# Recommendation of Complementary Material during Chat Discussions

Stanley Loh*

Lutheran University of Brazil and Catholic University of Pelotas
Lutheran University of Brazil, Department of Computer Science
Av. Farroupilha, 8001, Canoas - RS, 92425-900, Brazil
Catholic University of Pelotas, Centro Politécnico
R. Félix da Cunha, 412, Pelotas - RS, 96010-000, Brazil
E-mail: sloh@terra.com.br

Daniel Lichtnow

Federal University of Rio Grande do Sul and
Catholic University of Pelotas
Federal University of Rio Grande do Sul
Instituto de Informática
Av. Bento Gonçalves, 9500
Caixa Postal 15064,
Porto Alegre – RS, 91501-970, Brazil
Catholic University of Pelotas, Centro Politécnico
R. Félix da Cunha, 412, Pelotas - RS, 96010-000, Brazil
E-mail: dlichtnow@inf.ufrgs.br

Adriana Justin Cerveira Kampff

Lutheran University of Brazil and Colégio Nossa Senhora do Rosário
Lutheran University of Brazil, Department of Computer Science
Av. Farroupilha, 8001, Canoas - RS, 92425-900, Brazil
E-mail: adriana@maristas.org.br

José Palazzo Moreira de Oliveira

Federal University of Rio Grande do Sul
Instituto de Informática
Av. Bento Gonçalves, 9500
Caixa Postal 15064,
Porto Alegre – RS, 91501-970, Brazil
E-mail: palazzo@inf.ufrgs.br

*Corresponding author

**Abstract:** In the context of Internet, there are many tools that allow sharing knowledge. Examples of these tools are Web chats. However it is possible to use Web chats in a more effective way. In this sense, this paper presents a system that analyzes the themes discussed in a chat room and then recommends information sources according to the context of the discussion. In order to produce recommendations, the system considers users' profiles to complement the knowledge of each individual, reaching what Vygotsky called zone of proximal development. Another important feature is related to the fact that, after the chat discussion session, it is possible to generate statistical analyses. These analyses allow evaluating the discussion (e.g. how many different subjects were discussed, discussion deviate) and thus the knowledge of the whole community and of each member (e.g. about what subject a participant is talking). The system uses text mining techniques to identify the themes discussed in the chat room.

**Keywords:** Chat, Recommendation, Recommender System, Web-based Learning, Knowledge Sharing.

**Biographical notes**: Stanley Loh is a Professor in the Catholic University of Pelotas and in the Lutheran University of Brazil. He has a Ph.D. degree in Computer Science. His interests include recommendation systems, text mining, and knowledge management.

Daniel Lichtnow is a Ph.D. student of Computer Science at Federal University of Rio Grande do Sul - UFRGS. He is lecturer and researcher at the Catholic University of Pelotas (UCPEL). He obtained a Master degree in Computer Science at the Federal University of Santa Catarina (UFSC) in 2001. His interests include databases systems, text mining, knowledge management, recommender systems and Web information quality.

Adriana Justin Cerveira Kampff is lecturer and researcher at the Lutheran University of Brasil (ULBRA). She has a PhD in Computer Science obtained at the Federal University of Rio Grande do Sul (UFRGS) in 2009. Her interests include computer-supported learning and teaching, adaptive systems and distance education.

José Palazzo Moreira de Oliveira is full professor of Computer Science at Federal University of Rio Grande do Sul - UFRGS. He as a doctor degree in Computer Science from Institut National Politechnique - IMAG (1984), Grenoble, France, a M.Sc. degree in Computer Science from PPGC-UFRGS (1976) and has graduated in Electronic Engineering (1968). His research interests include information systems, e-learning, database systems and applications, conceptual modeling and ontologies, applications of database technology and distributed systems. He has published about 160 papers, has being advisor of 11 Ph.D. and 51 M.Sc. students.

## 1. Introduction

The number of technology-based environments that support knowledge sharing is growing up very fast. In the context of the World Wide Web, such environments enable the raising of Virtual Learning Communities, that gather people geographically distant but with similar interests. People in these communities exchange knowledge, documents, bibliographic references and other information sources about similar topics. People usually do that using digital libraries (indirect communication) or online discussions (as in forums and chats). As a good consequence, people can share knowledge and complement individual experiences.

This paper presents a recommender system for online discussions. The system consists in a web chat, where users exchange messages. The textual messages posted to the chat are analyzed so that complementary information can be recommended during the chat session according to the topics being discussed. Recommendations are personalized to each user's profile. Doing that, the system stimulates the extension of the user's knowledge, enabling the collective learning (social or interpersonal level) and the personalized learning (individual or intrapersonal level). Recommendations include electronic documents and links to web pages, stored in a private digital library. The system also recommends past discussion sessions stored in historical records and people with special expertise in the discussion topics. The system is based on a social-interactionist approach. The socio-historical theory from Vygotsky (1984) characterizes the knowledge elaboration as a collective construction. People internalize and generate new knowledge from interactive acts, within a social and cultural context. According to Vygotsky (1984), the social interaction is the base of the learning; the intellectual development appears first in the social level (interpersonal) and then in the individual level (intrapersonal). In this point of view, the communication process is fundamental to concretize the learning, and the language is a mediator sign that allows people to analyze, abstract and generalize, consequently, establishing and categorizing concepts.

