Analyzing and Visualizing Web Opinion Development and Social Interactions With Density-Based Clustering

Christopher C. Yang and Tobun Dorbin Ng, *Member, IEEE*

***Abstract*—Due to the advancement of Web 2.0 technologies, a large volume of Web opinions is available on social media sites such as Web forums and Weblogs. These technologies provide a platform for Internet users around the world to communicate with each other and express their opinions. Analysis of developing Web opinions is potentially valuable for discovering ongoing topics of interests of the public like terrorist and crime detection, understanding how topics evolve together with the underlying social interaction between participants, and identifying important participants who have great influence in various topics of discussions. Nonetheless, the work of analyzing and clustering Web opinions is extremely challenging. 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 unsatisfactory performance. In this paper, we investigated the density-based clustering algorithm and proposed the scalable distance-based clustering technique for Web opinion clustering. We conducted experiments and benchmarked with the density-based algorithm to show that the new algorithm obtains higher microaccuracy and macroaccuracy. This Web opinion clustering technique enables the identification of themes within discussions in Web social networks and their development, as well as the interactions of active 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*—Density-based clustering, information visualization, social media analytics, social network analysis, web opinion mining.**

I. Introduction

THE Internet facilitates communication between people not limited to geographical boundaries. For example, 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 in discussion threads [19]. There, Web users are able to share their personal details to a circle of friends, amplify their voices and sentiment, establish online communication in a topic of interest, and promote an ideology [8]–[10], [12], [20]. The continuous user interaction on a Web forum becomes a virtual community for members to share thoughts on subjects of their interest without face-to-face contact with each other. The messages in a Web forum typically do not have strong factual content as information-rich news sites such as CNN or BBC. Nonetheless, the factual content is usually hidden within user's subjectivity in opinions. In addition, there are factual connections that reflect the focuses of discussions among the forum members of a thread.

Web forum members express their opinions virtually on all kinds of topics such as political and social issues, religion, entertainment, movies, music, traveling experiences, consumer products, sports, health, and technology. For example, to an extreme, the Gray Web Forum in the recent years has focused on topics that might potentially state and encourage biased, offensive, or disruptive behaviors and might disturb the society, or threaten the public or even national safety [15], [20], [23]. By analyzing the content development and visualizing the social interactions in Web forums, we want to identify the focuses of public attention and their sentiments as well as their interaction patterns in the virtual communities efficiently and effectively. Such knowledge will be valuable for detecting, understanding, and tracking the social responses to popular and sensitive issues.

In this paper, we present Web opinion clustering and information visualization techniques, which are components of an ongoing project of Web opinion analysis and understanding [21]. The framework of the overall project is depicted in Fig. 1 with 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 messages are analyzed with the emphasis on these 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

Manuscript received August 1, 2009; revised April 15, 2010; accepted June 16, 2010. Date of publication March 24, 2011; date of current version October 19, 2011. This paper was recommended by Editor W. Pedrycz.

C. C. Yang was with The Chinese University of Hong Kong, Shatin, NT, Hong Kong. He is now with the College of Information Science and Technology, Drexel University, Philadelphia, PA 19104 USA (e-mail: Chris.Yang@drexel.edu).

T. D. Ng is with the Digital Library Laboratory, The Chinese University of Hong Kong, Shatin, NT, Hong Kong (e-mail: tbng@se.cuhk.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCA.2011.2113334

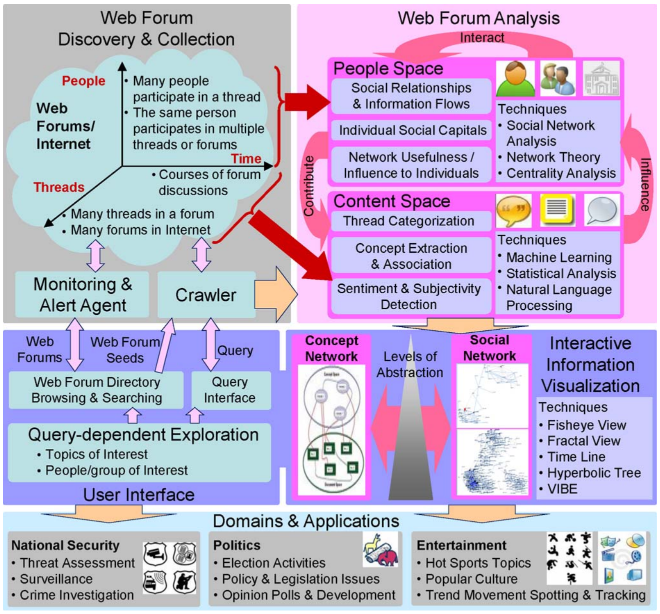


Fig. 1. Framework of the Web opinion analysis and understanding project.

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 Web opinion 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. Traditional document clustering techniques that work well in clustering regular documents usually do not work well in Web opinion clustering.

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 $(d_i,\ c_j)$, where $D = \{d_1, d_2, \ldots, d_{|D|}\}$ and $C = \{c_1, c_2, \ldots, c_{|C|}\}$. The Boolean value assigned to each $(d_i,\ c_j)$ determines whether d_i is assigned to c_j. 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.

In order to make use of the new clustering results, we also developed interactive information visualization tools to explore the interactions between Web users to understand the network structure of each extracted topic of discussion. Fisheye views and fractal views are applied to extract pattern of interactions and identify users with the highest influence.

A. Related Work

Applying document clustering techniques on Web opinions is not appropriate because of the Web opinion properties and the design of these clustering techniques. Many document clustering techniques such as K-means [4] and EM [5] require prespecified number of clusters and then classify all documents in the collection to one of the prespecified clusters. In Web opinion clustering, we cannot predefine or predict the number of clusters. The clusters change from time to time. In addition, many Web opinions are noisy, and therefore, they are not assigned to any cluster. In our preliminary studies [20], it is found that over 50% of Web opinions are noise. Due to the sparseness of terms appearing in Web opinions, the distance measured by document vectors is usually large although the corresponding documents are related. All of these reasons cause the poor performance of Web opinion clustering when document clustering techniques are applied directly.

Recent literature reported some techniques for clustering short texts. Most of them rely heavily on external sources, such as search engine. Sahami and Heilman [13] submitted each short text as a query to a Web search engine and used the search results to create a context vector. Similarity between short texts will then be measured by the similarity between their corresponding context vectors. Bollegala *et al.* [3] also adopted search engines and used the page counts and lexico-syntactic patterns to measure the similarity of words. Such measurement of word similarity can be utilized to categorize short texts in which the words in common are less. Banerjee *et al.* [1] expanded the short text representation by additional features from Wikipedia concepts. The Wikipedia concepts are the titles of the matched Wikipedia articles. It showed that it improved the result of clustering substantially in most cases. Wen *et al.* [17] made use of user logs of selected documents from search results to augment a set of corresponding queries on a search engine in order to perform query clustering for improving the performance of question answering systems. The similarity between two queries might be deducted from the common documents the users selected for them. Phan *et al.* [11] collected a large-scale external data collection defined as universal data set to build a classifier on a set of training data and a set of hidden topics from universal data set. According to Phan *et al.* [11], the universal data set must be large enough to cover a large number of concepts for classification and must be consistent with the training data and future unseen data.

Although these techniques showed improvement in clustering short texts, relying on external sources is not feasible in many cases. It creates excess network traffic each time the system expands the representation of a short text. Considering the huge number of Web opinions in the Web, it is inefficient and costly. At the same time, it also overloads the external sources such as search engines or Wikipedia. A risk exists when these external sources terminate the connections from these systems in order to provide consistent service to other users. Moreover, these resources are usually lagging behind the changes of term usage in Web opinion communities. Web users actively create new terms, and the topics evolve rapidly. Unless using raw data, it would be impossible to rely on external sources to gain new and updated knowledge. The objective of this paper is to develop an effective algorithm for clustering

Web opinions that overcomes the weaknesses of the existing document clustering techniques.

