each terms:

$$tfidf \times wc^2$$

where *tfidf* is the same as the original TFIDF calculation in monogram vectors, and *wc* is the word count of a term. The multiplier of the square of word count reflects and emphasizes the specificity of a bigram in a vector. The introduction of bigrams or specificity into the vector representation does contribute a certain level of uniqueness into each vector. In addition, the specificity brings in stronger links through bigrams between vectors and removes weaker links represented by monograms.

In the forum data set extracted from myspace.com used in the experiment, there are, at most, 100,128 (*n (n-1)/2*) similarity scores between all pairs of 448 threads that we have collected from MySpace in a period of one week. The details of the experiment are in Section VI. By applying the first criterion of top 20 terms, there are 5,728 similarity pairs left for clustering analysis. By applying the second criterion of terms appearing in at least two documents, the introduced cohesiveness expands the similarity pairs to 7,891. After employing the third criterion of mixing bigrams into vector formation, the uniqueness reduces the similarity pairs down to 4,096. This resulting set of similarity scores has the characteristics of stronger cohesiveness between a pair of vectors because of the use of both bigrams and terms appearing in at least two documents, as well as stronger uniqueness to separate similar pairs or groups. These two characteristics work synchronously with the DBSCAN's capability of deeming areas of higher density (stronger cohesiveness) of points as clusters and those of lower density (stronger uniqueness) of points as noise.

V. INTERACTIVE INFORMATION VISUALIZATION

Social network visualization is helpful in exploring the communication between participants in a Web forum. Our interactive visualization tool provides an effective exploration through selection of focus nodes and applying fisheye view to explore the area of interest and fractal view to abstract the network so that interesting pattern can be extracted efficiently [6]. Figure 2 presents the social network of MySpace Web forum in news and politics collected between May 23 and May 30, 2007. Each node represents a participant in the Web forum. Each link represents an interaction between two participants. The direction of a link from *A* to *B* corresponds to a response from *A* to *B*. A large number of in-degrees means that the participant receives a lot of responds from other participants. On the other hand, a large number of out-degrees means that the participant is responding to many messages posted by other participants. The color of the node represents the topic that the user participates in based on our clustering result. Each topic is a cluster generated by the DBSCAN algorithm. If the node is filled with a solid color, the participant only participates in the corresponding topic. If the node is filled in white with a border of another color, the participants participate in more than one topic and the border color corresponds to the topic that he is most active in. For example, some of the topics with large number of participations are "Al Qaeda", "Authorizing Eavesdropping Program", "Climate Change", "Congress Approves Iraqi War Budget", "Immigration Policy", "Memorial Day", "New World Bank Chief", "Nuclear Weapon", "Oil", "President Election", "Raymond Ronald Karczewski's Articles", "Tainted Food from China", and "Venezuela Shut down RCTV".