Vygotsky (1984) states that each person has a real knowledge level (what he/she dominates) and a potential knowledge (what he/she can do with the help of others). The difference between these two levels is called Zone of Proximal Development. The proposed system intends to extend the zones of each person, acting as mediator in the discussion, suggesting complementary sources and thus enabling knowledge extension.

## 2. Related Work

A recommender system is a software system whose main goal is to aid in the social collaborative process of indicating or receiving indication, when the number of options is huge (Resnick & Varian, 1997). Recommender systems are proactive devices and their goal is to supply people with information useful for decision making. This information may be about books, documents, music, restaurants and whatever (Resnick & Varian, 1997). Recommender systems are successfully used in commerce and for merchandising, allowing companies to offer products, services and information to help customer in the decision. This kind of system is especially useful when there are many options to choose and users have little information about those options. The great benefit is that recommender systems can supply information without people having to search, query or search for it.

Lawrence, Almasi, Kotlyar, Viveros, and Duri (2001), Schafer, Konstan, and Riedl (2004) and Srivastava, Cooley, Deshpande, and Tan (2000) discuss many applications of recommender systems to commerce and marketing in general. There are works on recommendation in communication environments. For example, Terveen and Hill (2001) discuss the PHOAKS system, which extracts addresses of Web pages from messages in the Usenet newsgroup for future recommendations. Other system, proposed by Viegas and Donath (1999) apud Terveen and Hill (2001) analyzes messages in the Usenet intending to later recommend group of messages according to some attributes (for example, presence of certain themes or discussions with greater number of participations).

Recommender systems are becoming an important alternative to support knowledge acquisition. GroupLens system uses collaborative filters to help people to find useful information (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994). The technique collects user's feedback to select new articles that can be interesting to the user. Walker, Recker, Lawless, and Wiley (2004) presents the Altered Vista system, whose aim is provide recommendations of Web resources based on reviews from teachers and students. Users submit reviews about the quality and usefulness of Web resources and these ratings become part of the recommendation database. After that, users can search the recommendations of other users or can request personalized recommendations from the system, thus avoiding less useful sites.

In the specific case of chat communications, some works analyze the use of chats in learning processes. Mock (2001) studied many tools to support communication in the classroom. The work found that "*student participation was generally low unless the students were either motivated or were given an explicit assignment using the tool*." The chat tool does not ease the organization of the students together simultaneously. By other side, instant messaging tools were favored over online chat rooms due to their asynchronous nature and notification features. However, these conclusions date of 2001 and certainly a lot of changes occurred since then. Lonchamp (2005) found similar conclusions. Standard chat tools suffer from important coordination and coherence deficiencies. The solution proposed in that paper is a generic framework for building structured chat applications, providing new ways of controlling chat sessions: end users can strengthen (or relax) constraints when it becomes necessary during a chat session. In the educational context, this means that teachers can control many parameters such as the content of the turns, the flow of turns (with a predefined protocol), who are the participants and their roles, solving problems such as lack of participation, "flying fingers" domination, control of disturbing persons, etc. Mu, Marchionini, and Pattee (2003) present Smartlink, a new concept for synchronizing various information components or "channels" (video player with storyboard, shared browser and text chat room). The conclusions point that the combination of three information presentation components provided an effective and more comfortable collaboration and learning environment.

There are works that analyze chat messages. Khan, Fisher, Shuler, Wu, and Pottenger (2002) 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. Anjewierden, Kolloffel, and Hulshof (2007) presents a tool that provide feedback and guidance to learners about the nature and patterns of their communication in a chat room. To achieve this, the tool identifies different types of messages, for example, transformative (domain), regulative, technical, and off-task (social) messages. Holmer (2008) presents the DSA system that analysis the log of messages exchanged during a chat in order to study the discourse structure. The structure is created by analyzing references among messages, forming branched threads. After that, some metrics are applied, as for example complexity and gaps in the structure, and metrics about participants' individual behavior and about social interaction. Neuage (2005) presents seven case studies concerning dialogues from chat rooms. These studies analyze features peculiar to on-line chat, demonstrating that chat "*texted talk*" combines face-to-face chat with text-based communication. Wu, Khan, Fisher, Shuler, and Pottenger (2002) presents a text mining tool for analysis of chat-room conversations, in order to answer questions such as "what topics are being discussed", "who is discussing which topics" and "who is interacting with whom".

