The Analysis of Online Communities using Interactive Content-based Social Networks

Anatoliy Gruzd

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, agruzd2@uiuc.edu

Caroline Haythornthwaite

Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign, haythorn@uiuc.edu

Introduction

Today the Internet has become a convenient and ubiquitous platform for anyone with access to publish their thoughts and ideas, express their opinions, argue with their peers on various issues and, most importantly, organize and form online communities. Although video and image formats, and 3D environments are rapidly gaining popularity, the majority of user-generated discussions on the Internet are still text-based. Many of these discussions are archived and readily available to organizers, developers and researchers of online communities. Being able to evaluate the internal processes of such communities is important for helping users seeking to join a community, managers or instructors hoping to facilitate participation and discussion in communities, and to researchers exploring the nature of social processes online. However, as yet, there is still no easy or cost effective method to study and analyze this ever growing mountain of textual data in real-time. This presentation describes a new, automated procedure and a prototype of our web application for making sense of online activities in real time.

To study online communities via their textual exchanges, it is important to know (1) who is talking to whom, and (2) what they are talking about. To address the first part, researchers usually rely on social network analysis. To address the second part, researchers rely on some form of text analysis. Traditionally, these two methods of analysis are conducted and studied independently. However, research on online communities will greatly benefit if these two types of analysis can be merged to form one comprehensive and coherent method for studying online communities. This work presents a combined method of analysis that takes advantage of the strengths of each of these two types of analysis.

The test case for the evaluation of this combined approach consists of threaded discussions of online conversations, and specifically those that occur within bulletin boards of online classes. The analysis technique is used to reveal the social networks within these e-learning communities and relies heavily on automated discovery and analysis of the online conversations. The research questions addressed in this work are:

- Can Natural Language Processing aid researchers (and members of online communities) to analyze and visualize online social networks from just the textual data found in online threaded discussion postings?
- What syntactic and semantic features found within online threaded discussion postings help to uncover explicit and implicit ties between network members, and can these features be used to provide a reliable estimate of the strengths of interpersonal ties among the network members?

Method

As a way to address these research questions, we proposed a content-based approach for inferring social networks from postings in threaded discussions, dubbed 'name network'. The approach starts by finding all mentions of personal names in the postings and uses them as nodes in the name network. Once all the nodes are identified, the next step is to discover how these names/nodes are connected to each other in order to derive 'who talks to whom' data. To accomplish this, the algorithm works under the assumption that the chance of two people sharing a social tie is proportional to the number of times each of them mentions the other in his/her postings either as an addressee or a subject person. As a way to quantify this assumption, the algorithm adds a nominal weight of 1 to a tie between a poster and all names found in the postings. After processing all postings, only those ties that have weights higher than a pre-defined threshold (to be determined experimentally) are included as part of the name network. And, finally, to make the name network better reflect e-learning processes, tie strength is assigned based on pre-defined relations that have shown to predict success in e-learning communities such as Information Exchange. (See Gruzd & Haythornthwaite (2008) for a more detailed description of this algorithm.)

To evaluate the content-based method of building social networks, and to identify what is gained from using this more elaborate method, social networks derived using this method will be compared against those derived from other means, specifically (1) traditional 'who *replies* to whom' networks and (2) members' perceived (self-reported) social networks. The traditional 'who *replies* to whom' network data will be built automatically using students' posting behaviors, and the self-reported social networks data will be collected via online questionnaire from six online courses that are participating in the study.

The 'name network' method as proposed and evaluated in this work provides one more option for understanding and extracting social interaction networks from online discussion boards. The preliminary results demonstrate that name networks address some of the shortcoming of traditional 'who replies to whom' networks. But more research is needed to test the generalizability of the 'name network' method with regards to datasets from other domains or genres.

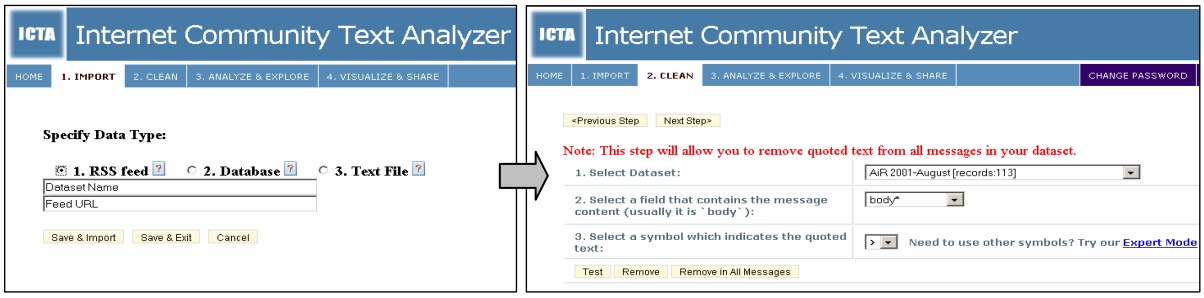
We expect that name networks will be a useful diagnostic tool for instructors to evaluate and improve lesson plans, and to identify students who might need additional help or students who may provide such help to others. This is possible because of two important features associated with the 'name network' method. First, name networks take into account only those messages that contain personal references to others in a group. These messages tend to be more interactive and argumentative and as a result are considered to be good indicators of collaborative learning. Second, by operationalizing and measuring information exchange for the purpose of assigning tie strengths, the method increases weights for postings that are believed to be better contributors to shared knowledge construction.