The discussion of content clustering in Section II reviews distance-based clustering techniques and introduces the scalable distance-based clustering (SDC) algorithm. Section III illustrates the interactive information visualization techniques that present clustered Web opinions. Then, Section IV details the experiments on selecting suitable parameters for density-based clustering and comparing the performance between the established density-based spatial clustering of application with noise (DBSCAN) and the new SDC. Section V concludes this paper and explores future direction.

II. Content Clustering

The value of performing content clustering 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 partition-based differential evolution algorithm determines the optimal number of partitions to cluster all data in a data set [4]. 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 a unique or rare topic will not be able to form a cluster or be assigned into any cluster. Some threads may remain unique over time, whereas some may become the leads to form new clusters. The content clustering in this paper focuses on grouping threads gathered from a snapshot based on users' query-dependent exploration, which does not require the time factor to identify and discover event episodes from document sequences [16].

A. DBSCAN

DBSCAN is a density-based cluster algorithm that can discover the clusters and filter the noise into a spatial database [6], [14]. 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. Contrary, the density within the noisy regions is lower than the density in any of the clusters. There are two factors or parameters to quantify 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 being in a neighborhood. In DBSCAN, this boundary or distance is called an eps-neighborhood of a point (eps). The second is the minimum number of points $(MinPts)$ needed in a neighborhood, i.e., the neighborhood of a given radius, i.e., eps-neighborhood, for each point in a cluster has to contain at least a predetermined 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 [2], [14]. A point q is *directly* density-reachable from a point p if there are $MinPts$ number of points within a distance of eps from p, and q is one of these points. A point q is density-reachable from a point p if there is a chain of points $p, p_1, p_2, \ldots, p_n$, q such that p_1 is directly density-reachable from p, p_{i+1} is directly density-reachable from p_i, and q is directly density-reachable from p_n. A cluster is formed by DBSCAN for all points that are density-reachable. In the clustering process, 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 expanding a cluster or merging clusters in DBSCAN relies on the density-reachability mechanism, some of the resulting clusters may be in nonconvex or elongated shape. The complexity of DBSCAN is $O(\mathrm{N})$, which is ideal for clustering Web opinions as the number of Web opinions grows exponentially.

Shared nearest neighborhood (SNN) is an enhanced density-based clustering algorithm [7]. The major difference between SNN and DBSCAN is the definition of similarity between pairs of points. SNN defines the similarity between a pair of points as the number of nearest neighbors the two points share. The density is measured by the sum of the similarities of the k nearest neighbors of a point. Points with high density are selected as core points, and points with low density are identified as noise and removed. All remaining points that have a similarity higher than a threshold to a core point are clustered together. SNN is also able to form clusters in nonconvex and elongated shapes, but it removes the problem of varying density in DBSCAN [7]. It was also reported that SNN outperformed DBSCAN in a number of data set [7]. However, the complexity of SNN is $O(\mathrm{N}^2)$. Although SNN may achieve higher performance, it is not preferred in clustering Web opinions due to the fast growing number of Web opinions in social media.

B. Core Thread Concept Selection for Clustering

The words found in forum messages are relatively noisy 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 top-ranked 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 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. Noncomparable terms that appear only in a document do not play any role in the comparison process at all.
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 nonedited 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 does. After extracting bigrams and monograms from a document, we use the following modified TFIDF formula to score each term:

$$tfidf \times wc^2 \tag{1}$$

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. Given the text content of "nuclear weapon," "nuclear" and "weapon" are extracted as monograms, whereas "nuclear weapon" is extracted as a bigram simultaneously. Both monograms "nuclear" and "weapon" will have $tfidf$ as their TFIDF values because the wc^2 factor is just one. The bigram "nuclear weapon" will have $4tfidf$ as its TFIDF value because the wc^2 factor is 2^2 or 4. 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.

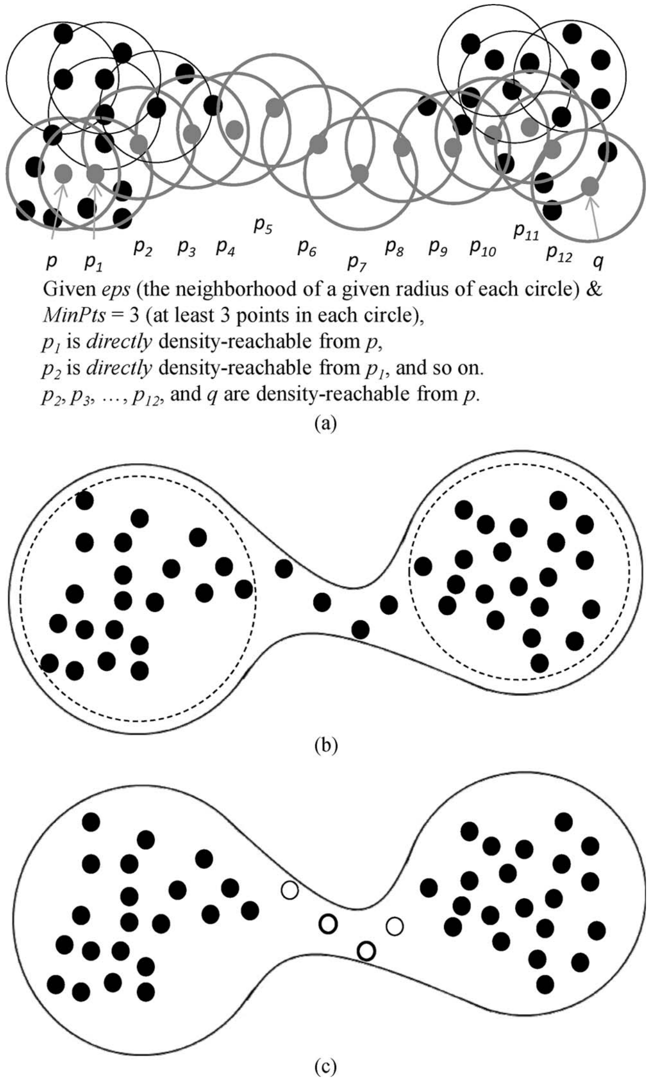
In Web opinion clustering, we represent each unit of Web opinions as a weighted vector of core concepts. In our paper, we consider a thread in a Web forum as a unit of Web opinion. Only 20 terms with the highest scores are selected to represent a thread with common terms being removed and representative terms being retained. A similarity function is utilized to measure the distance between any two vectors of threads. The cosine similarity function that is commonly used in information retrieval is selected as the distance in this paper. Conceptually, the cosine similarity function bears an inverse relationship to an ordinary distance function such as Euclidean distance. The cosine similarity score (i.e., distance) is higher when the similarity between threads increases. The cosine similarity between Web opinions wo_i and wo_j is computed as follows:

$$sim(wo_i, wo_j) = \frac{\sum_{k=1}^{n} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2 \times \sum_{k=1}^{n} w_{jk}^2}} \tag{2}$$

where w_{ik} is the weight of term k in Web opinion i, w_{jk} is the weight of term k in Web opinion j, and n is the total number of terms.

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 IV. By applying the first criterion for selecting top 20 terms to form document vectors, 5728 similarity pairs are left for clustering analysis. By applying the second criterion for selecting top 20 terms appearing in at least two documents, the introduced cohesiveness expands the similarity pairs to 7891. After further employing the third criterion of mixing bigrams appearing in at least two documents into vector formation, the uniqueness reduces the similarity pairs down to 4096. 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 capability of deeming areas of higher density (stronger cohesiveness) of points as clusters and those of lower density (stronger uniqueness) of points as noise.

