An Integrative Framework for Knowledge Extraction in Collaborative Virtual Environments

Robert P. Biuk-Aghai
University of Macau
Faculty of Science and Technology
P.O. Box 3001
Macau S.A.R.
China
+853-3974365
fst.robert@umac.mo

Simeon J. Simoff
University of Technology, Sydney
Faculty of Information Technology
P.O. Box 123
Broadway, NSW 2007
Australia
+61-2-95141103
simeon@it.uts.edu.au

ABSTRACT
Collaborative virtual environments are becoming an intrinsic part of professional practices. In addition to providing collaboration support, they have the potential to collect vast amounts of data about collaborative activities. The aim of this research is to utilize this data effectively, extract meaningful insights out of it and feeding discovered knowledge back into the environment. The paper presents a framework for integrating knowledge discovery techniques with collaborative virtual environments, starting from early conceptual development. Discovered patterns are deposited in an organizational memory which makes these available within the virtual environment. Two examples of the application of the framework are included.

Categories and Subject Descriptors
H.1.m [**Models and Principles**]: Miscellaneous; H.2.8 [**Database Management**]: Database Applications – Data mining; H.4.1 [**Information Systems Applications**]: Office Automation – Groupware.

General Terms
Design, Experimentation, Human Factors, Theory.

Keywords
Knowledge extraction, collaborative virtual environment, virtual collaboration.

1. INTRODUCTION
Collaborative virtual environments have become increasingly popular in recent years. There are numerous approaches and techniques for arranging such environments for collaborative projects. The most common approach is to extend the desktop environment to include tools for meeting and sharing files. This approach takes the individual work environment and adds tools for communicating with others. Unfortunately, the toolbox approach handles only limited collaborative activity when usually people proceed effortlessly between different working styles, in terms of time, place and representations. This introduces a gap that interferes with collaborative activity. More formally, such gaps are defined as the physical or perceptual boundaries within the computer environment that either distract participants from the work they are doing, or that block them from crossing spatial, temporal or functional boundaries, inherent in collaborative work [15]. Trying to bridge these gaps, Gutwin and Greenberg [13] recognized the importance of the ontology of "place" and virtual environments that follow such ontology. These environments range from simple desktop-style places to sophisticated virtual reality worlds (for an excellent taxonomy of the latter see [7]). Despite their variety and difference in functionality these environments have several key concepts in common:

the concept of "*inserting people*" in the networked environment, in other words, representing them as some entities. These representations span from the so-called "characters" in text-based MOO/MUD and Web-based WOO environments [19] to the "avatars" in the 3D virtual worlds;

the concept of "*structuring the space*" in the networked environment, in other words, providing some way of structuring the place, separating and handling different information within the units of this structure, and some reference system for orientation and navigation. These structures span from the "room" approach in MOO/MUDs, to the "squares of land" and "worlds" in ActiveWorlds universes;

the concept of "*a feasible set of actions*" that can be performed in the networked environment. This set defines to

what extent the environment under consideration can be used for conducting collaborative projects in a particular domain.

1.1 People

The establishment of the identity of the people in the virtual place occurs through the representation of individuals as objects or avatars that possess various properties, and through the behaviour of that representation. Object representations of a person include characteristics such as a textual description, messages about their movements in the place, and links to web pages to help establish their identity and personality. An important aspect of people's representation is the variety of "rights" that can be assigned to them. Different environments use different terms for this – privileges, roles, permissions. Thus, the representations are potential sources of preliminary information about a person's individuality. However, in collaborative projects it is important to be able to make judgments about the collaboration of the people in a team or in different teams and to reuse such knowledge when forming teams in other collaborative projects. The preliminary information is not always sufficient for establishing successful work. Data mining techniques can be applied for extracting information about the functioning of groups of individuals and discovering patterns of collaboration based on project communication between them. This knowledge can be reused for configuring groups in new projects.