Although the quality of the referenced works, they do not provide recommendation of contents during the discussion in a chat session. The basic contribution of the system proposed in this paper is to recommend information sources stored in a digital library, in real time during chat sessions and according to the context of the themes discussed in the chat. The extended contribution of the system is to support learning, acting as mediator in the discussion, suggesting complementary sources and thus enabling knowledge extension.

## 3. Description of the Proposed System

The goal of this recommendation system is to provide people with useful information during a collaboration session. To do that, the system analyzes textual messages sent by users when interacting in a private Web chat, identifies topics (subjects/themes/concepts) inside the messages and recommends items catalogued in a private database, previously classified in the same topics. Figure 1 presents the architecture of the system with its main components.

The text mining module analyzes each message posted to the chat. The words present in the message are compared against terms present in a domain ontology. After that, it passes the identified concept to the recommender module, that searches in the database for items to suggest. The database is composed by:

1. A Digital Library, containing electronic documents, Web links and bibliographic references;
2. A base of Past Discussions, containing historical discussions; and
3. A Profile base, containing registered users with their profiles.

According to the classification of Terveen and Hill (2001), the system is a content-based recommender system because the context of the messages is matched against the content of items in the database.

One difference of the proposed system from others is that it is not necessary to store a profile for a user to use the system and receive recommendations. Messages sent by users are enough for the system deciding what recommend. The profile stored in the database is only used to improve recommendations, but it is not a necessary feature. If a profile exists associated to a user, this user will receive personalized recommendations during the discussion and that will stimulate the development of new knowledge, according to the theory about the Zone of Proximal Development.

Figure 2 shows a snapshot of the system in a real use. There is an area where the users logged in the chat appear (*users*), an area where the messages can be viewed (*messages*), an area where the recommendations appear individually for each user (recommendations – topics and documents) and an other area for the user writing the messages (*your message*).

In the next sections, each component of the system is described in details.

### 3.1. The Web Chat

The chat works like traditional chats over the Web. The difference is that it is specially constructed for the proposed system and it is not open to non-registered users. Thus, users have to be authenticated for using the system. There is no limit for the number of persons interacting at the same time.

At the moment, only one chat channel is allowed. Thus a discussion session concerns all messages sent during a day. In the future, this restriction will be eliminated.

## 3.2. The Text Mining Module

The main component of the system is a Text Mining Module. It works as a *sniffer*, examining each message sent in the chat. This module is responsible for identifying themes or subjects in the messages. Themes are identified by comparing words present in the message against terms defined in the ontology. Generic terms like prepositions (called *stopwords*) will be disregarded. Each message is compared online against all concepts in the ontology. The concepts identified in the messages represent the topics being discussed and are forwarded to the Recommender Module.

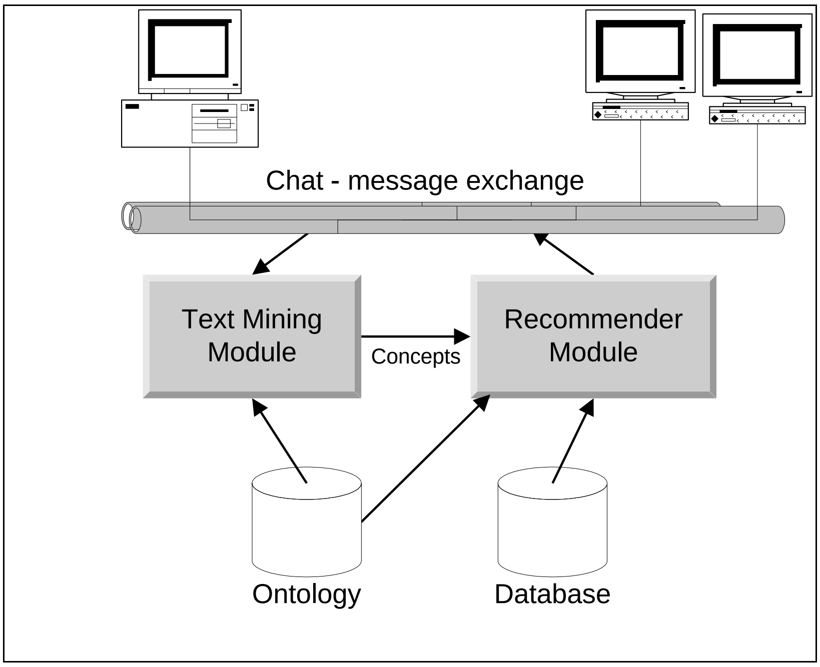
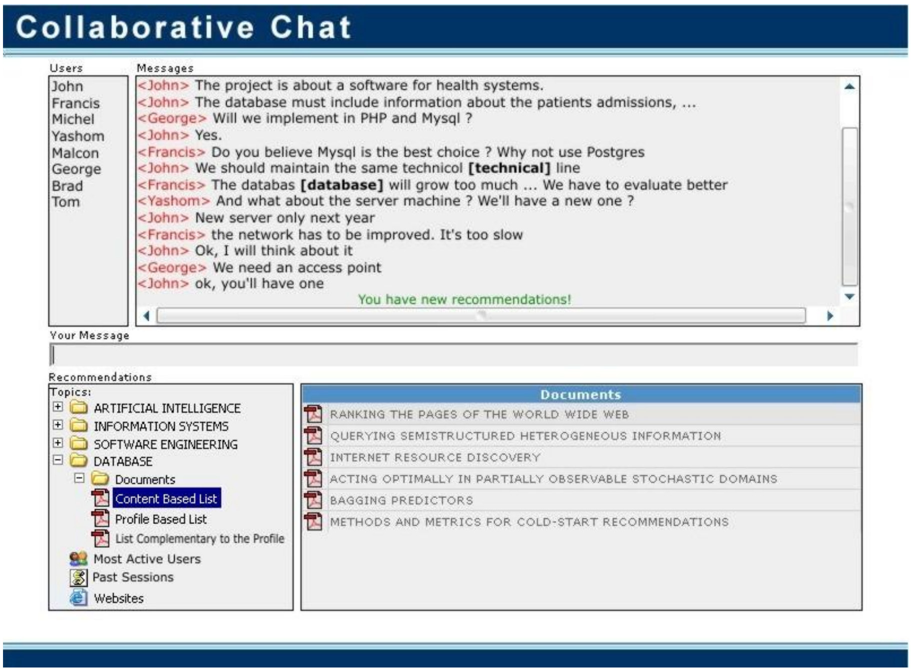