The advantages of no prespecified number of clusters and noise filtering for DBSCAN perfectly fit the requirements of clustering Web opinions. The Web opinions are evolving, and the number of topics of discussions cannot be exactly predicted. Among all Web opinions, many of them are noise that do not fall into any particular topic of discussions. However, DBSCAN forms clusters of any shape based on the density-reachability mechanism, which is not practical in Web opinion clustering. Considering an elongated shape of cluster, a concept at one end of the elongated region is less relevant to another concept at the other end of the elongated region although these concepts are density-reachable to each other. For example, a thread or message about "2008 Olympics" is density-reachable to another thread about the Olympics celebration in "Beijing" and "Tian An Men." On the other hand, a thread about "Tian An Men" is density-reachable to another thread about "Tian An


p
p1
p2
p3
p4
p5
p6
p7
p8
p9
p10
p11
p12
q
Given eps (the neighborhood of a given radius of each circle) &
MinPts = 3 (at least 3 points in each circle),
p1 is directly density-reachable from p,
p2 is directly density-reachable from p1, and so on.
p2, p3, …, p12, and q are density-reachable from p.

(a)

(b)

(c)

Fig. 2. Illustrations of DBSCAN and SNN algorithms. (a) Clustering weakly relevant concepts at both left and right ends in an elongated shape of DBSCAN-identified cluster. (b) Two high-density subclusters identified by the dotted circles in the same DBSCAN-identified cluster. (c) SNN: Two clusters of (solid) core points are joined because the connecting points have similarities to some core points of both clusters that are higher than the threshold.

Men Square Protests of 1989." However, "2008 Olympics" is not relevant to "Tian An Men Square Protests of 1989." In this paper, we tackle the weakness of DBSCAN and propose the SDC algorithm. Instead of including all the density-reachable points to a cluster, SDC increases the radius of a cluster with high density gradually until no other points can be reached by further increment of distance from the centroid of a cluster. It ensures that all points within a cluster are reachable by all other points in a reasonable distance instead of being reachable by only some points of the cluster.

Fig. 2(a) illustrates how a cluster is formed when all of these points are density-reachable, although some points are not directly density-reachable by some other points, e.g., the points $p_2, p_3, \ldots, p_{12}$, and q is not directly density-reachable by the point p. Centering at the point p, a circle defined by a given eps-neighborhood containing at least two other points $(MinPts = 3)$ is drawn at the bottom left corner in Fig. 2(a). The point p_1 is directly density-reachable from p, and it is used as a center point to draw the next circle with the same eps-neighborhood and $MinPts$ criteria. The same neighborhood-forming mechanism repeatedly applies to points p_2 (which is directly density-reachable from p_1), $p_3, \ldots, p_{12}$, and q in sequence, as shown by the gray circles. However, in Web opinion clustering, it may create problems. Some of the Web opinions may be directly density-reachable to other Web opinions due to some connection in the weak concepts; that is, two clusters may be put together even though they are not as relevant to each other. Fig. 2(b) illustrates the two potential clusters that may be grouped together by DBSCAN. Although it is reported that SNN performs better than DBSCAN, it is possible that SNN puts two clusters together due to the similarities between the connecting points and the core points. The similarity is defined as the number of nearest neighbors two points are sharing. A connecting point may share many nearest neighbors with core points in either one or both of the two originally disconnected clusters. Fig. 2(c) presents the core points (solid points) that create two clusters. In Fig. 2(c), 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.

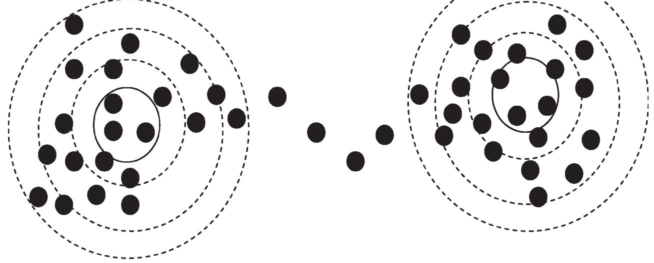


Fig. 3. Illustration of a scalable distance-based algorithm.

C. SDC Algorithm

In this paper, we propose a distance-based algorithm 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 a 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 $O(\mathrm{N})$, 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 by including other

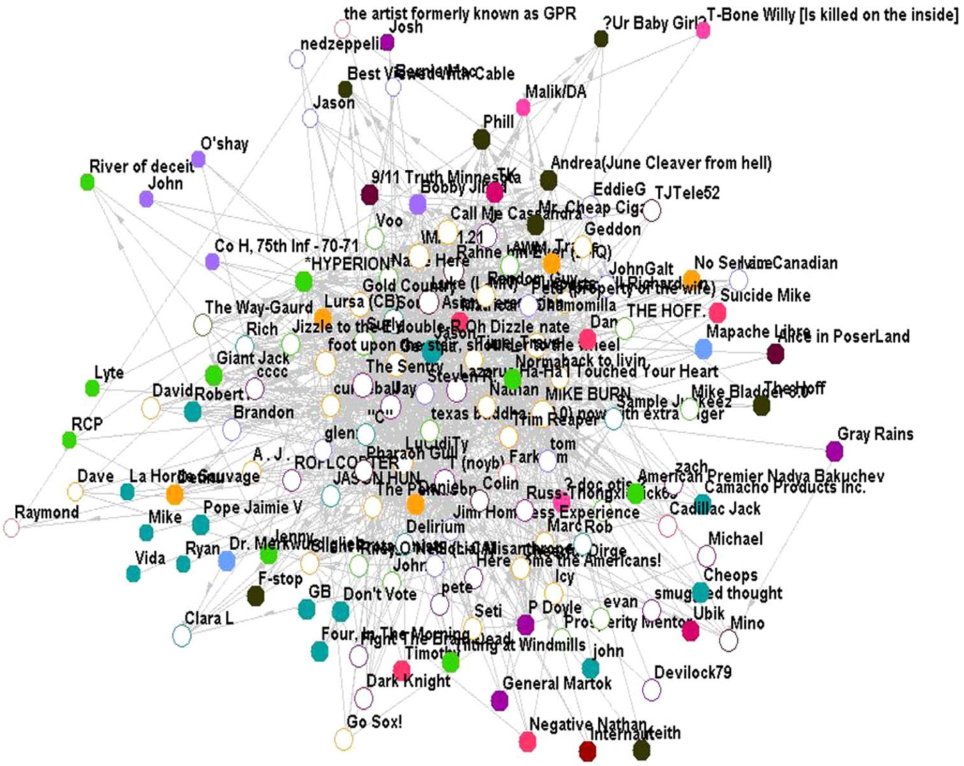


Fig. 4. MySpace social network in political subjects.

density-reachable points, SDC increases the radius of the identified clusters iteratively until it cannot further expands.

The notion of density-reachability in DBSCAN is maintained in SDC. 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 in (1). 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 Δeps. Δeps is a constant, which is 1/10 of the initial eps. For instance, if the initial eps is 0.2, Δeps will be 0.02.

SDC algorithm:

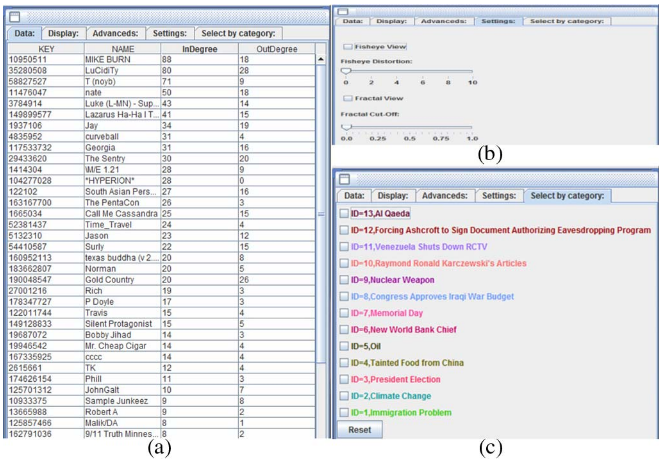
```
Initialize all points as unclassified points, $S = \{p_1, p_2, \ldots, p_n\}$
Repeat
  Randomly select a point $p_i$ in $S$ as a seed
  If the number of points in eps-neighborhood of $p_i \geq MinPts$
    Then, create an initial cluster $C_j$ by including the seed and all its eps-neighborhood points
      $S = S - C_j$
    Else $p_i$ is classified as $X$
      $S = S - p_i$
Until $S = \varnothing$
For each initial cluster $C_j$
  Repeat
    Find the centroid
    $eps = eps - \Delta eps$
    Add points from $X$ in which the distance from the centroid of the cluster is larger than $eps$
  Until no other points are found
The points remaining in $X$ are considered noise.
```

By using SDC, Web opinions that have similar content are clustered and identified as a theme of discussions. Web 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. By analyzing the social network of the extracted themes, we can further investigate the leading and active participants in a theme of discussion.