1.2 Space Structuring

The ways of structuring of the environment's space depend on a number of factors, including the ontology (what kind of place the environment is), purpose of the environment, the embedded functionality, the preferable communication and collaboration mode [18], underlying technologies and their integration [22]. For example, The Virtual Campus (Faculty of Architecture, University of Sydney) shown in Figure 1 is organized according to the ontology of a university campus. The space is structured in terms of "rooms", "levels" and "buildings", which follows the ontology of building design. The reference system and the topology of the space are based on the purpose of the "buildings" and the "rooms" in them. This ontology defines the partition of space [19]. The space structure of a virtual environment usually evolves according to the needs of a project. One way to approach this problem is to create "design prototypes" according to the ontology of the environment. In our example, a prototype of a faculty building can be a "building" with four "levels": "Classrooms", with rooms for each subject; "Offices", with rooms for staff members; "Library", with rooms that keep information from past subjects; and "Common level" which can accommodate general purpose meeting rooms, practice rooms and other functional spaces. This approach does not capture the results of the actual use of the virtual environment – which parts of it were used more intensively, what are the "neighbouring" relations (e.g., co-visited rooms), and other relations. Data mining techniques can be applied for discovering such relations. Discovered knowledge can be reflected in variations of the space structuring of the "design prototypes", resulting in building a library of such prototypes and reusing them according to the requirements of the new project space.

1.3 Feasible Actions

The ontology of the virtual environment can provide substantial *a priori* knowledge not only about the navigation, but also about the set of feasible actions in such an environment. Usually the initial set of actions is derived from the design requirements. This initial set can provide substantial *a priori* knowledge for the analysis and discovery of patterns of collaboration. The real set of actions used in different projects may vary substantially. The overlapping set of actions forms the common kernel set and the rest is the individual component. In the long term, this provides a potential for designing pro-active prototypes supporting different types of projects. Data mining techniques can be applied for composing such action sets. Discovered action sets and space structures will form pro-active design prototypes, resulting in a library of such prototypes and their reuse according to the requirements of a new project.

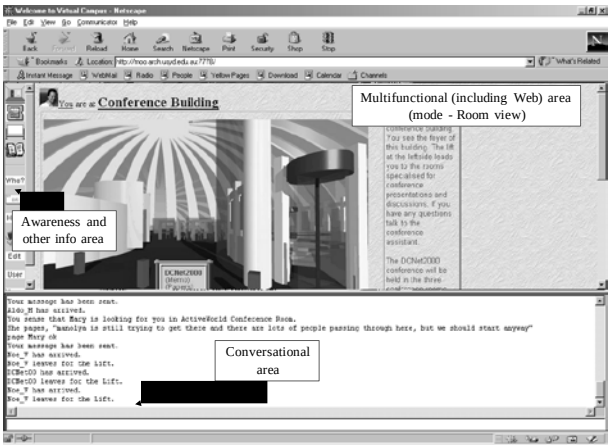


Figure 1. A "Conference building" in the Virtual Campus (http://www.arch.usyd.edu.au:7778/)

The design of collaborative virtual environments (CVE) remains a craftwork. In this paper we propose to combine CVE and data mining technology to develop more coherent and consistent environments.

Collaborative virtual environments have the potential to provide professional working environments that can support collaborative projects in different disciplines independent of geography. Consequently, they can provide researchers with enormous amounts of data about various aspects of computer-mediated collaboration. Unfortunately, the design of earlier environments did not pay much attention to the issues of data collection [12]. Thus, the application of data mining methods had to struggle with translating data collected for other purposes, for example, a server log used usually for correct recovery after a failure, into data useful for the goals of data mining. Consequently, the earlier application of data mining methods in collaborative virtual environments has been focused mainly on the analysis of communication transcripts – whether recorded in synchronous collaborative sessions or over a bulletin board in asynchronous mode.

In this paper we present a framework for integrating data mining in the design of collaborative virtual environments, in a way that facilitates not only the data collection and analysis, but also the application of discovered knowledge. This framework differs from the approach presented by Chen [8], who uses graphical capabilities of the virtual environment to support the visual exploration of external data within the environment itself.