**Figure 1. Architecture of the Recommender System**

The text mining method employed in this work (a kind of classification task) was first presented in Loh, Wives, and Oliveira (2000). Instead of using Natural Language Processing (NLP) to analyze syntax and semantics, the method is based on probabilistic techniques: themes can be identified by cues. Using a *fuzzy* reasoning about the cues found in a text, it is possible to calculate the likelihood of a theme or subject being present in that text. The algorithm is based on Rocchio's and Bayes' algorithms (Lewis, 1998), (Ragas & Koster, 1998), (Rocchio, 1966), since it uses a prototype-like vector to represent texts and concepts. The method evaluates the relationship between a text and a concept of the ontology using a similarity function that calculates the distance between the two vectors. The vectors representing texts and concepts are composed by a list of terms with a weight associated to each term. In the case of texts, the weight represents the relative frequency of the term in the text (number of occurrences divided by the total number of terms in the text). And the weight in the concept vector represents the probability of the term being present in a text of that theme. The next section (the ontology) describes how concept weights are defined.

The text mining method compares the vector representing the text of a message against vectors representing concepts in the ontology. The method multiples the weights of common terms (those present in both vectors). The overall sum of these products is the degree of relation between the text and the concept, meaning the relative probability of the concept presence in the text or that the text holds the concept with a specific degree of importance. The decision concerning if a concept is present or not depends then on the threshold used to cut off undesirable degrees. This threshold is dependent on the domain ontology used in the system and is previously set by experts after some initial evaluations.

When two or more concepts are identified in the same message, the degree of relationship between the message and a concept is used to form a ranking. Only the top concept in the ranking is considered. New terms, used in the messages but not present in the ontology, are stored for future analysis. An orthographic corrector is used during the chat to avoid misspellings. Figure 2 presents examples of corrections on the words "technicol" and "databas".



<John> The project is about a software for health systems.
<John> The database must include information about the patients admissions, ...
<George> Will we implement in PHP and Mysql ?
<John> Yes.
<Francis> Do you believe Mysql is the best choice ? Why not use Postgres
<John> We should maintain the same technicol **[technical]** line
<Francis> The databas **[database]** will grow too much ... We have to evaluate better
<Yashom> And what about the server machine ? We'll have a new one ?
<John> New server only next year
<Francis> the network has to be improved. It's too slow
<John> Ok, I will think about it
<George> We need an access point
<John> ok, you'll have one
You have new recommendations!

RANKING THE PAGES OF THE WORLD WIDE WEB
QUERYING SEMISTRUCTURED HETEROGENEOUS INFORMATION
INTERNET RESOURCE DISCOVERY
ACTING OPTIMALLY IN PARTIALLY OBSERVABLE STOCHASTIC DOMAINS
BAGGING PREDICTORS
METHODS AND METRICS FOR COLD-START RECOMMENDATIONS

**Figure 2. A Snapshot of the Recommender System**

## 3.3. The Ontology

An ontology is a formal and explicit definition of concepts (classes or categories) and their attributes and relations (Noy & McGuinness, 2002). A domain ontology is a description of "things" that exist or can exist in a domain (Sowa, 2002) and contains the vocabulary related to the domain (Guarino, 1998).

In the proposed system, the ontology is implemented as a set of concepts in a hierarchical structure (a root node, parent-nodes and child-nodes). Each concept has associated to it a list of terms and their respective weights. Weights are used to state the relative importance or the probability of the term for identifying the concept in a text. The relation between concepts and terms is many-to-many, that is, a term may be present in more than one concept and a concept may be described by many terms.