III. 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 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 [18], [22]. 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.

Fig. 4 presents the social network of MySpace Web forum in news and politics collected between May 23 and May 30,



KEY	NAME	InDegree	OutDegree
10950511	MIKE BURN	88	18
35280508	LuCidiTy	80	28
58827527	T (noyb)	71	9
11476047	nate	50	18
3784914	Luke (L-MN) - Sup...	43	14
149899577	Lazarus Ha-Ha I T...	41	15
1937106	Jay	34	19
4835952	curveball	31	4
117533732	Georgia	31	16
29433620	The Sentry	30	20
1414304	\M/E 1.21	28	9
104277028	*HYPERION*	28	0
122102	South Asian Pers...	27	16
163167700	The PentaCon	26	3
1665034	Call Me Cassandra	25	15
52381437	Time_Travel	24	4
5132310	Jason	23	12
54410587	Surly	22	15
160952113	texas buddha (v 2....	20	8
183662807	Norman	20	5
190048547	Gold Country	20	26
27001216	Rich	19	3
178347727	P Doyle	17	3
122011744	Travis	15	4
149128833	Silent Protagonist	15	5
19687072	Bobby Jihad	14	3
19946542	Mr. Cheap Cigar	14	4
167335925	cccc	14	4
2615661	TK	12	4
174626154	Phill	11	3
125701312	JohnGalt	10	7
10933375	Sample Junkeez	9	8
13665988	Robert A	9	2
125857466	Malik/DA	8	1
162791036	9/11 Truth Minnes...	8	2

☐ Fisheye View

Fisheye Distortion: 0 2 4 6 8 10

☐ Fractal View

Fractal Cut-Off: 0.0 0.25 0.5 0.75 1.0

☐ ID=13,Al Qaeda
☐ ID=12,Forcing Ashcroft to Sign Document Authorizing Eavesdropping Program
☐ ID=11,Venezuela Shuts Down RCTV
☐ ID=10,Raymond Ronald Karczewski's Articles
☐ ID=9,Nuclear Weapon
☐ ID=8,Congress Approves Iraqi War Budget
☐ ID=7,Memorial Day
☐ ID=6,New World Bank Chief
☐ ID=5,Oil
☐ ID=4,Tainted Food from China
☐ ID=3,President Election
☐ ID=2,Climate Change
☐ ID=1,Immigration Problem

Reset

Fig. 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.

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 to a node means that a participant receives a lot of responses from other participants. On the other hand, a large number of out-degrees from a node means that a participant responds 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 clustering algorithm. If a node is filled with a solid color, the corresponding participant only participates in a single topic. If a node is filled in white with a border of another color, the participant participates in more than one topic, and the border color corresponds to the topic that he is most active in. The topics extracted by the clustering algorithm 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 Shuts down RCTV."

In order to effectively visualize the busy network presented in Fig. 4, the visualization tool provides a set of controls through various control panels as shown in Fig. 5. Fig. 5(a) shows a list 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. Fig. 5(b) shows a panel that allows users to select fisheye view or fractal view and change corresponding parameters to facilitate visual examination of network connections. Fig. 5(c) displays a list of extracted discussion topics that users can select one or multiple discussion topics.

By selecting a topic in our user interface, we can visualize the social network of a particular topic. For example, Fig. 6 shows the social network of the topic "Al Qaeda," and Fig. 7 shows the window that presents the messages in this cluster. Users can click on a node to see the messages posted by the corresponding forum participant and the threads that he has involved.

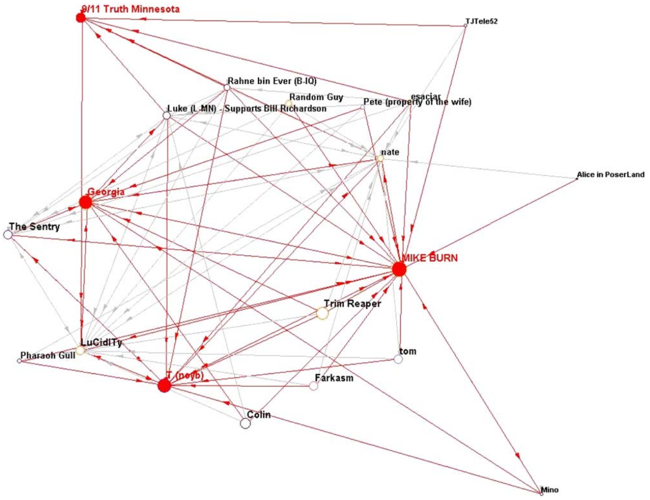
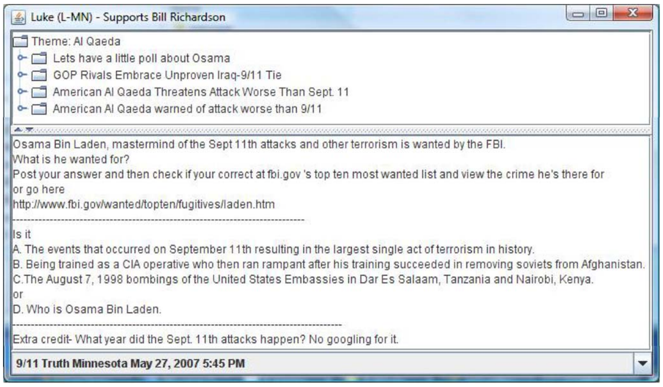


Fig. 6. Social network of "Al Qaeda" cluster.



Luke (L-MN) - Supports Bill Richardson

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

Fig. 7. User interface of presenting messages in "Al Qaeda" cluster.

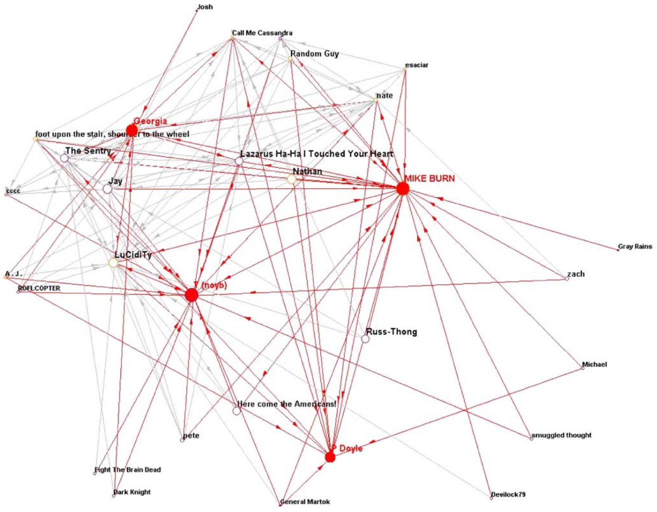


Fig. 8. Social network of "Nuclear Weapon" cluster.