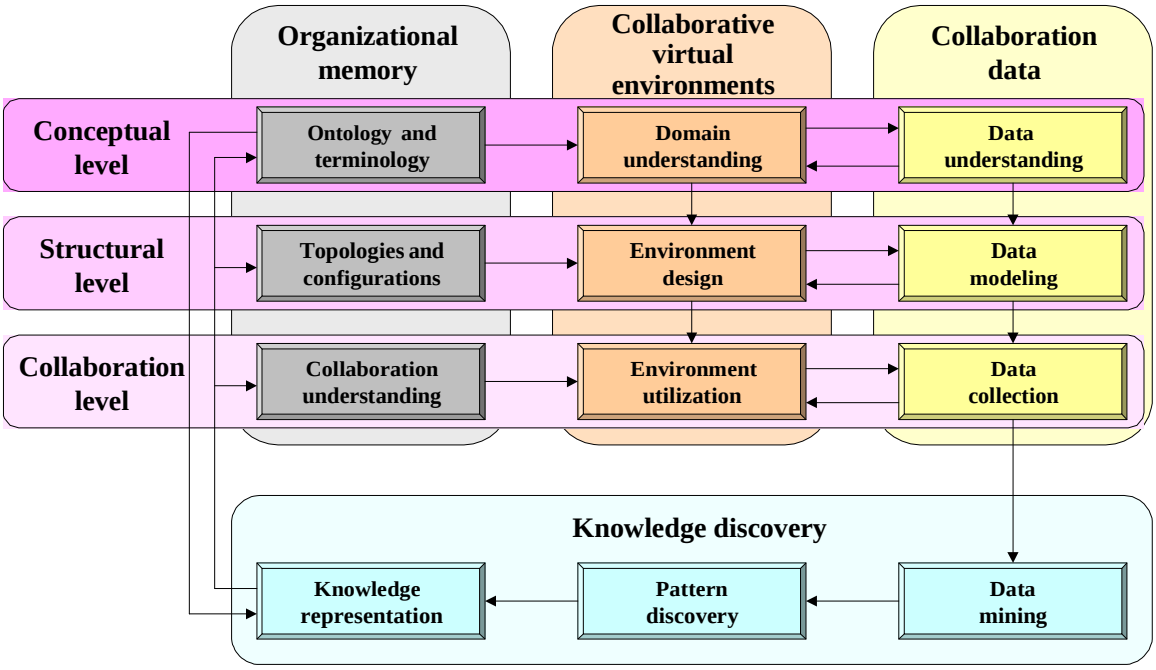


Figure 2. Framework for integrating data mining in the design and application of collaborative virtual environments and knowledge extraction from them

After presentation of the framework, we illustrate its application to two cases:

monitoring and visualization of team collaboration based on the data collected over project discussion boards and incorporating that knowledge

monitoring and extracting knowledge from collaborative activities and incorporating that knowledge

2. THE FRAMEWORK

The presented framework embeds knowledge discovery in collaborative virtual environments. Its two primary goals are:

1. To influence the design of collaborative virtual environments so as to provide the data necessary for mining and analysis of collaboration.
2. To feed extracted knowledge back into the use of collaborative virtual environments.

As a result, data design and design of the collaborative virtual environment are seen as complementary and parallel activities, affording the opportunity to control data collection to a greater extent. Knowledge obtained from collaboration data is a further contributor to CVE design. A number of related research efforts are underway in the direction of controlled data collection, carried out mainly in the field of e-commerce and Web data mining [23].

The framework is shown in Figure 2. It includes four major groups of inter-woven components:

1. *Collaborative virtual environments*.
2. *Collaboration data*.
3. *Knowledge discovery*.
4. *Organizational memory*.

Moreover, the three components appearing in the upper part of the figure consist of three parts, at different levels of abstraction:

1. *Conceptual level.*
2. *Structural level*.
3. *Collaboration level*.