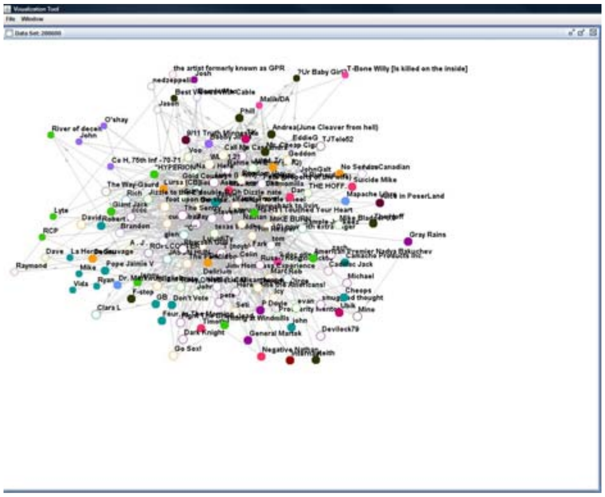


Figure 2. MySpace social network in political subjects

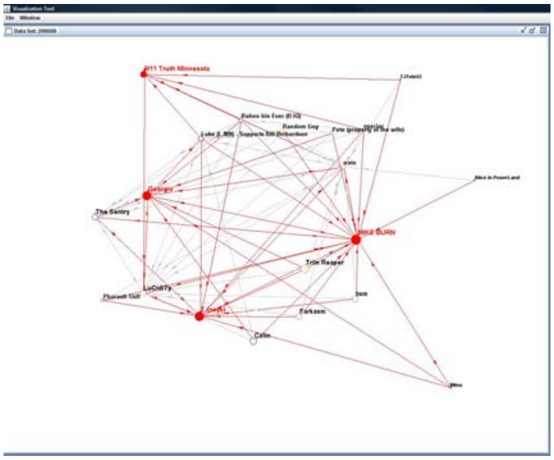
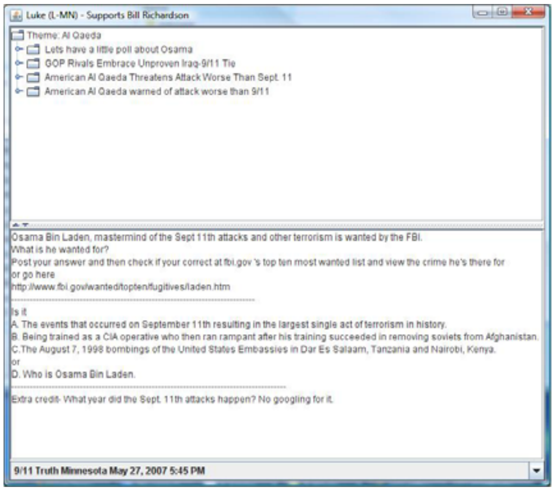


Figure 3. Social network of "Al Qaeda" cluster



Theme: Al Qaeda
- Lets have a little poll about Osama
- GOP Rivals Embrace Unproven Iraq-9/11 Tie
- American Al Qaeda Threatens Attack Worse Than Sept. 11
- American Al Qaeda warned of attack worse than 9/11

Osama Bin Laden, mastermind of the Sept 11th attacks and other terrorism is wanted by the FBI.
What is he wanted for?
Post your answer and then check if your correct at fbi.gov 's top ten most wanted list and view the crime he's there for
or go here
http://www.fbi.gov/wanted/topten/fugitives/laden.htm

Is it
A. The events that occurred on September 11th resulting in the largest single act of terrorism in history.
B. Being trained as a CIA operative who then ran rampant after his training succeeded in removing soviets from Afghanistan.
C.The August 7, 1998 bombings of the United States Embassies in Dar Es Salaam, Tanzania and Nairobi, Kenya.
or
D. Who is Osama Bin Laden.

Extra credit- What year did the Sept. 11th attacks happen? No googling for it.

9/11 Truth Minnesota May 27, 2007 5:45 PM

Figure 4. User interface of presenting messages in "Al Qaeda" cluster.

By selecting a topic in our user interface, we can visualize the social network of a particular topic. For example, Figure 3 shows the social network of the topic "Al Qaeda" and Figure 4 shows the window that present the messages in this cluster.