Figs. 8 and 9 present the social networks of two other topics: "Nuclear Weapon" and "Venezuela Shuts down RCTV." Comparing the social networks in Figs. 6 and 8, we find many participants who are involved in both topics, particularly the active participants. However, such pattern cannot be found when we compare Fig. 6 with Fig. 9. Using such patterns, we can identify the hidden association between topics of discussion. The participants who are concerned about the terrorist group Al Qaeda are also active in the discussion of the nuclear weapon issue. Although the content in these topics does 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

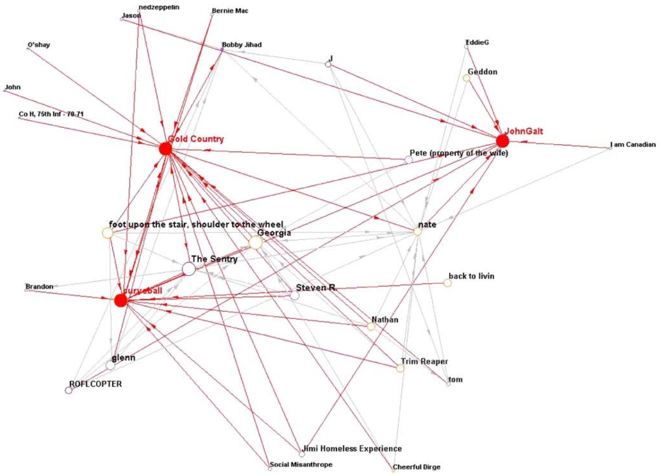


Fig. 9. Social network of "Venezuela Shuts down RCTV" cluster.

national security. When we examine the topic on "Venezuela Shuts down RCTV," its relationships to "Al Qaeda" and "Nuclear Weapon" are not strong in terms of participants and their interactions. "Venezuela Shuts down RCTV" 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 Shuts down RCTV." The interactions are also dissimilar.

To illustrate the interactive visualization, we present two case studies to further elaborate the interactions involved.

— Case Study 1: Fig. 4 presents the social network for all the extracted topics. However, the social network is too intricate that it is difficult for users to easily visualize the details. To understand the forum social network, one may want to identify the most active participants and visualize their neighborhoods and filter less active and relevant participants in the network. Using the control panel, users can select the participants who have received the most number of responses by sorting their in-degrees. Fig. 10(a) shows that "Mike Burn" and "LuCidiTy" are selected. Using these two nodes as foci, applying fisheye view and fractal view generates the views as shown in Fig. 10(b). We can identify the topics participated by these active participants and their neighbors which are "Al Qaeda," "nuclear weapon," "climate change," and "immigration problem." The topics are denoted by the color.

— Case Study 2: Figs. 6 and 8 present the social networks of "Al Qaeda" and "nuclear weapon," respectively. By selecting both topics, one can see the combined network as shown in Fig. 11(a). By selecting the overlapped nodes, one can find the participants who are active in both discussions and examine their messages to identify the connection between two topics. Fig. 11(b) and (c) shows two messages—one in "Al Qaeda" and another one in "nuclear weapon." These messages are related to Bush policies in military strategy against Islamic countries and nuclear weapon. Both messages are closely related to terrorism. The participants in these two threads are highly overlapped, and therefore, the content is related although each of them has its focus on "Al Qaeda" or "nuclear weapon."

IV. Experiment

We conducted an experiment that investigated 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 a neighborhood $(MinPts)$ affect the performance. Both eps and $MinPts$ are the important parameters determining the density for clustering.

The microaccuracy and macroaccuracy are used as the metrics to measure the performance of SDC and benchmark with the performance of DBSCAN. Microaccuracy measures the overall average clustering accuracy, whereas macroaccuracy measures the average of the clustering accuracy of all clusters

$$microaccuracy = \frac{\sum_{i=1}^{|C|} |H_i|}{N} \tag{3}$$

$$macroaccuracy = \frac{\sum_{i=1}^{|C|} |H_i|/|C_i|}{|C|} \tag{4}$$

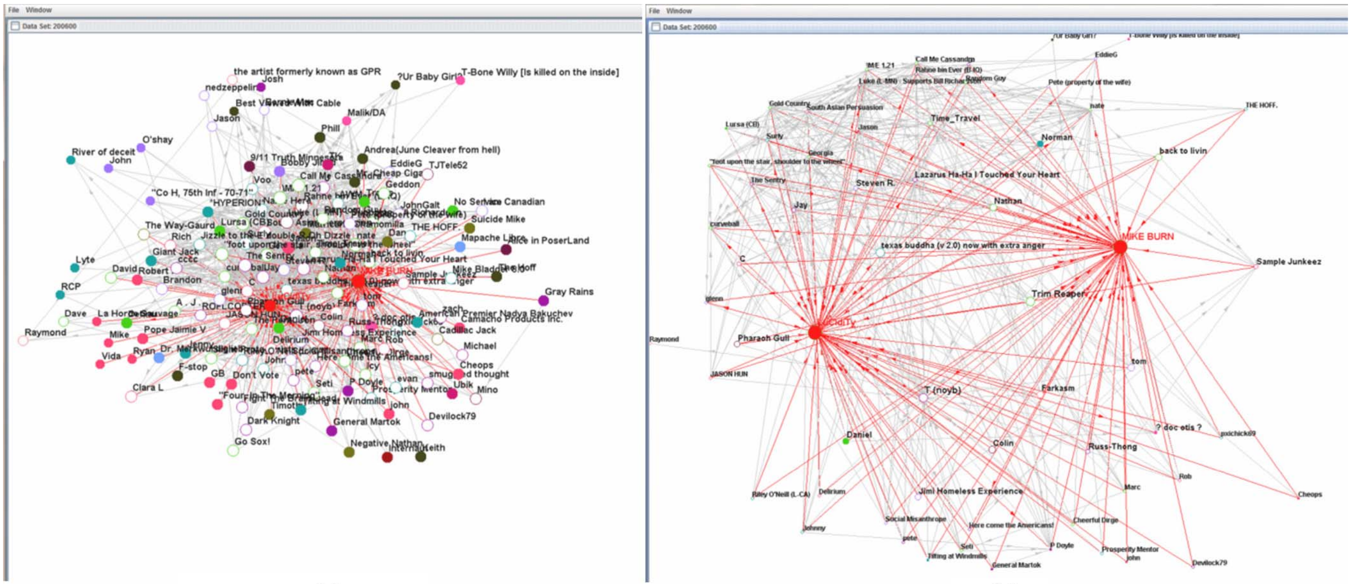
where $|C|$ is the number of clusters created, $|H_i|$ is the number of threads that is correctly classified in the cluster C_i, $|C_i|$ is the number of threads in the cluster C_i and $|C_i|$ is greater than one, and N is the total number of threads.

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 interrater reliability test, was measured to determine the degree of agreement between the two judges. A high degree of agreement was obtained $(\text{Cohen's Kappa} = 0.96)$.

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.

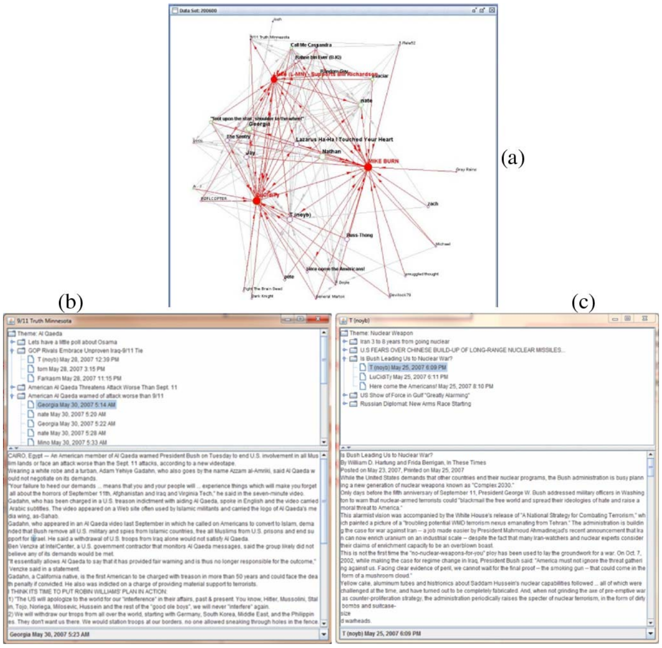
We first investigate the effect of $MinPts$ on microaccuracy and macroaccuracy of DBSCAN by setting eps as 0.17 and 0.19. Figs. 12 and 13 show the microaccuracy and macroaccuracy with $MinPts = 2$ to 7. The best microaccuracy and macroaccuracy are obtained at $MinPts = 4$ when $eps = 0.19$, and the best microaccuracy and macroaccuracy are obtained at $MinPts = 5$ when $eps = 0.17$.