Below, we discuss the components of the framework in more detail.

2.1 Collaborative Virtual Environments

Collaborative virtual environments (CVEs) are the support systems within which collaboration is carried out. As mentioned earlier, CVEs are becoming increasingly part of professional practice. Such environments aim to support certain work practices, hence are domain-specific. For each domain, an understanding of the domain-dependent requirements for the CVE has to be obtained. On a conceptual level this activity identifies the concepts to be supported by the environment: the structuring metaphor employed, navigation facilities, representation of people and their abilities, artefacts and tools provided in the environment, etc.

On the structural level, this initial step is followed by the actual design of the CVE when the relationship between the identified concepts is established and their detail is elaborated. Once designed (and implemented), the CVE is utilized by its users on the collaboration level.

2.2 Collaboration Data

The activities related to collaborative virtual environments are paralleled by those related to collaboration data. Within the

presented framework, collaboration data is that portion of data which facilitates knowledge discovery within the domain of collaboration, regardless of whether it is of direct use within the CVE. Traditionally, virtual collaboration systems did not provide any particular support for data collection aimed at knowledge discovery. Data was seen as an internal aspect of the system and only internally required data was maintained. The presented framework emphasizes the need for additional data that can enable knowledge discovery in the collaboration domain, and therefore within this framework collaboration data is treated separately from the CVE.

On a conceptual level, domain understanding within the CVE sphere and data understanding within the sphere of collaboration data are mutually complementary: once domain understanding identifies a concept to be supported, data understanding identifies the necessary data elements.

On the structural level, during environment design, data modeling identifies details of and relationships among the collaboration concepts and data.

Finally, on the collaboration level, the CVE is utilized and produces collaboration data which is collected for subsequent data mining.

2.3 Knowledge Discovery

The knowledge discovery in this framework differs slightly from the classical schema [10] – the selection and data pre-processing stages are implicitly embedded in the data design. Therefore, collected data is expected to be ready for data mining.

Knowledge discovery starts with applying traditional data mining algorithms to collaboration data leading to the discovery of patterns in the data. As a further difference to the classical knowledge discovery schema, a step of knowledge representation is included at the end. Its purpose is to map discovered knowledge back into the CVE's representation.

Knowledge discovery aims to produce a better understanding of computer-mediated collaboration, and to enable the usage of discovered knowledge to improve structural features. For example, through analysis of the structuring of virtual environments, templates of structures of these environments can be collected, implying certain navigation behaviour. Collecting data about actual navigation within the environment can provide a source for discovering traversal patterns, which can provide indicators for improving the topology (structuring) of the environment. Other possibilities for improvement of the environment exists according to particular collaboration needs. This is something difficult to know ahead of time. In both cases, some necessary indicators for improvement of the structure are required, which knowledge discovery provides.

2.4 Organizational Memory

Over the past decade, the CSCW community and related areas have taken a keen interest in organizational memory (OM) [1],[3],[9]. This suggests that there is value in retaining and later drawing on historical records of virtual collaboration. Such records could be referenced when setting out on new virtual collaboration, to "see how others have done it", and perhaps to reuse and re-enact those collaboration instances. Unlike conventional work settings where details of collaboration have to be collected manually through effort-intensive and sometimes intrusive methods, CVEs are an ideal source of data on collaboration, particularly when work is predominantly or entirely carried out virtually, as such environments can automatically record a great amount of detail on the collaboration.

While much work in organizational memory concerns itself with the content of collaboration, or the *declarative memory*, little work has been done on harnessing the *procedural memory*, or knowledge about how work has been carried out. The importance of utilizing this aspect of organizational memory in groupware systems has been pointed out relatively early [9], and again more recently within the context of virtual team effectiveness [11]. The presented framework makes the procedural portion of organizational memory an integral part of the collaboration support environment by maintaining knowledge extracted from collaboration environments and making it available within the environment.