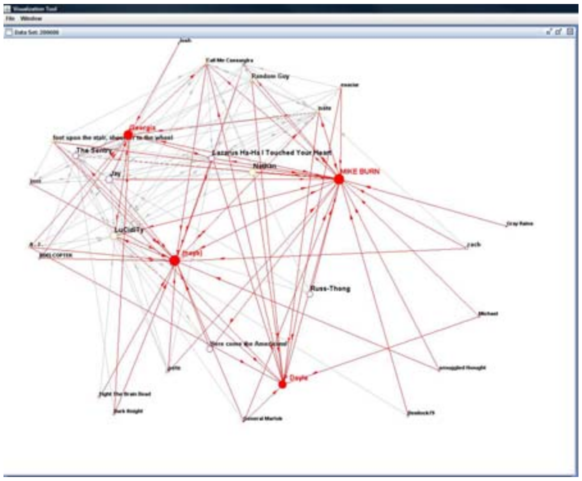


Figure 5. Social network of "Nuclear Weapon" cluster

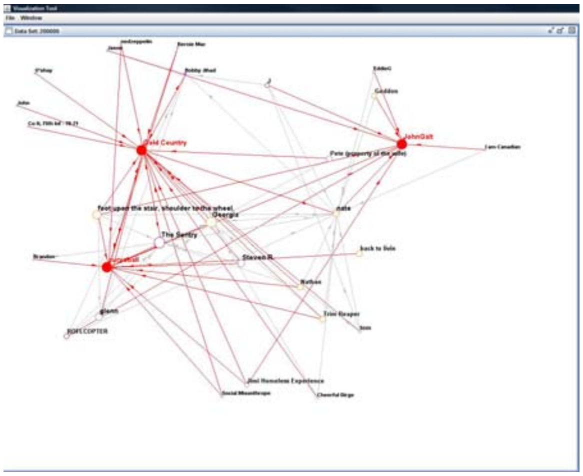


Figure 6. Social network of "Venezuela Shuts down RCTV" cluster

Figure 5 and Figure 6 presents the social networks of two other topics, "Nuclear Weapon" and "Venezuela Shuts down RCTV". Comparing the social networks in Figure 3 and Figure 5, we find many participants who involve in both topics, especially the active participants. However, such pattern cannot be found when we compare Figure 3 with Figure 6. Using such patterns, we can identify the hidden association between topics of discussion. The participants concerns about the terrorist group Al Qaeda, they are also active in the discussion of the nuclear weapon issue. Although the content in these topics do not show any relationship explicitly, the participants and their interactions in these two social networks show a certain degree of association. Indeed, these two topics are strongly related to national security. When we compare with the topic on "Venezuela Shut down RCTV", its relationship to "Al Qaeda" and "Nuclear Weapon" in terms of participants and their interactions is not strong. "Venezuela Shut down RCT" is an issue on the freedom of speech and the internal political policy of Venezuela rather than national security. As a result, the active members in "Al Qaeda" are not the active members in "Venezuela Shut down RCTV". The interactions are also dissimilar.

VI. EXPERIMENT

We have conducted an experiment to investigate the effectiveness of the DBSCAN algorithm in clustering topics in Web forum and analyze how the parameters of EPS and n affect the performance. Both EPS and n are the important parameters determining the density for clustering.

Micro accuracy and macro accuracy are used to measure the quality of the generated clusters by DBSACN. Micro accuracy is the total number of threads that have been correctly clustered dividing by the total number of threads that have been clustered. Macro accuracy is the average of the number of correctly clustered threads in cluster *i* dividing by the number of threads in cluster *i* for all clusters.

DBSCAN does not require specifying the number of clusters to be formed. As a result, changing eps and n will also affect the number of clusters being generated in addition to the accuracy.