Web Application

As part of this work, a web-based system for content and network analysis called the Internet Community Text Analyzer (ICTA) is being developed. The main goal is to provide researchers and other interested parties with an automated system for analyzing text-based communal interactions with the help of various interactive visualizations. ICTA's web-based architecture will stimulate collaborative research by allowing researchers to access and analyze datasets remotely from anyplace where there is web access and share their preliminary results with their collaborators. Another benefit of a web-based software implementation is the ability to outsource data processing and have all the heavy computing be done on a speedier remote server. For example, once data has been entered on a stand-alone website, it can then be sent to ICTA to be analyzed in real-time, and then immediately returned and presented using useful visualizations to a community's web space.

In its current state, ICTA is a prototype designed to test and evaluate the effectiveness of different text mining and social network discovery techniques. The goal at this stage is to identify a range of optimal values for various parameters that control automated procedures. Eventually, these optimal values will be used as default settings in a future simplified single-step "one-button" version of ICTA. Below is a brief description of ICTA's current multi-steps interface and functionalities.

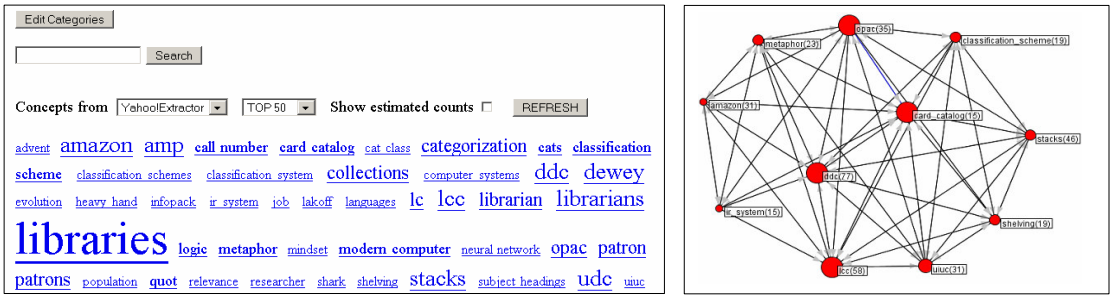
First, a user starts by importing a dataset. To do this, he/she can upload a file or specify the location of an external repository (See Figure 1a). Currently, ICTA can parse email-based or forum-based online communication that has been stored in one of three data formats: XML (e.g., RSS feeds), MySQL database or CVS text file. After the data is imported, the second step is to remove any text that may be considered as noise (See Figure 1b). This is an optional step that is primarily designed to remove redundant or duplicate text that have been carried forward from prior messages. To accomplish this, ICTA simply removes all lines that start with a symbol commonly indicating quotation such as ">" or ":". But a user is not restricted to just these two symbols. In fact, in 'Expert Mode", it is possible to remove almost any text patterns such as URLs or email addresses from messages using a mechanism called *regular expression*.



(a) **(b)**

Figure 1. Importing (a) & Cleansing (b) Dataset

After the data importing and cleansing steps are completed, the data is then ready to be analyzed. In this stage, ICTA will build very concise summaries of the communal textual discourse. This is done by extracting the most descriptive terms (usually nouns and noun phrases) and presenting them in the form of interactive concept clouds and semantic maps (See Figure 2). With a summary in hand, a researcher or a member of an online group can quickly identify emerging community interests and priorities as well as patterns of language and interaction that characterize a community. (See Haythornthwaite & Gruzd (2007) for more details on this type of text analysis and some preliminary results.)



(a) **(b)**

Figure 2. An example of a Concept Cloud (a) and a Semantic Map (b)

The second part of the analysis stage consists of building two types of social networks: (1) the 'who replies to whom' network and (2) the name network as described above. When building these networks from interaction data, there are a lot of different parameters and thresholds choices to select from. To find the most optimal configuration for a particular type of datasets, ICTA's interface allow us to fine tune many of the available parameters and thresholds. For example, one of the choices that are likely to influence network formation is how to estimate tie strengths between individuals. ICTA provides a range of options for doing this estimation: from a simple count of the number of messages exchanged between individuals to an estimation based on the amount of information exchanged between individuals. Another important parameter that is also likely to influence network formation is the decision whether to include or exclude particular types of messages from the analysis. This is especially important because different messages may expose different types of relations between community members. With ICTA, a user can quickly decide what messages from the dataset can be included or excluded from a particular analysis. For example, if a relation being studied is *agreement*, then a researcher might want to decide to ignore all messages that are neutral in nature and keep only those that suggest agreement or disagreement. This way, it becomes a lot easier to make assertions about the quality of tie strengths, interpret their values, and study group interactions from competitive or complimentary perspectives. Each relation is described and stored in the system as a set of linguistic markers (e.g., words, phrases, patterns). A relation can be defined manually using ICTA's interface or selected from a list of pre-defined ones. A list of pre-defined relations include linguist markers shown to be useful in the literature when identifying instances of cognitive and meta-cognitive processes such as decision-making, problem-soling, questions-answering, etc (see, for example, Alpers et al., 2005; Corich et al., 2006; Pennebaker & Graybeal, 2001).