The ontology is used to identify themes in textual messages, to automatically classify items of the digital library and to relate people to subjects for identifying interest areas (stored in the use's profile). The ontology is also used to retrieve items from the Digital Library or to search the base of past discussions. These two last operations are performed out of the chat, in specific modules.

A software tool is used to manage and configure the ontology, including functions to visualize the structure of concepts and the list of terms, to insert a new concept and its respective list of terms, to insert/remove terms and to modify the weights. New terms, found by the Text Mining module, may be added as a new concept or they may be added to term list of an existing concept or the new terms may be added to the stopword list. A group of experts should be responsible for creating and updating the ontology. Their first task is to define the concepts of the domain ontology and the relationships among concepts (the hierarchy). Software tools support experts in identifying terms for each concept. Terms and weights may be defined using the supervised learning strategy (for machine learning): experts select texts about the concepts present in the ontology and a software tool identifies the most important terms for each concept, establishing the weights using the TFIDF method (Salton & McGill, 1983). A normalization method is applied over the weights to avoid a great variation in the limits from one concept to other.

Currently, the system uses a domain ontology for Computer Science, but other ontologies can be used. For this purpose, the domain ontology has a root node called "ontology". Under this node, other ontologies may be aggregated. Concepts and the hierarchy were based on the ACM classification for Computer Science. Approximately 100 texts for each concept were used in the learning step. The texts where extracted from Citeseer digital library (www.researchindex.org) by experts. After, experts reviewed the ontology adding word variations with the same weight as the principal. This last task was important since texts were written in English and Portuguese. So, the terms used in the ontology come from these two languages.

## 3.4. The Digital Library

The Digital Library is a repository of information sources, especially created for the recommendation system, including electronic documents, links to Web page and bibliographic references. The inclusion (upload) of items in the Digital Library is responsibility of authorized people and can be made offline in a specific module.

The classification of the electronic documents is made automatically by software tools, using the same text mining method used in the Text Mining Module and the same ontology. A difference is that a document may be related to more than one concept. Thus, the relation between concepts and documents in the Digital Library is many-to-many. The relationship degree between the document and the concept is also stored and a threshold is used to determine which concepts can be regarded.

At the moment, Web links and bibliographic references are not classified automatically. We are studying how to make this possible. One alternative is to use the classification method used for electronic documents applied over texts extracted from the Web pages and over abstracts of the bibliographic references.

### 3.5. The Profile Database

The profile base contains identification of authorized users. Besides administrative data like name, institution, department, e-mail, etc., the profile base also stores the interest areas of each person, as well an associated degree, informing the degree of interest of user in the area. These areas are related to concepts in the ontology. The initial degree is set by system administrators, but there is a strategy of points for increasing this degree. The increase of points associated to a person in a specific area/concept depends on the activities of that person, using the system. Each time a person participates in a discussion, the subjects appearing in that session produce an increment in the user profile subject, indicating that his/her interest in that area is increasing. In addition, when somebody reads, opens or downloads items from the digital library, his/her profile is update. In this case, the system asks the user for a rating about the item (very useful, useful or not useful); if the user rates the item as “useful” or “very useful”, the corresponding concepts (where the items are classified) are incremented in the user profile; if the user rates the item as “not useful”, the corresponding concepts are decremented in the profile. In the future (after an ongoing study), points will be reduced if a person does not execute any operation in the system, during a specified time period. The scale of points, that is, how much each operation in the system should increment the profile of a person is under experimental evaluation and will be formally presented in a future paper.

The profile base is similar to a Knowledge Map or Yellow Pages (Stewart, 1998). In this sense, it is used by the Recommender Module to indicate people interested in an area (those with great interest on the area).

Associated to each person, the profile base also records the items accessed (uploaded, added, read or downloaded) by a person or recommended to him/her. This information is useful to avoid recommending known items. In the future, the profile base will be used for collaborative filtering, grouping people with similar characteristics, in order to recommend cross items.

### 3.6. The Base of Past Discussions

This base records everything that occurs in the chat, during a discussion session. Discussions are stored by sessions, identified by data. Associated to the session, the base must store who participated in the session, all the messages exchanged (with a label indicating who sent it), the concept identified in each message, the recommendations made during the session for each user and documents downloaded or read during the session.

The list of concepts identified during the session compose an interesting order, allowing users to analyze the path followed the participants during the discussion. For example, it is important to observe:

1. The depth of the discussion: if the discussion went deep, down in the concept hierarchy, or occurred superficially at a higher conceptual hierarchical level;
2. The extension or coverage: how many different subjects were discussed;
3. If the discussion deviate from the main sub-tree (if a node with no common parent was reached); and
4. What was the central point of the discussion (subject with more messages associated).

This base also allows users to review after the session the recommendations made during the discussion.