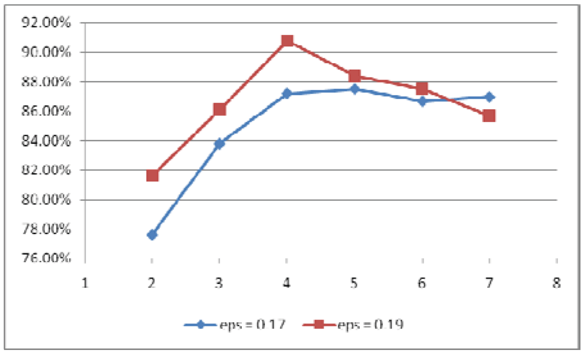


Figure 7. Micro accuracy vs n for esp = 0.17 and 0.19

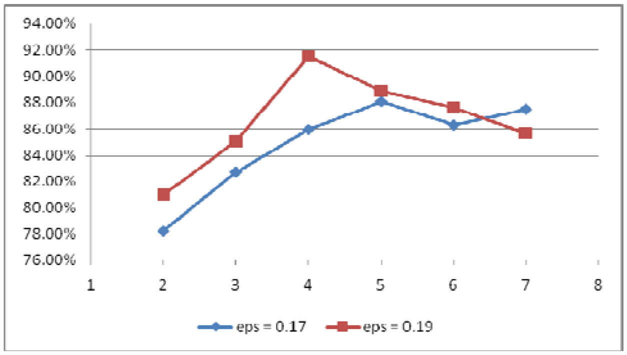


Figure 8. Macro accuracy vs n for eps = 0.17 and 0.19

We first investigate the effect of n on micro and macro accuracies by setting eps as 0.17 and 0.19. Figure 7 and 8 show the micro and macro accuracies with n = 2 to 7. The best micro and macro accuracies are obtained at n = 4 when eps = 0.19 and the best micro and macro accuracies are obtained at n = 5 when eps = 0.17.

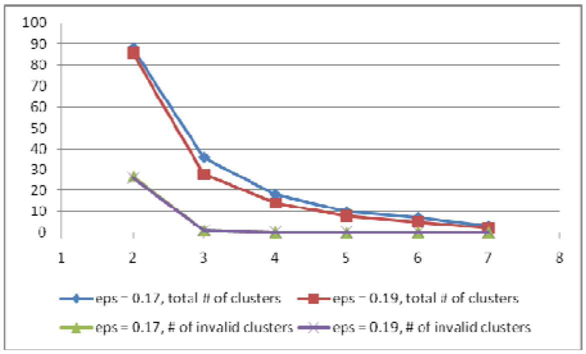


Figure 9. Number of clusters and number of invalid clusters vs n for eps = 0.17 and 0.19

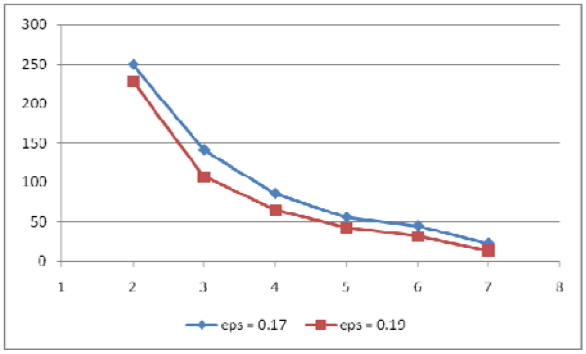


Figure 10. Total number of threads vs n for eps = 0.17 and 0.19

Figure 9 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 conisdered as 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 foucs in the discussion among the threads. A cluster is considered as valid if a theme is identified and only some threads are considered as noise. In Figure 9, it shows that there are many invalid clusters when n = 2. If n is greater than or equal to 3, there is zero or one invalid clusters. Almost all the generated clusters are valid. However, as n continues to increase, the number of valid clusters decreases. The number of valid clusters decreases significantly for each increment of n until the number of valid clusters is 3 when n = 7. The parameter n in DBSCAN restricts the minimum size of clusters. When n = 4, all clusters with size of 3 or smaller will be discarded regardless of the validity of the clusters. As a result, as n increases from 4, it starts to discard valid clusters of smaller size. The total number of threads in all clusters also decreases significantly for each increment of n as shown in Figure 10.

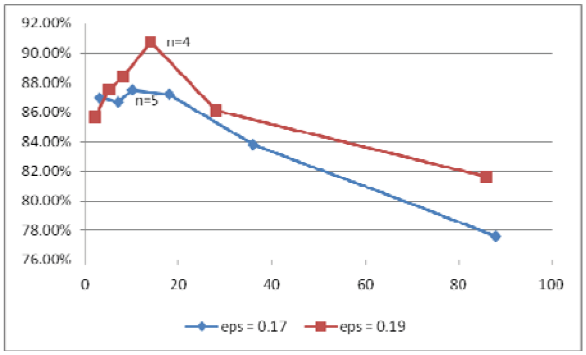


Figure 11. Micro accuracy vs Total number of clusters for eps = 0.17 and 0.19

As shown in Figure 11, the micro accuracy increases as the total number of clusters decreases until it reaches the optimal at 91% and 87% when n = 4 and n = 5, respectively. The micro accuracy decreases as the total number of clusters contintute to decrease. However, when we reach the optimal accuracy, we are sacrificing the valid clusters of smaller size.