In the final part of the analysis**,** networks are visualized using a proprietary Java application based on JUNG (Java Universal Network/Graph Framework), a set of java libraries for drawing and manipulating graphs (http://jung.sourceforge.net). In addition to a number of basic visualization features such as scaling, changing graph layouts, selecting cut off points to hide "weak" nodes or ties, ICTA can also display excerpts from messages exchanged between two individuals to show the context of their relations. This ability to call up and display excerpts from messages makes it a lot easier to "read" a network and understand why a particular tie exists. This feature is activated by simply moving a mouse over an edge connecting two nodes (see Figure 3 below). ICTA is also capable of simultaneously displaying two different types of networks of the same group on the same graph using different colors to display edges from different networks. The latter makes it easier to study the quality of and differences/similarities between different networks.

When fully developed, ICTA will become a general purpose tool for automated real-time analysis of the interaction patterns within online communities. It may also be used by e-learning communities to improve the e-learning experience for both instructors and students, as well as by those examining other online communities. For example, faculty and administration will be able to use ICTA to gain insight into class interactions about the online learning processes in their classes, and to develop more appropriate and effective strategies for the next generation of students.

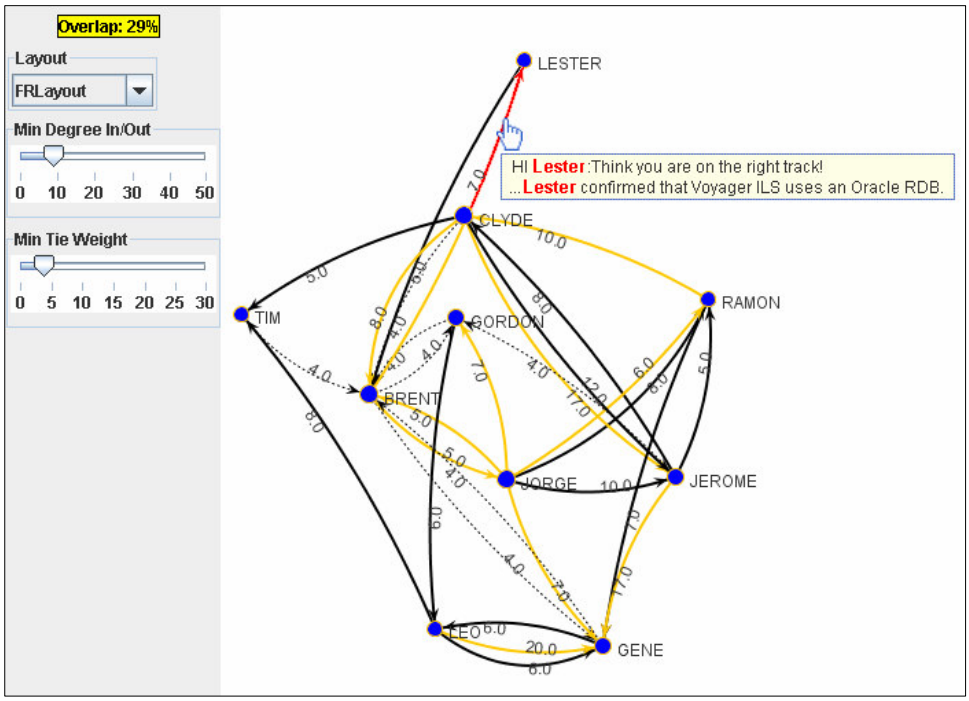


Figure 3. Social Network Visualization

References

Alpers, G.W., Winzelberg, A.J., Classen, C., Roberts, H., Dev, P., Koopman, C. and Barr Taylor, C. (2005). Evaluation of Computerized Text Analysis in an Internet Breast Cancer Support Group. *Computers in Human Behavior, 21*(2), 361-376.

Corich, S., Kinshuk and Hunt, L.M. (2006). Measuring Critical Thinking within Discussion Forums Using a Computerised Content Analysis Tool. In the Proceedings of *Networked Learning*.

Gruzd, A. and Haythornthwaite, C. (2008). Automated Discovery and Analysis of Social Networks from Threaded Discussions. *Paper presented at the International Network of Social Network Analysts*, St. Pete Beach, FL, USA.

Haythornthwaite, C. and Gruzd, A.A. (2007). A Noun Phrase Analysis Tool for Mining Online Community. In the Proceedings of *the 3rd International Conference on Communities and Technologies*, Michigan State University.

Pennebaker, J.W. and Graybeal, A. (2001). Patterns of Natural Language Use: Disclosure, Personality, and Social Integration. *Current Directions in Psychological Science*, *10*(3), 90-93.