On the collaboration level, this knowledge relates to an understanding of the collaboration. For example, it can identify what main types of activities were conducted within a virtual environment, how the activities were carried out over time, what differences exist in the activity of different people within the environment, etc. This knowledge can be utilized within the environment itself, leading for instance to an adaptation of the environment itself and/or its interface in order to facilitate the execution of predominant activities. It can also serve as a management and control instrument, which is of particular value when collaboration is completely virtual and traditional management methods are severely limited.

On the structural level, representations of the environment's topology are maintained. Where structural patterns are discovered in a set of environments, this too is deposited in the organizational memory in the form of different topologies and configurations available for reuse. Such information may feed back into environments in use, for instance to rearrange the environment's topology if its current arrangement is discovered to encumber work.

Use of collaborative virtual environments may, over time, also lead to the emergence of new concepts, or an application of existing concepts in ways that were not previously anticipated. These are deposited on the conceptual level as modifications to the underlying ontology, and feed into the ongoing development of a CVE. An example of this is where an environment lacks a certain feature, but where users discover workarounds that, though cumbersome, allow the feature to be supported. Discovery of such cases can be of use in the development of the next version of the CVE to explicitly support the feature.

3. MONITORING AND VISUALIZATION OF TEAM COLLABORATION

A typical scenario from a participatory design session in the Virtual Design Studio is shown in Figure 3. Such an environment can provide rich multimedia data, including data about the evolving geometry of a design and transcripts of the corresponding discussions of the ideas on each step; data about the allocation and behaviour of participants; web content of project documentation; even audio and video records. The transcripts of the "conversations" (the chat logs) during the collaborative sessions are a rich data source. A methodology for pre-processing and analysing such transcripts and for deriving

measures for estimating participation in synchronous collaborative sessions was presented in detail in [20],[21], and is beyond the scope of this paper. This approach has been extended to incorporate the on-line analysis of project communications via a discussion board system. Personal contributions to a collaborative session can be evaluated using text analysis of transcripts [21] and multimedia analysis of related web pages.

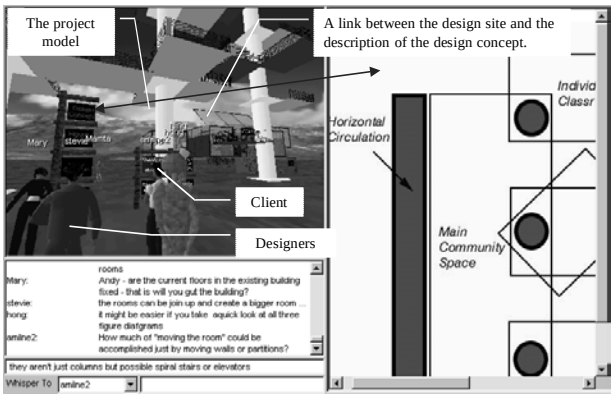


Figure 3. An example of a collaborative project in the virtual design studio

3.1 Environment Design

The design of a bulletin board is part of the environment design and depends on the project scenario and its collaboration mode. Figure 4 shows the configuration of a board for a case where 6 teams are developing their projects independently.

3.2 Data Understanding, Modeling and Collection

Figure 5 presents a fragment from a team bulletin board. The messages on the board are grouped in threads. A threefold split of the thread structure of e-mail messages in discussion archives was proposed in [4], [5], in order to explore the interactive threads. It included (i) reference-depth: how many references were found in a sequence before this message; (ii) reference-width: how many references were found which referred to this message; and (iii) reference-height: how many references were found in a sequence after this message. The threefold split was extended in [24] to include the time variable explicitly. This model, expressed graphically as a tree, allows the comparison of the structure of discussion threads both in a static mode (for example, their length and width at corresponding levels) and in a dynamic mode (for example, detecting moments of time when one thread dominates another in multi-threaded discussions).

3.3 Data Mining and Visualization

Based on this model, on-line visualization techniques have been developed, which are modified versions of the nested set visualization of tree structures [16]. Figure 6 shows an example of such visualization applied to threads "A" and "B" from Figure 5.