### 3.7. The Recommender Module

The goal of the Recommender Module is to offer information stored in the different bases to the chat participants. The module uses a content-based technique, where only items classified in the concepts identified in the discussion are recommended.

The action of this module starts when it receives a concept from the Text Mining Module. Then, it searches the different bases for items classified in the same concept. Each time the Text Mining Module identifies a concept in a message, it sends this concept to the Recommender Module that searches the database for items to recommend.

Since the discussion in the chat is synchronous, recommendations should not interrupt the users. So, indications are given in a separate frame and not inside the chat window (see Figure 2). Recommendations are particular of each user. Thus, each user receives a different list of suggestions in the screen. For each concept identified in the discussion by the text mining module, a different list of items are suggested (the concepts are classified inverse chronological order, as seen in the Figure 2, box in the left and bottom side). The system does not recommend:

1. The same item twice in the same section;
2. Items already associated to the user.

We are studying if is good or bad to recommend items already suggested in past discussions. There is a button to eliminate some items from the list, reducing the overload in the recommendation frame. Another button can lead the user to details of the item being recommended (information stored in the digital library, as title, authors, abstract and the electronic document itself).

For each concept, three kinds of recommendations are done. In all the three, the items that appear are those classified in the identified concept. A threshold previously set by experts is used to minimize the list of suggestions avoiding the information overload. The difference between the three kinds of recommendations is on the way the system ranks the items in the list:

1. The content-based list ranks the items according to the degree of relationship between the item and the concept, putting the items with greater degrees in the top;
2. The profile-based list ranks the items according to the profile of the user; the same list as in the content-based is considered but the degrees of each concept appearing in each document are multiplied by the corresponding concept degrees in the user's profiles; for example, considering a document X with concepts and degrees as { Database/0.9, Neural Nets/0.6, Software Engineering/0.3 }, a document Y with { Database/0.2, Neural Nets/0.9, Software Engineering/0.8}, for a profile as { Database/0.2, Neural Nets/0, Software Engineering/0.7 }, the ranking will be document Y in first place (with a total degree of 0.6) and document X in second place (with a total degree of 0.39);
3. The complementary list ranks the items according to the inverse profile of the user, that is, the concepts with less weight in the profile receive

more privilege; the intention is to complement the formation of the user, bringing to top documents that treat about themes where the user is not yet an expert; the method inverts the ranking in the profile disregarding concept with weigh zero; the concept with the highest degree will receive the minor degree and vice-versa; other concepts will maintain the same relation (proportional difference) between the limits; for example, assuming the same documents X and Y as above and the same profile, the ranking will be document X in first place (with a total degree of 0.69) and document Y in second place (with a total degree of 0.3).

Regarding the personalized recommendations, they can only be generated if there is a profile for the user. If there is yet no profile for a user, he/she will receive only the first kind of recommendations (the content-based list). As the user participates in chat discussions or accesses the digital library, a profile will be generate and the user will receive personalized recommendations.

## 4. Experiments

Currently, the proposed system uses an ontology for the Computer Science domain with 57 concepts and more than three thousands words (including terms in English and Portuguese).

The quality of the recommendations depends directly on the quality of the ontology and on the text mining method used on the chat messages and on the documents in the digital library. We carried out evaluations of each method separately. First, the orthographic corrector was evaluated on real discussions occurred in the chat; the error rate was 19%.

Second, we evaluated the text mining method; that also includes the evaluation of the ontology, since the text mining method depends on the ontology. The method was evaluated on messages posted in the chat during real sessions. A sample of 10 discussion sessions was selected for this evaluation. For each individual message, a concept was identified. Experts marked the concepts correctly identified and marked the messages were a concept should be identified but did not (actually some messages did not have concepts from the ontology). The evaluation resulted in 67.5% of precision (concepts correctly identified divided by total number of identified concepts) and 50% of recall (concepts correctly identified divided by total number of concepts that should be identified).

One raised assumption is that concepts can be identified with more precision when the messages are more specific and objective, restricted to one theme and when more than one message is analyzed. To evaluate this assumption, other method was evaluated in the same sample of sessions. This second method evaluated a group of 10 messages (the last 10 messages posted in the chat in order to determine the context). Participants of each chat judged the concepts identified in each group of messages. Table 1 shows the comparative results. The threshold used for both methods was 0.001. Comparing methods 1 and 2, it is possible to conclude that the latter reaches better results because take in account the context of the discussion (a group of messages). The assumption that a group of messages can better identify a subject is correct. A single message can lead to ambiguity or has too much uncertainty to allow the identification of a subject. These results confirm the importance of the context and the self-contained characteristic of the textual message.

Other interesting observation is that the second method (considering a group of chat messages) achieved a precision of 95% in one session analyzed individually. This allows concluding that the text mining method can reach good results on real chat sessions when the context is analyzed.