Fig. 14 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. However, if a cluster has five threads, where two of them discuss Al Qaeda's ideology, one of them discusses the new development of the original twin towers site in New York but not about the 911 attack, two of them discuss issues in Afghanistan but not about Al Qaeda or terrorism, this thread will be considered invalid. In this experiment, a cluster is considered invalid when less than half of the total number of threads in a cluster is of the same theme. In Fig. 14, it shows that there are many invalid clusters when $MinPts = 2$. When $MinPts$ is greater than or equal to three, there is zero or one invalid cluster. Almost all the generated



(a) (b)

Fig. 10. (a) Selecting two nodes as foci. (b) Applying fisheye view and fractal view.



9/11 Truth Minnesota

Theme: Al Qaeda
Lets have a little poll about Osama
GOP Rivals Embrace Unproven Iraq-9/11 Tie
T (noyb) May 28, 2007 12:39 PM
tom May 28, 2007 3:15 PM
Farkasm May 28, 2007 11:15 PM
American Al Qaeda Threatens Attack Worse Than Sept. 11
American Al Qaeda warned of attack worse than 9/11
Georgia May 30, 2007 5:14 AM
nate May 30, 2007 5:20 AM
Georgia May 30, 2007 5:22 AM
nate May 30, 2007 5:28 AM
Mino May 30, 2007 5:33 AM

CAIRO, Egypt — An American member of Al Qaeda warned President Bush on Tuesday to end U.S. involvement in all Muslim lands or face an attack worse than the Sept. 11 attacks, according to a new videotape.
Wearing a white robe and a turban, Adam Yehiye Gadahn, who also goes by the name Azzam al-Amriki, said Al Qaeda would not negotiate on its demands.
"Your failure to heed our demands ... means that you and your people will ... experience things which will make you forget all about the horrors of September 11th, Afghanistan and Iraq and Virginia Tech," he said in the seven-minute video.
Gadahn, who has been charged in a U.S. treason indictment with aiding Al Qaeda, spoke in English and the video carried Arabic subtitles. The video appeared on a Web site often used by Islamic militants and carried the logo of Al Qaeda's media wing, as-Sahab.
Gadahn, who appeared in an Al Qaeda video last September in which he called on Americans to convert to Islam, demanded that Bush remove all U.S. military and spies from Islamic countries, free all Muslims from U.S. prisons and end support for Israel. He said a withdrawal of U.S. troops from Iraq alone would not satisfy Al Qaeda.
Ben Venzke at IntelCenter, a U.S. government contractor that monitors Al Qaeda messages, said the group likely did not believe any of its demands would be met.
"It essentially allows Al Qaeda to say that it has provided fair warning and is thus no longer responsible for the outcome," Venzke said in a statement.
Gadahn, a California native, is the first American to be charged with treason in more than 50 years and could face the death penalty if convicted. He also was indicted on a charge of providing material support to terrorists.
I THINK ITS TIME TO PUT ROBIN WILLIAMS' PLAN IN ACTION:
1) "The US will apologize to the world for our "interference" in their affairs, past & present. You know, Hitler, Mussolini, Stalin, Tojo, Noriega, Milosevic, Hussein and the rest of the "good ole boys", we will never "interfere" again.
2) We will withdraw our trops from all over the world, starting with Germany, South Korea, Middle East, and the Philippines. They don't want us there. We would station troops at our borders. no one allowed sneaking through holes in the fence.

Georgia May 30, 2007 5:23 AM

T (noyb)

Theme: Nuclear Weapon
Iran 3 to 8 years from going nuclear
U.S FEARS OVER CHINESE BUILD-UP OF LONG-RANGE NUCLEAR MISSILES...
Is Bush Leading Us to Nuclear War?
T (noyb) May 25, 2007 6:09 PM
LuCidiTy May 25, 2007 6:11 PM
Here come the Americans! May 25, 2007 8:10 PM
US Show of Force in Gulf "Greatly Alarming"
Russian Diplomat: New Arms Race Starting

Is Bush Leading Us to Nuclear War?
By William D. Hartung and Frida Berrigan, In These Times
Posted on May 23, 2007, Printed on May 25, 2007
While the United States demands that other countries end their nuclear programs, the Bush administration is busy planning a new generation of nuclear weapons known as "Complex 2030."
Only days before the fifth anniversary of September 11, President George W. Bush addressed military officers in Washington to warn that nuclear-armed terrorists could "blackmail the free world and spread their ideologies of hate and raise a moral threat to America."
This alarmist vision was accompanied by the White House's release of "A National Strategy for Combating Terrorism," which painted a picture of a "troubling potential WMD terrorism nexus emanating from Tehran." The administration is building the case for war against Iran -- a job made easier by President Mahmoud Ahmadinejad's recent announcement that Iran can now enrich uranium on an industrial scale -- despite the fact that many Iran-watchers and nuclear experts consider their claims of enrichment capacity to be an overblown boast.
This is not the first time the "no-nuclear-weapons-for-you" ploy has been used to lay the groundwork for a war. On Oct. 7, 2002, while making the case for regime change in Iraq, President Bush said: "America must not ignore the threat gathering against us. Facing clear evidence of peril, we cannot wait for the final proof -- the smoking gun -- that could come in the form of a mushroom cloud."
Yellow cake, aluminum tubes and histrionics about Saddam Hussein's nuclear capabilities followed ... all of which were challenged at the time, and have turned out to be completely fabricated. And, when not grinding the axe of pre-emptive war as counter-proliferation strategy, the administration periodically raises the specter of nuclear terrorism, in the form of dirty bombs and suitcase-
size
d warheads.

T (noyb) May 25, 2007 6:09 PM

Fig. 11. (a) Social networks of two topics, namely, "Al Qaeda" and "nuclear weapon." (b) Message extracted from "Al Qaeda." (c) Message extracted from "nuclear weapon."

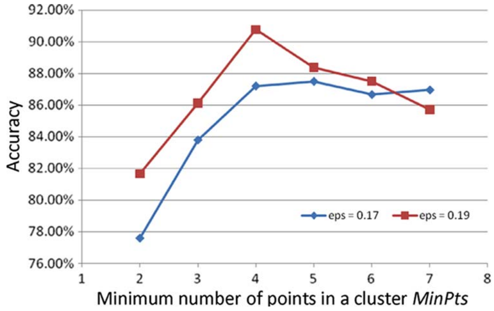


Fig. 12. Microaccuracy versus $MinPts$ for $eps = 0.17$ and 0.19.

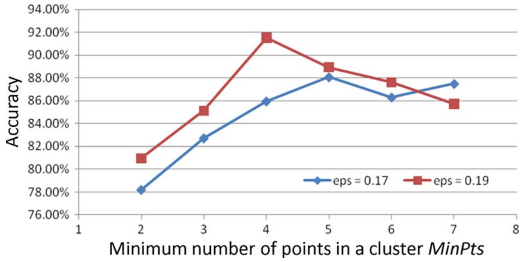


Fig. 13. Macroaccuracy versus $MinPts$ for $eps = 0.17$ and 0.19.

