R.Rajasekaran *et al*, International Journal of Computer Science and Mobile Applications,
Vol.2 Issue. 1, January- 2014, pg. 180-188 **ISSN: 2321-8363**

COLLECTION OF WEB OPINION INFORMATION AND VISUALIZATION BASED ON SDC

Mr.R.Rajasekaran[1], Mr.S.Rajesh[2]

[1]PG Student, [2]Assistant Professor, Department of Computer Science and Engineering,

PRIST University, Trichy District, India

([1] rajaravi.90@gmail.com)

Abstract

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 project, investigated the density-based clustering algorithm and proposed the scalable distance-based clustering technique for Web opinion clustering. This Web opinion clustering technique enables the identification of themes within discussions in Web social networks and their development, as well as the interactions of active participants. This 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.

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 .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. 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.

©2014, IJCSMA All Rights Reserved, www.ijcsma.com

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. 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.

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. The framework of the overall project 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. This special properties of Web opinions that do not exist in regular documents include the following: 1) the content of messages is less focused; 2) the messages are usually short in length ranging from a few words to a couple paragraphs; 3) the terms used in the messages are sparse because different users may use different terms to discuss the same topic; 4) the messages contain many unknown terms or slang that do not exist in typical dictionary or ontology, e.g., iPhone and Xbox; 5) there are many noises like unrelated text or typographical error so that many Web opinions do not fall into any categories; 6) the volume of 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. s

©2014, IJCSMA All Rights Reserved, www.ijcsma.com

Given a collection of documents *D* document clustering techniques identify a set of clusters *C* and assign a Boolean value to each pair (*di*, *cj*), where *D = {d*1, *d*2, . . . , *d|D|}* and *C ={c*1, *c*2, . . . ,*c|C|}*. The Boolean value assigned to each (*di*, s*cj)* determines whether *di* is assigned to *cj* . However, the set of clusters is not predefined in the setting of Web opinions because the topics of discussion are always evolving and usually not known in advance. Therefore, the cluster analysis in this paper employs the unsupervised learning approach in which the set of clusters is not predefined and samples of documents for each cluster are not available.

Specifically, we propose a scalable distance-based algorithm for clustering Web opinions. 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.

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 and *EM* 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, 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. Similarity between short texts will then be measured by the similarity between their corresponding context vectors. Such measurement of word similarity can be utilized to categorize short texts in which the words in common are less. The Wikipedia concepts are the titles of the matched Wikipedia articles. It showed that it improved the result of clustering substantially in most cases.

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. 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, 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

©2014, IJCSMA All Rights Reserved, www.ijcsma.com

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.

II LITERATURE SURVEY

2.1 Statement of project

Evolution of Blogspace

The culture of blog space focuses on local community interactions among a small number of bloggers, from, say, three to 20. Members of such an informal community might list one another's blogs in a "blogroll" (a sidebar within a particular blog listing the other blogs the blogger frequents) and might read, link to, and respond to content in other community members' blogs. This sequence of responses often take place during a brief burst of activity as an interesting topic arises, jumps to Prominence then recedes. We observed and modeled this highly dynamic, temporal community Structure in order to reveal the evolution of blogs pace over time. To do so, we considered each blogger as more than a static object, extending our view of the individual blog to include a temporal component, reflecting the fact that blog entries are posted over time. It is difficult to capture the particular topics covered by each entry, as the entries lack structure; even the definition of a topic is subjective. However, we've observed that bloggers in a community often link to and cross-reference one another's postings, so we can infer community structure by analyzing the linkage patterns among blog entries. In this view of the worldwide blogging network, a "community" is a set of blogs linking back and forth to one another's postings while discussing common topics. Each community may exhibit different levels of activity over particular periods of time; for example, a community may show a burst of rapid-fire discussion during a three-week period, then lie dormant for several more weeks before the next burst of activity.

Locality Models of Evolution

Our goal is to model how the communication structure evolves, and equilibrates to its observed steady state. Our model specifies how nodes update their edges in response to the observed communication activity. In specifying this model of evolution, we take as given the out-degree distribution of the Blogograph.