**Table 1. Results of the Text Mining Evaluation**

| Method | Average Precision | Average Recall |
|---|---|---|
| **1** | 67,5% | 50,0% |
| **2** | 79,0% | 86,1% |

Experiments were carried out with undergraduate students of a Computer Science course utilizing the system in some classes of different disciplines. The class was conducted by the teacher using the chat. After each session, students were asked to talk about the system and the process. The majority of the students reported benefits when using the system, since they did not have to search the digital library for documents and the system returned new and interesting documents. However, none student was comfortable to read an entire document during the session. Some reported that, when viewing the content of recommended documents, they lost part of the discussion. By other side, they reported that this is not a disadvantage of the system because in some way the process is like searching web with search engines (they receive a list of items and have to click in each one to verify the content). From that, we can conclude that the system is better suited for retrieving documents to the user, hoping that the user will see the documents after the chat session. The system has an option to retrieve past discussions by date, name of the participant or by words present in the messages.

## 5. Concluding Remarks

Feigenbaum *apud* Davies (1989) compares the current libraries against the future ones. The current libraries are repositories of passive objects, while the future libraries will be composed by active objects, helping people to discover new knowledge, providing hidden associations, analogies and new concepts, without people having to state what they need. In this sense, the proposed system intends to complement the knowledge people have recommending documents about person’s interest areas, discovered by analyzing chat discussions. Based on the experiments, we can say that the system helps the collective and the individual construction of knowledge. When the system generates a recommendation, users can share opinions about the recommended items during the discussion. On the other hand, the list of recommended items is different for each user, depending on his/her interest areas and on the degree of interest or knowledge. The complementary list (using the inverse of the user’s profile) is useful to present documents in areas where the user has less activity or knowledge.

The philosophical approach used by the system follows the social-interactionist theory of Vygotsky (1984), defending the knowledge construction as a collective task. Collaboration is present in the system through the chat and the digital library. Messages posted in the chat contribute to knowledge exchange, especially when people are physically distant. The construction of the digital library is a collective task and allows people to share information sources, since items can be included in the digital library by different people.

Furthermore, the system can help each person to achieve his/her potential knowledge by acting in the Zone of Proximal Development. The system acts as a mediator suggesting complementary sources and thus enabling knowledge extension. The possibility of retrieving the discussion (messages and recommendations) after the chat session is important to users review topics that could not be understood during the session, especially when the discussion deviates from the main topic. This is also a way to transform tacit knowledge in explicit knowledge since ideas concretized as textual messages are recorded and can be retrieved later.

The quality of the recommendations depends on the quality of the text mining method that identifies concepts in the chat messages and on the quality of the digital library. The performance of the method in online messages (79% of precision) in not the ideal but is enough to generate recommendations with a certain precision. Regarding the digital library, the effort is on populating the base with quality documents. The hard work is to find good documents, because the system automatically indexes the document and extracts some attributes as title, authors, e-mails and abstracts. The responsibility of finding good documents relies on users participating in the virtual community that uses the system. We believe that the collaborative work of these members can generate a good base. For example, the current digital library for Computer Science has more than one thousand items (gathered in few years). An ongoing work is investigating ways to collect documents in the Web and filtering them by quality measures.

One of the main characteristics of chat messages is the conciseness and informality to reduce the response time in a conversation. As a result, there is a lot of implied information embedded in the messages (the context). For example, when discussing about “computer networks”, people tend to use only the term “networks”. Analysis of context helps people to understand the missing information. This paper showed that analyzing a group of messages allows the system to handle the context of the discussion and then to improve the identification of subjects in a chat discussion. A precision of 79% in identifying concepts, as obtained in the experiments, is a good result. However, method 2, described in this work, still needs to be improved. A future work will investigate the best window of messages (number of messages in a group to be analyzed jointly). One question is if it is better to consider a group composed of the N last messages or a message group defined by a time interval (for example, the messages sent in the last 2 minutes).

The drawbacks of the system are related to dynamism of the chat and recommendations, making difficult to keep attention on the messages (discussion) and on the recommendations generated by the system at the same time. This problem is minimized by an alternative option in the system, where users can review past discussions (sent messages and recommendations).

Another drawback is the process for constructing the domain ontology. This is a work that consumes much time and effort. The main task is to define the relevant concepts of the domain and the relations among them (the hierarchy of concepts). After that, the task is to find terms and weights for each concept. This task is made by a supervised learning process, where experts select textual documents about each concept and a software tool identifies the relevant terms and establishes the weights using the TFIDF method (Salton & McGill, 1983).

Finally, the application of the described procedures may benefit some applications systems for identifying subjects in chat messages. That is special important for systems that:

1. Identify expertise in chat discussions, in order to recommend authorities in certain subjects;
2. Make personalized offers, for example recommending items in a responsive way to the interest of the participants;
3. Advertise information within the context of the discussion.

## Acknowledgements

This research is partially supported by CNPq, an entity of the Brazilian government for scientific and technological development (Project DIGITEX - Editoração, Indexação e Busca em Bibliotecas, grant number 550845/2005-4), and FAPERGS, Foundation for Supporting Research in Rio Grande do Sul state (Project Rec-Semântica - Plataforma de Recomendação e Consulta na Web Semântica Evolutiva, grant number 0408933).