If the objective is obtaining the optimal accuracy regardless of the number of clusters formed, we may choose a larger n such as 4 or 5. However, if the objective is removing the invalid clusters and maximizing the number of valid clusters, choosing n = 3 is more reasonable.

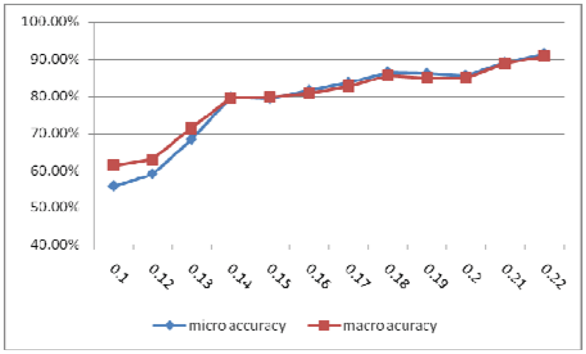


Figure 12. Micro accuracy and macro accuracy vs eps for n = 3.

We further investigate the effect of eps by setting n as 3. As shown in Figure 12, the micro and macro accuracy continues to increase as eps increases from 0.1 to 0.22. eps controls the minimum similarity between the threads in a cluster. As we increases the minimum requirement of similarity, the quality of the generated clusters will improve. However, as shown in Figure 13, the total number of clusters also decreases. The number of invalid clusters decreases until it reaches 0 when eps

= 0.18. The number of valid clusters increases from 35 to 42 as eps increases from 0.11 to 0.14 since the number of invalid clusters decreases significantly in this range. The number of valid clusters decrease from 42 to 25 as eps increases from 0.14 to 0.22.

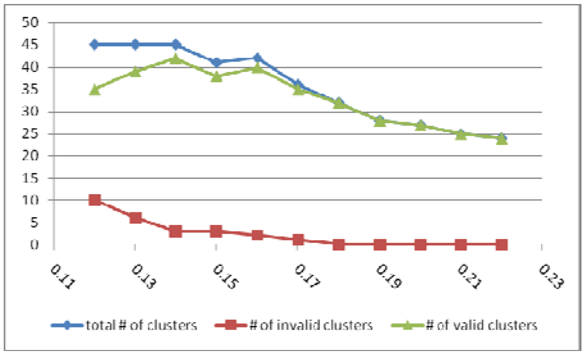


Figure 13. Total # of clusters, # of invalid clusters, and # of valid clusters vs eps for n = 3

Figure 14 shows the plots of the number of valid clusters against the micro accuracy. When eps = 0.18, all invalid clusters are removed and the micro accuracy reaches 87% and it is a balance between the number of valid clusters and micro accuracy. Figure 15 shows the plots of the average number of threads per cluster and the maximum number of threads against eps. The average number of threads is around 4 when eps is between 0.14 and 0.22. The maximum number of threads is around 10 when eps is between 0.15 and 0.20.

Our experiment shows that DBSCAN achieves a promising performance in clustering threads in Web forum although there a large number of noise. By setting a higher value of eps and a higher value of n up to 4, it can achieve micro and macro accuracies above 90% but it will discard smaller clusters and remove less relevant threads from clusters. By reducing eps to around 0.18 and setting n as 3, it will identify more smaller size clusters. DBSCAN is a promising clustering technique to extract the important themes in Web forum. By futher applying the visualization tool, we are able to understand the social interactions among the forum participants and their common interests.