The justification for this is that, while the outdegree distribution would be an interesting object to model, it mainly reflects the individual properties of the users in the network (the level of energy and involvement of the user). Such quantities tend to be innate to a user (different people have different social habits: some manage to communicate with hundreds of people while

©2014, IJCSMA All Rights Reserved, www.ijcsma.com

others interact with only a small group), and hence out-degrees should be specified either *ab initio* (e.g. from social science theory) or extracted directly from the observed data. We will take the latter approach to specifying the out-degree distribution when it comes to testing our model. Given the out-degrees for all nodes, the task is now to specify how to attach the out-edges of the nodes, and in particular, to obtain the in-degree distribution. It is the indegree distribution that characterizes the global communication structure of the network (for example, who is considered by others to be important). Clearly, the out-degree distribution of a graph alone does not determine its in-degree distribution. Algorithms for generating undirected random graphs with a prescribed degree distribution However, even if those algorithms can be expanded to the domain of directed graphs, they would still be insufficient for our purpose of modeling evolution which requires repeated generation of the next graph given the previous one.

III - EXISTING SYSTEM

3.1 Overview

Search engines could personalize their results to match not only a person's stated interests, but also the interests inferred from the user's social network. Knowing a person's social network would allow one to infer what likes and dislikes a person may have, what advertisements they may be more likely to take note of, etc.

It could also lead to more intelligent chat clients that, for example, recommend a new friend to join a chat based on the interests shared by the friends already chatting. But I couldn't able to correlate the personal behavior from social networks.

3.2 Disadvantages

1. Cannot predefine or predict the number of clusters.
2. Poor performance of Web opinion clustering when document clustering techniques are applied directly.
3. Clustering short texts, relying on external sources is not feasible in many cases.
4. Huge number of Web opinions in the Web, it is inefficient and costly.

IV PROPOSED SYSTEM

4.1 Overview

In order to analyze the relation between communication and personal behavior, we need two sources of data:

(1) who communicates with whom, and

(2) the characteristics of each person in the communication network.

For the first, we use an instant messaging network, and for the second, we use data from people's search history and their demographics. In this project we apply data mining techniques to study this relationship. We also present Web opinion clustering and information visualization

©2014, IJCSMA All Rights Reserved, www.ijcsma.com

techniques, which are components of an ongoing project of Web opinion analysis and understanding