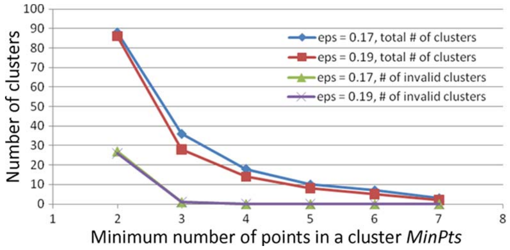


Fig. 14. Number of clusters and number of invalid clusters versus $MinPts$ for $eps = 0.17$ and 0.19.

clusters are valid. However, as $MinPts$ continues to increase, the number of valid clusters decreases. The number of valid clusters decreases significantly for each increment of $MinPts$ until the number of valid clusters is three when $MinPts = 7$. The parameter $MinPts$ in DBSCAN restricts the minimum size of clusters. When $MinPts = 4$, all clusters with size of three or smaller are discarded regardless of the validity of the clusters. As a result, as $MinPts$ increases from four, it starts to discard valid clusters of smaller size. The total number of threads in all clusters also decreases significantly for each increment of $MinPts$ as shown in Fig. 15.

As shown in Fig. 16, the microaccuracy increases as the total number 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 clusters continues to decrease. However, when we reach the optimal accuracy, we are sacrificing the valid clusters of smaller size.

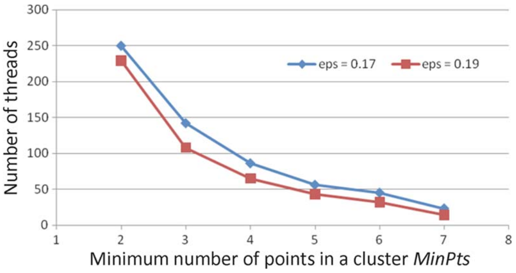


Fig. 15. Total number of threads versus $MinPts$ for $eps = 0.17$ and 0.19.

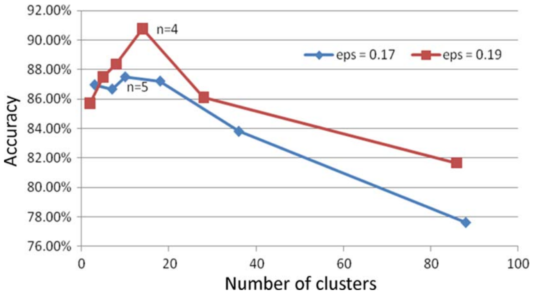


Fig. 16. Microaccuracy versus total number of clusters for $eps = 0.17$ and 0.19.

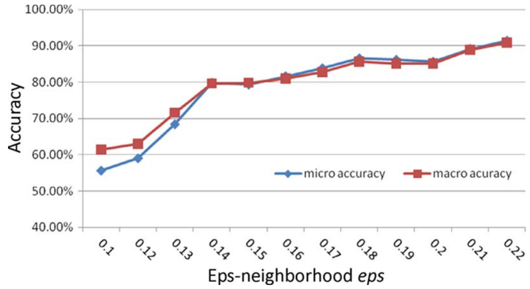


Fig. 17. Microaccuracy and macroaccuracy versus eps for $MinPts = 3$.

If the objective is to obtain the optimal accuracy regardless of the number of clusters formed, we may choose a larger $MinPts$ such as four or five. However, if the objective is to remove the invalid clusters and maximize the number of valid clusters, choosing $MinPts = 3$ will be more reasonable.

We further investigate the effect of eps by setting $MinPts$ as three. As shown in Fig. 17, the microaccuracy and macroaccuracy continue to increase as eps increases from 0.1 to 0.22. The value of eps controls the minimum similarity between the threads in a cluster. As we increase the minimum requirement of similarity, the quality of the generated clusters will improve. However, as shown in Fig. 18, the total number of clusters also decreases. The number of invalid clusters decreases until it reaches zero 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 decreases from 42 to 25 as eps increases from 0.14 to 0.22.

Fig. 19 shows the plots of the number of valid clusters against microaccuracy. When $eps = 0.18$, all invalid clusters are removed, and the microaccuracy reaches 87%. This combination reveals a balance between the number of valid clusters and microaccuracy. Fig. 20 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 four when eps is between 0.14 and 0.22. The maximum number of threads is around ten when eps is between 0.15 and 0.20.

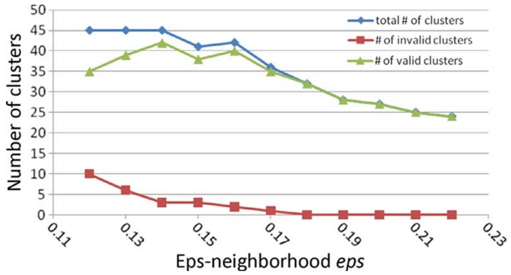


Fig. 18. Total number of clusters, number of invalid clusters, and number of valid clusters versus eps for $MinPts = 3$.

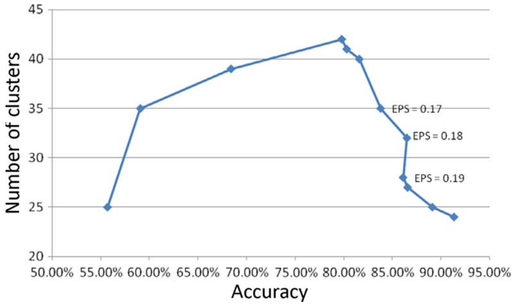


Fig. 19. Number of valid clusters versus microaccuracy.

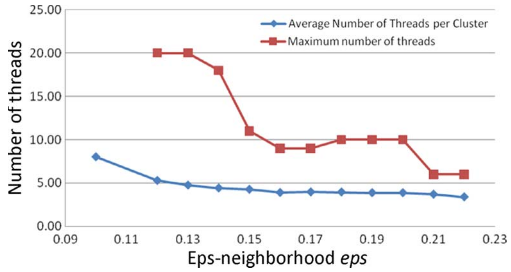


Fig. 20. Average number of threads per cluster and maximum number of threads versus eps for $MinPts = 3$.

Our experiment shows that DBSCAN achieves a promising performance in clustering threads in Web forum although there is significant amount of noise. By setting a higher value of eps and a higher value of $MinPts$ up to four, it can achieve microaccuracy and macroaccuracy above 90%, but it will discard smaller clusters and remove less relevant threads from clusters. By reducing eps to around 0.18 and setting $MinPts$ as three, it will identify more smaller size clusters. DBSCAN is a promising clustering technique to extract the important themes in Web forum. By further applying the visualization tool, we are able to understand the social interactions among the forum participants and their common interests.

We have further conducted an experiment to investigate the performance of SDC algorithm in clustering Web opinions and compare its performance with DBSCAN's. In this experiment, $MinPts$ is set as three since it is tested and shown to achieve the optimal performance in both DBSCAN and SDC. We have investigated how SDC performs as we adjust the parameter eps, which is the initial eps value to identify the initial clusters. Fig. 21 shows the total number of clusters and the number of valid clusters constructed by SDC versus eps.

The experiment shows that the total number of clusters decreases when eps increases. Similarly, the number of valid clusters decreases when eps increases. However, the difference

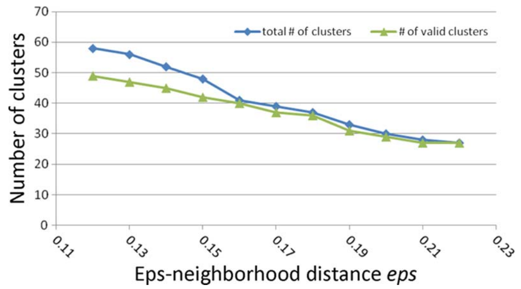


Fig. 21. Total number of clusters and number of valid clusters by SDC for different values of eps.

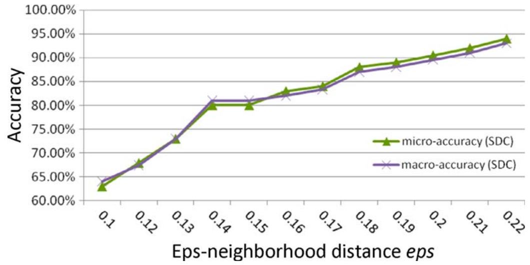


Fig. 22. Microaccuracy and macroaccuracy achieved by SDC for different values of eps.