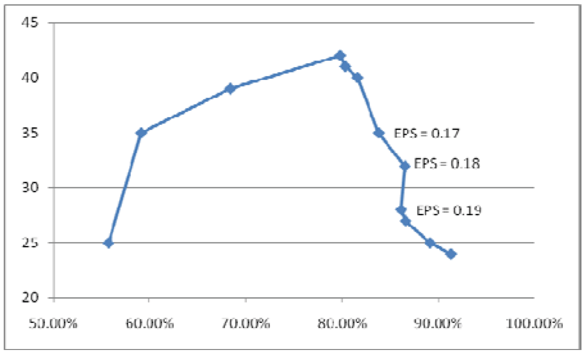


Figure 14. # of valid clusters vs micro accuracy

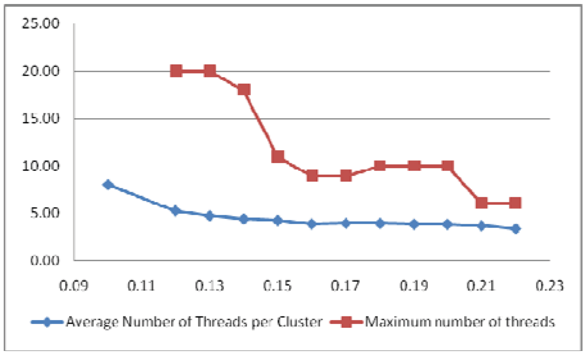


Figure 15. Average number of threads per cluster and maximum number of threads vs eps for n = 3

VII. CONCLUSION

Web forums are virtual communities where forum members communicate with each other without face-to-face interaction and disclosing their true identities. Members express their opinions and discuss with other members in the topics that they have common interests. In some of these Web forums, there are issues related to terrorism and crime. Monitoring and analyzing these forums help us to understand the public interest and extract sensitive topics. It also helps to extract the subgroups that are active in a particular topic and the interactions between the subgroups. In this work, we utilize DBSCAN to cluster the major themes in MySpace forum. The result shows that it is promising to extract clusters of threads with important topics and filter the noise. Using the visualization tools, we are able to analyze the interaction patterns in each cluster and across clusters.

REFERENCES

[1] B. Bicici and D. Yuret, "Locally Scaled Density Based Clustering," *Proceedings of ICANNGA*, 2007.

[2] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," *Proceedings of International Conference o Knowledge Discovery and Data Mining* (*KDD*),1996.

[3] J. Sander, M. Ester, H. Driegel, and X. Xu, "Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications," *Data Mining and Knowledge Discovery Archive*, 2(2), June, 1998.

[4] J. Wang, T. Fu, H. Lin, and H. Chen, "A Framework for Exploring Gray Web Forums: Analysis of Forum-Based Communication in Taiwan," *Proceedings of IEEE International Conference on Intelligence and Security Informatics*, San Diego, CA, May, 2006, pp.498-503.

[5] J. Wen, J. Nie, and H. Zhang, "Query Clustering using User Logs," *ACM Transactions on Information Systems*, 20(1), January, 2002, pp.59-81.

[6] C. C. Yang, N. Liu, and M. Sageman, "Analyzing the Terrorist Social Network with Visualization Tools," *Proceedings of IEEE International Conference on Intelligence and Security Informatics*, San Diego, CA, May, 2006, pp.331-342.

[7] C. C. Yang, T. D. Ng, J. Wang, C. Wei, and H. Chen, "Analyzing and Visualizing Gray Web Forum Structure," *Proceedings of Pacific-Asia Workshop on Intelligence and Security Informatics*, Chengdu, China, 2007, pp.21-33.

[8] C. C. Yang and T. D. Ng, "Terrorism and Crime Related Weblog Social Network: Link, Content Analysis and Information Visualization," *Proceedings of IEEE International Conference on Intelligence and Security Informatics*, New Brunswick, NJ, 2007.

[9] Y. Zhou, E. Reid, J. Qin, G. Lai, and H. Chen, "U.S. Domestic Extremist Groups on the Web: Link and Content Analysis," *IEEE Intelligent Systems*, 20(5), 2005, pp.44-51.