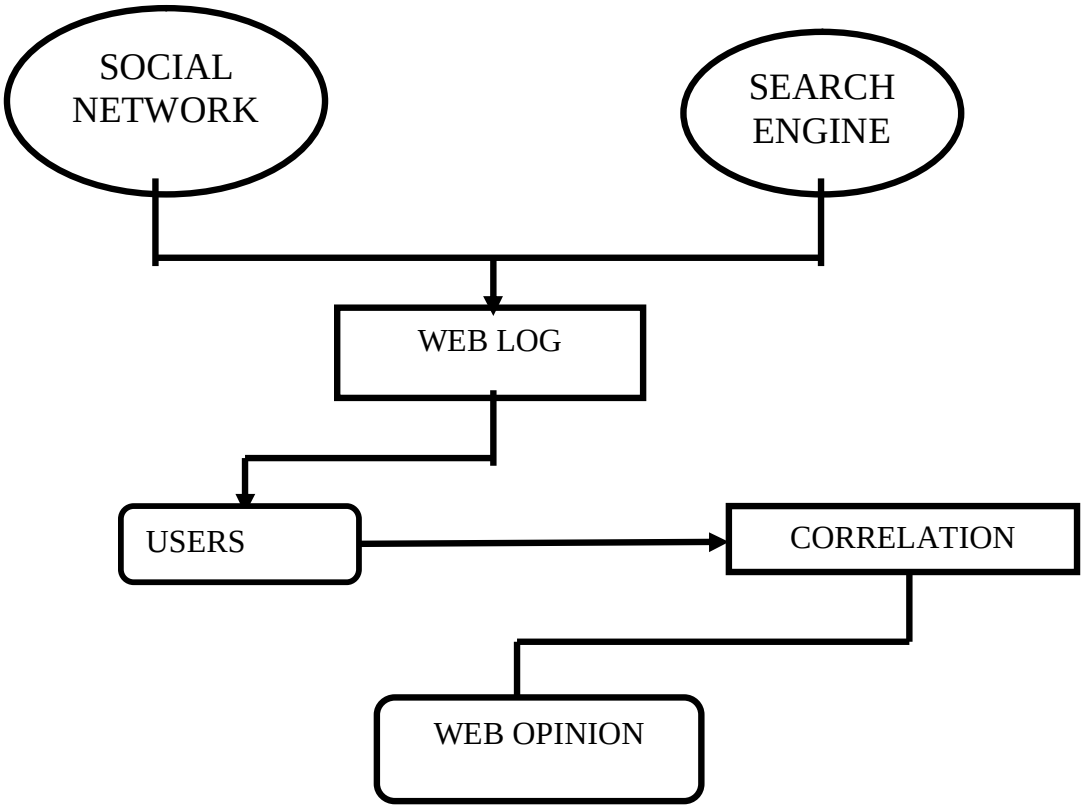
4.2 Advantages

The overall clustering result is to provide a high level content summarization of the underlying threads in forums.

It may be useful for them to identify other participants whom they have never interacted with but share with similar ideologies.

Clustering web opinions due to the fast growing number of Web opinions in social media.

V DATA FLOW DIAGRAM



©2014, IJCSMA All Rights Reserved, www.ijcsma.com

VI MODULES

6.1 MODULE DESCRIPTION

1. Social Network

Chat Server:

A Chat Server contains all the information of the client such as client name, logon time of each client and the messages that are transferred between the clients and also it contains the user logout time.

Chat Client:

Client chat Application can be used easily by all users to sends ans Instant messages to single client or multiple clients. This allows the user to use the similes to share his/her feelings in a swift and humorous way to the clients. Also it displays the logout time of the other clients.

Store chat sessions for later retrieval:

Messages are stored in a database for later retrieval. If a connection to the database isn't possible or corrupt, the recorder will soar messages to a local media (local hard drive). Next time the connection to the database is present, the locally stores messages will be sent to the database.

2. Web Crawler

In this module, we built the local web crawler used to gather specific types of information from Web pages.

3. Interactive Information Visualization

In this module we built the interactive information visualization tool. In that tool, the DBSCAN is a density-based cluster algorithm that can discover the clusters and filter the noise into a datasets. The words found in forum messages are relatively noisy because the content usually consists of no edited and conversation-like material.

4. Content Clustering Module

In this module, input the log from Social Network and Web Crawler, apply DBSCAN that can discover the clusters from the log and filter the noise.

5. Correlation-An Analysis

Social Network Data

In this module, we used data on user interactions in the MSN Messenger network, for a period of time. The raw data logged each event on the network, such as joining a chat session, chat invites, leaving a chat session etc., along with corresponding time-stamps. We obtained the raw data, which we processed to extract out relevant attributes.

Aggregated Messenger Session

©2014, IJCSMA All Rights Reserved, www.ijcsma.com

userid1 userid2 #sessions #sent1 #sent2 duration #sessions denotes the total number of individual sessions involved in the aggregate. #sent1, #sent2 denote the total numbers of messages sent by each user aggregated over all the sessions. Duration is the total duration of all the sessions combined.

Personal Interests Data

The module is based on personal interests. For this, we turned to Web search behavior sa sampling of these searches demonstrates that users reveal personal interests and information through what they search for. For any interest that an Internet user has, it is very likely that he/she has at some point used a Web search engine to learn more about it. This makes the search engine query logs an ideal source of information about users' personal interests and behavior.

Aggregated Search Session

Userid query list main-category list sub-category list age group gender zip

Joining the Data

Once we obtained the messenger data and the search data, they were joined together into one dataset where each tuple has the information about the aggregated messenger session as well as the searches for each user in the pair. This was simply done by scanning through the aggregated messenger data and appending to each tuple the aggregated search entries for the corresponding user IDs from the search data.

CONCLUSION

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 micro accuracy and macro accuracy. 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.

©2014, IJCSMA All Rights Reserved, www.ijcsma.com

REFERENCES

- S. Banerjee, K. Ramanathan, and A. Gupta, "Clustering short texts using wikipedia," in *Proc. ACM SIGIR*, Amsterdam, The Netherlands, 2007,pp. 787–788.
- B. Bicici and D. Yuret, "Locally scaled density based clustering," in *Proc. ICANNGA*, 2007, pp. 739–748.
- 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.

©2014, IJCSMA All Rights Reserved, www.ijcsma.com