Each first message in a level is represented by a corresponding rectangle, labeled in this example to illustrate the message correspondence. Thus, there are four nested rectangles in Figure 6a. When messages are at the same level, the thickness of the line is estimated based on the content-analysis of the message, including the text, graphics and images. As a reasonable approximation, each of the relevant messages on the same level can be represented as additional 0.5 pt to the base line thickness. In Figure 6b the base line thickness is 1 pt, thus rectangle "M2B" has thickness 2.5 pt.

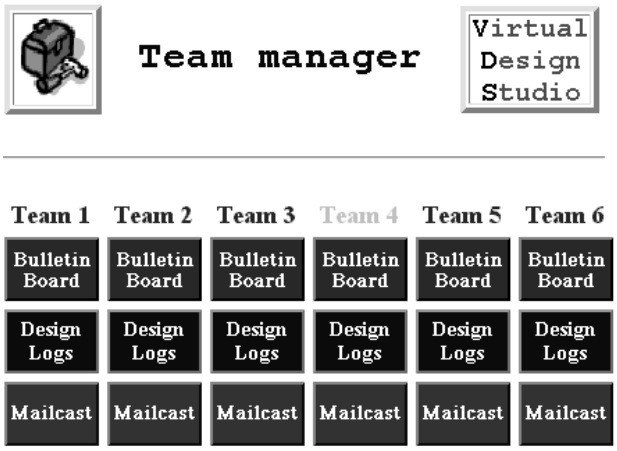


Figure 4. Team bulletin board configuration

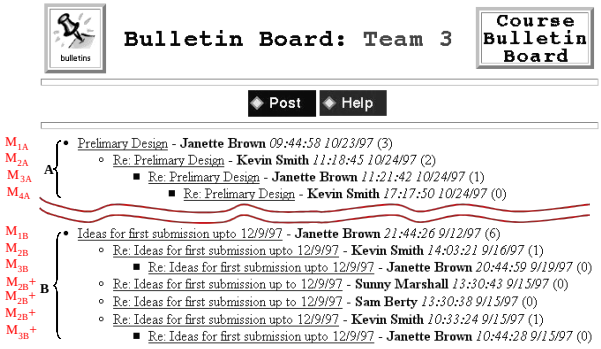
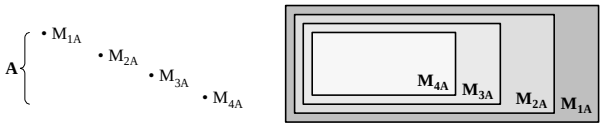
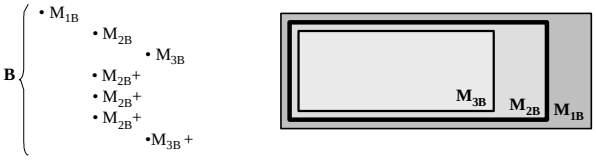


Figure 5. Fragments from an asynchronous communication on the bulletin board in a virtual world

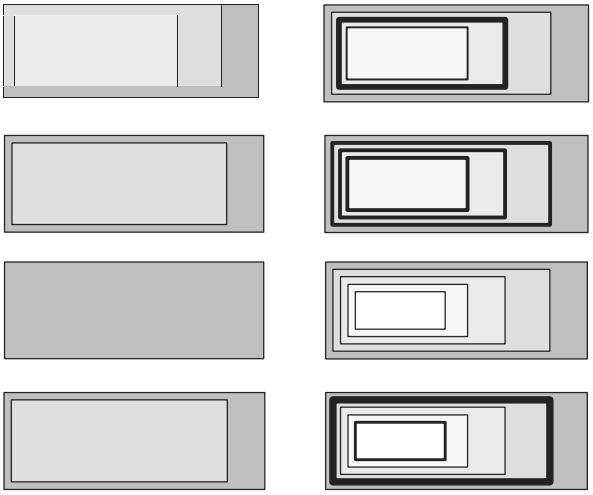


a) Nested rectangles for single message per level.



b) Nested rectangles when there are multiple messages on some levels.