## References

**1** Anjewierden, A., Kolloffel, B., & Hulshof, C. (2007). Towards educational data mining: using data mining methods for automated chat analysis to understand and support inquiry learning processes. *Proceedings of International Workshop on Applying Data Mining in e-Learning*, 27-36.

**2** Davies, R. (1989). The creation of new knowledge by information retrieval and classification. *Journal of Documentation*, *45(4)*, 273-301.

**3** Guarino, N. (1998). Formal Ontology and Information Systems. *Proceedings of International Conference on Formal Ontologies in Information Systems - FOIS'98*, 3-15.

**4** Holmer. T. (2008). Discourse structure analysis of chat communication. *Language@Internet*, *(5)*. Retrieved from http://www.languageatinternet.de/articles/2008/1633

**5** Khan, F. M., Fisher, T. A., Shuler, L., Wu T., & Pottenger, W. M. (2002). *Mining chat-room conversations for social and semantic interactions*. Technical Report, LU-CSE-02-011, Lehigh University. Retrieved from http://www3.lehigh.edu/images/userImages/cdh3/Page_3456/LU-CSE-02-011.pdf

**6** Lawrence, R. D., Almasi, G. S., Kotlyar, V., Viveros, M. S., & Duri, S. S. (2001). Personalization of Supermarket Product Recommendations. *Data Min. Knowl. Discov. 5(1-2)*, 11-32.

**7** Lewis, D. D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In: *European Conference on Machine Learning. Lecture Notes in Computer Science, Vol. 1398, ECML '98: Proceedings of the 10th European Conference on Machine Learning* (pp. 4-15). London, UK: Springer-Verlag doi: 10.1007/BFb0026666

**8** Loh, S., Wives, L. K., & Oliveira, J. P. 2000. Concept-based knowledge discovery in texts extracted from the Web. *SIGKDD Explor. Newsl.* 2(1), 29-39.

**9** Lonchamp, J. (2005). A structured chat framework for distributed educational settings. *Proceedings of The 2005 Conference on Computer Support For Collaborative Learning: Learning 2005: the Next 10 Years!*, 403-407.

**10** Mock, K. (2001). The use of internet tools to supplement communication in the classroom. *J. Comput. Small Coll.*, *17(2)*, 14-21.

**11** Mu, X., Marchionini, G., and Pattee, A. (2003). The interactive shared educational environment: user interface, system architecture and field study. *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, 291-300.

**12** Neuage, T. (2005). *Conversational analysis of chatroom talk*. South Australia University, 452p. (PhD Thesis).

**13** Noy, N. F. & McGuiness, D. L. (2002). *Ontology Development 101: a guide to creating your first ontology*. Retrieved from http://protege.stanford.edu/publications/

**14** Ragas, H. & Koster, C. H. A. (1998). Four text classification algorithms compared on a Dutch corpus. *Proceedings of International ACM-SIGIR Conference on Research and Development*. Melbourne, 369-370.

**15** Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: an open architecture for collaborative filtering of netnews. *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 175-186.

**16** Resnick, P. & Varian, H. R. (1997). Recommender systems. *Commun. ACM 40(*3), 56-58.

**17** Rocchio, J. J. (1996). *Document retrieval systems - optimization and evaluation*. Phd Thesis, Harvard University, Cambridge.

**18** Salton, G. & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

**19** Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-Commerce Recommendation Applications. *Data Min. Knowl. Discov. 5(1-2)*, 115-153. doi: http://dx.doi.org/10.1023/A:1009804230409

**20** Sowa, J. F. (2002) *Building, sharing, and merging ontologies*, AAAI Press / MIT press, 3-41.

**21** Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. (2000). Web usage mining: discovery and applications of usage patterns from Web data. *SIGKDD Explor. Newsl.*, *1(2)*, 12-23.

**22** Stewart, T. A. (1998). *Intellectual Capital: The New Wealth of Organizations*. New York: Bantam Books.

**23** Terveen, L. & Hill, W. (2001). Beyond recommender systems: helping people help each other. In J. Carroll (Ed.), *Human computer interaction in the new millennium*. (pp. 487-509). Addison-Wesley.

**24** Vygotsky, L. S. (1984). *The social formation of mind* (in portuguese). São Paulo, Brasil: Martins Fontes.

**25** Walker, A., Recker, M. M., Lawless, K., & Wiley, D. (2004). Collaborative Information Filtering: A Review and an Educational Application. *Int. J. Artif. Intell. Ed. 14(1)*, 3-28.

**26** Wu, T., Khan, F. M., Fisher, T. A., Shuler, L. A., Pottenger, W. M. (2002). *Error-driven boolean-logic-rule-based learning for mining chat-room conversations*. Lehigh CSC 2002 Technical Reports.