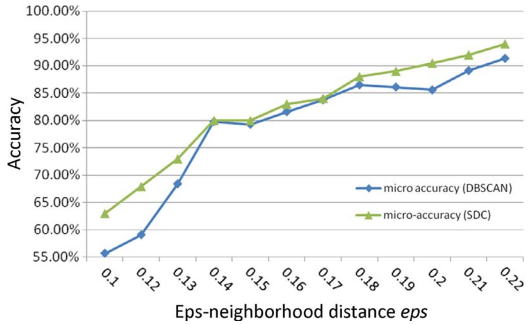


Fig. 23. Microaccuracy achieved by DBSCAN and SDC for different values of eps.

between the total number of clusters and the number of valid clusters decreases when eps increases. That means that the number of invalid clusters decreases when eps increases. The total number of clusters and the number of valid clusters are about the same when eps is higher than 0.16.

Fig. 22 shows that the microaccuracy and macroaccuracy of SDC increase as eps increases. The difference between microaccuracy and macroaccuracy of SDC is not substantial. It is less than 1%. That means that all clusters generated by SDC have similar accuracy.

Fig. 23 compares the microaccuracy of SDC with that of DBSCAN. Fig. 24 compares the macroaccuracy of SDC with that 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. SDC is also able to increase the size of a cluster gradually without including many noisy threads. It shows that the overall performance of SDC is substantially better than that of DBSCAN.

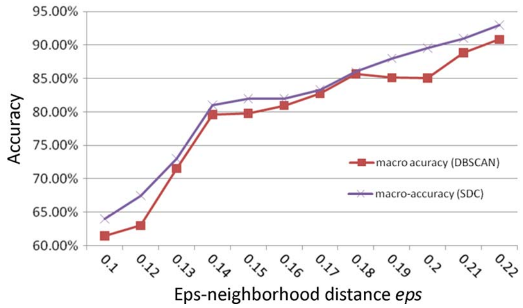


Fig. 24. Macroaccuracy achieved by DBSCAN and SDC for different values of eps.

It is found that SDC performs consistently better than DBSCAN in a wide range of eps no matter whether a higher purity of clusters is obtained when a higher eps value is used or a larger number of clusters are created when a lower eps value is used. A t-test shows that the difference between the performances of DBSCAN and SDC is significant at 0.05 level. Such finding ensures that a scalable distance-based approach is more suitable than a pure density-based approach for Web opinion clustering.

Although SDC achieves good performance in clustering Web opinions, 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

Web 2.0 technologies encourage individuals to share opinions with each other. Web users participate in Web forums or Weblog platform when expressing their opinions and interacting with other users of similar interests. Through this means, Web forums become virtual communities where forum members communicate with each other without face-to-face interaction and disclosing their true identities. Themes within Web opinions develop as participants are actively discussing specific topics. Unfortunately, the current Web communities are difficult to organize. Identifying such themes is not trivial. Nonetheless, monitoring and analyzing these forums help us understand the public interest, extract sensitive topics, and organize content into themes. They also help extract the subgroups of individuals that are active in a particular topic and the interactions between the subgroups. It would be particularly valuable to analyze some of these Web forums related to terrorism and crime activities. Clustering Web opinions is indeed important for intelligence and security informatics because many terrorist groups and criminal organizations are active in using such channels to propagandize their ideologies, by spreading threatening messages and recruiting members in organized crime.

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 less relevant 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 macroaccuracy. We have also found that there is a tradeoff between the number of identified clusters and the purity of clusters when we adjust the parameter *eps* in SDC and DBSCAN. 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.

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,” *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 38, no. 1, pp. 218–237, Jan. 2008.

[5] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley-Interscience, 2000.

[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. Data Mining (KDD)*, 1996, pp. 226–231.

[7] L. Ertöz, M. Steinbach, and V. Kumar, “Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data,” in *Proc. 2nd SIAM Int. Conf. Data Mining*, 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 and T. Heilman, “A web-based kernel function for measuring the similarity of short text snippets,” in *Proc. Int. WWW Conf.*, 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,” *ACM Trans. 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 Asia Workshop Intell. 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 in Web forum,” in *Proc. IEEE Int. Conf. Intell. Security Informat.*, 2008, pp. 25–30.

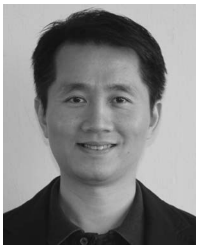
[22] C. C. Yang and M. Sageman, “Analysis of terrorist social networks with fractal views,” *J. Inf. Sci.*, vol. 35, no. 3, pp. 299–320, Jun. 2009.

[23] 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.

Christopher C. Yang received the B.S., M.S., and Ph.D. degrees in computer engineering from the University of Arizona, Tucson, in 1990, 1992, and 1997, respectively.

He is an Associate Professor with the College of Information Science and Technology, Drexel University, Philadelphia, PA. He has also been an Associate Professor with the Department of Systems Engineering and Engineering Management and the Director of the Digital Library Laboratory with The Chinese University of Hong Kong, Shatin, NT, Hong Kong, an Assistant Professor with the Department of Computer Science and Information Systems, The University of Hong Kong, Pokfulam, Hong Kong, and a Research Scientist with the Department of Management Information Systems, The University of Arizona, Tucson. His recent research interests include social media analytics, Web 2.0, health informatics, Web search and mining, cross-lingual information retrieval and knowledge management, security informatics, and electronic commerce. He has published over 200 refereed journal and conference papers in *Journal of the American Society for Information Science and Technology* (JASIST), *Decision Support Systems* (DSS), IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON ROBOTICS AND AUTOMATION, IEEE COMPUTER, IEEE INTELLIGENCE SYSTEMS, *Information Processing and Management* (IPM), *Journal of Information Science, Graphical Models and Image Processing, Optical Engineering, Pattern Recognition, International Journal of Electronic Commerce, Applied Artificial Intelligence, Intelligence and Security Informatics* (ISI), World Wide Web, Special Interest Group on Information Retrieval, International Conference on Information Systems, Conference on Information and Knowledge Management (CIKM), and more.

Prof. Yang has edited several special issues on social media, Web mining, multilingual information systems, knowledge management, and electronic commerce in JASIST, DSS, IPM, and IEEE Transactions. He chaired and served in many international conferences and workshops, including serving as a Program Cochair in IEEE ISI, Association for Computing Machinery (ACM) CIKM, International Conference on Electronic Commerce, and ACM Special Interest Group on Knowledge Discovery and Data Mining Workshop on Intelligence and Security Informatics. He has also frequently served as an invited panelist in the National Science Foundation and other government agencies' review panels.



Tobun Dorbin Ng (M'95) received the B.S. degree in business administration majoring in management information systems and finance and the M.S. and Ph.D. degrees in management information systems from The University of Arizona, Tucson, in 1990, 1993, and 2000, respectively.

He is a Research Scientist and Project Manager with the Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, NT, Hong Kong. He has also been a Systems Scientist and Assistant Professor with the Computer Science Department, Carnegie Mellon University, Pittsburgh, PA. His research interests include healthcare informatics, supply chain management, radio-frequency identification technologies, digital libraries, social media analytics, digital video analysis, multimedia information retrieval, semantic interoperability, large-scale knowledge discovery, high-performance supercomputing, web-based search engine, neural-network computing, and collaborative computing.

Dr. Ng has been a member of the IEEE Computer Society, Association for Computing Machinery, Association for the Advancement of Artificial Intelligence, and Association for Computational Linguistics. He served as an Associate Editor in the “Social Networks, Web 2.0, and Beyond” track in International Conference on Information Systems 2009 and a program committee member of Pacific Asia Workshop on Intelligence and Security Informatics 2009 and Workshop on the Sciences of the Artificial 2005.