Mining Chat Discussions

Stanley Loh
Catholic University of Pelotas, Brazil
Lutheran University of Brazil, Brazil

Daniel Licthnow
Catholic University of Pelotas, Brazil

Thyago Borges
Catholic University of Pelotas, Brazil

Tiago Primo
Catholic University of Pelotas, Brazil

Rodrigo Branco Kickhöfel
Catholic University of Pelotas, Brazil

Gabriel Simões
Catholic University of Pelotas, Brazil

Gustavo Piltcher
Catholic University of Pelotas, Brazil

Ramiro Saldaña
Catholic University of Pelotas, Brazil

INTRODUCTION

According to Nonaka & Takeuchi (1995), the majority of the organizational knowledge comes from interactions between people. People tend to reuse solutions from other persons in order to gain productivity.

When people communicate to exchange information or acquire knowledge, the process is named *Collaboration*. Collaboration is one of the most important tasks for innovation and competitive advantage within *learning organizations* (Senge, 2001). It is important to record knowledge to later reuse and analysis. If knowledge is not adequately recorded, organized and retrieved, the consequence is re-work, low productivity and lost of opportunities.

Collaboration may be realized through synchronous interactions (e.g., exchange of messages in a chat), asynchronous interactions (e.g., electronic mailing lists or forums), direct contact (e.g., two persons talking) or indirect contact (when someone stores knowledge and others can retrieve this knowledge in a remote place or time).

In special, chat rooms are becoming important tools for collaboration among people and knowledge exchange. Intelligent software systems may be integrated into chat rooms in order to help people in this collaboration task. For example, systems can identify the theme being discussed and then offer new information or can remember people of existing information sources. This kind of systems is named recommender systems.

Furthermore, chat sessions have implicit knowledge about what the participants know and how they are viewing the world. Analyzing chat discussions allows understanding what people are looking for and how people collaborates one with each other. Intelligent software systems can analyze discussions in chats to extract knowledge about the group or about the subject being discussed.

Mining tools can analyze chat discussions to understand what is being discussed and help people. For example, a recommender system can analyze textual messages posted in a web chat, identify the subject of the discussion and then look for items stored in a Digital Library to recommend individually to each participant of the discussion. Items can be electronic documents, web pages and bibliographic references stored in a digital library, past discussions and authorities (people with expertise in the subject being discussed). Besides that, mining tools can analyze the whole discussion to map the knowledge exchanged among the chat participants.

The benefits of such technology include supporting learning environments, knowledge management efforts within organizations, advertisement and support to decisions.

BACKGROUND

Some works has investigated the analysis of online discussions. Brutlag and Meek (2000) have studied the identification of themes in e-mails. The work compares the identification by analyzing only the subject of the e-mails against analyzing the message bodies. One conclusion is that e-mail headers perform so well

as message bodies, with the additional advantage of reducing the number of features to be analyzed.

Busemann et al. (2000) investigated the special case of messages registered in call centers. The work proved possible to identify themes in this kind of message, although the informality of the language used in the messages. This informality causes mistakes due to jargons, misspellings and grammatical inaccuracy.

The work of Durbin et al. (2003) has shown possible to identify affective opinions about products and services in e-mails sent by customers, in order to alert responsible people or to evaluate the organization and customers' satisfaction. Furthermore, the work identifies the intensity of the rating, allowing the separation of moderate or intensive opinions.

Tong (2001) investigated the analysis of online discussions about movies. Messages represent comments about movies. This work proved to be feasible to find positive and negative opinions, by analyzing key or cue words. Furthermore, the work also extracts information about the movies, like directors and actors, and then examines opinions about these particular characteristics.

The only work found in the scientific literature that analyzes chat messages is the one from Kahn et al. (2002). They apply mining techniques over chat messages in order to find social interactions among people. The goal is to find who is related to whom inside a specific area, by analyzing the exchange of messages in a chat and the subject of the discussion.

MAIN THRUST

Following, the chapter explains how messages can be mined, how recommendations can be made and how the whole discussion (an entire chat session) can be analyzed.

Identifying Themes in Chat Messages

To provide people with useful information during a collaboration session, the system has to identify what is being discussed. Textual messages sent by the users in the chat can be analyzed for this purpose. Texts can lead to the identification of the subject discussed because the words and the grammar present in the texts represent knowledge from people, expressed in written formats (Sowa, 2000).

An ontology or thesaurus can be used to help to identify cue words for each subject. The ontology or thesaurus has concepts of a domain or knowledge area, including relations between concepts and the terms used in written languages to express these concepts (Gilchrist, 2003). The ontology can be created by machine learning methods (supervised learning), where human experts select training cases for each subject (for example, texts of positive and negative examples) and an intelligent software system identifies the keywords that define each subject. The TFIDF method from Salton & McGill (1983) is the most used in this kind of task.

If considering that the terms that compose the messages compose a bag of words (have no difference in importance), probabilistic techniques can be used to identify the subject. By other side, natural language processing techniques can identify syntactic elements and relations, then supporting more precise subject identification.

The identification of themes should consider the context of the messages to determine if the concept identified is really present in the discussion. A group of messages is better to infer the subject than a single message. That avoids misunderstandings due to words ambiguity and use of synonyms.

Making Recommendations in a Chat Discussion

A recommender system is a software whose main goal is to aid in the social collaborative process of indicating or receiving indications (Resnick & Varian, 1997). Recommender systems are broadly used in electronic commerce for suggesting products or providing information about products and services, helping people to decide in the shopping process (Lawrence et al., 2001) (Schafer et al., 2001). The offered gain is that people do not need to request recommendation or to perform a query over an information base, but the system decides what and when to suggest. The recommendation is usually based on user profiles and reuse of solutions.

When a subject is identified in a message, the recommender searches for items classified in this subject. Items can come from different databases. For example, a Digital Library may provide electronic documents, links to Web pages and bibliographic references.

A profile database may contain information about people, including the interest areas of each person, as well an associated degree, informing the user's