Figure 6. Visualization of discussion threads



a) Collaboration without creating a shared understanding of the problem

b) Intensive collaboration for creating a shared understanding of the problem

Figure 7. Patterns of collaboration

Figure 7 illustrates the application of the technique for monitoring collaborative teams. Collaboration can be considered at different levels of task sharing. Two extreme approaches to sharing tasks during collaboration are identified in [17]: single task collaboration and multiple task collaboration. During single task collaboration, the product is a result of a continued attempt to construct and maintain a shared conception of the task. In other words, each of the participants has his/her own view over the whole problem and the shared conception is developed during intensive discussions. The basic assumption is that collaboration style influences the communication pattern. An example, of the visual pattern of such type of collaboration is presented in Figure 7b. It is characterized with relatively large numbers of nested rectangles, usually indicating also several messages in response to a particular message. During multiple task collaboration, the problem is divided among the participants so that each person is responsible for a particular portion of the product. Thus, multiple task collaboration does not necessarily require the creation of a single shared conception, thus messages are usually related to project management. Isolated messages and short threads dominate this collaboration style, as illustrated in Figure 7a.

3.4 Organizational Memory

The organizational memory in this case can consist of a selection of patterns that correspond to a specific collaboration. The patterns can assist the restructuring of the bulletin board system for similar types of projects. For example, if particular aspects of a new project are expected to generate intensive and long threads with important information, they can be allocated separate boards within the team discussion area. Content analysis of the messages can reveal some specific terminology. Apart from identifying participation and collaboration patterns, it has been difficult (if not impossible) to extract and analyse data that can provide insights about structuring the environment and the feasible set of actions without preliminary design of the action data to be recorded.

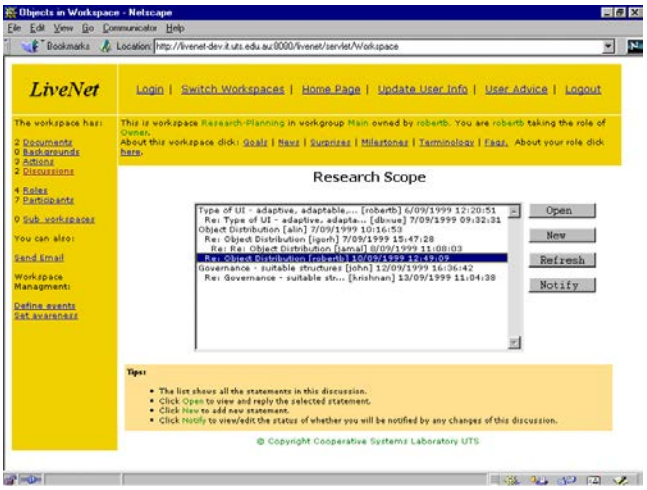


Figure 8. Typical LiveNet screen (web interface)

Figure 9. Simplified ontology of LiveNet

4. MONITORING AND EXTRACTING KNOWLEDGE FROM COLLABORATIVE ACTIVITIES

A second example of the application of the framework is related to the LiveNet virtual collaboration system. The framework was particularly applied in the areas of collaborative virtual environments, collaboration data, and knowledge discovery. We first introduce the LiveNet system, then show how the framework was applied.

4.1 LiveNet as a Collaborative Virtual Environment

The LiveNet system is a virtual collaboration system prototype developed at the University of Technology, Sydney [14]. It supports mainly asynchronous collaboration of distributed groups of people, i.e. different-time, different-place interactions, although its design does not limit it from other modes of collaboration. A central server is accessed across the network through one of several client interfaces, most commonly through a Web interface (as shown in Figure 8). LiveNet provides virtual workspaces which bring together people, artefacts (e.g. documents), communication channels, awareness facilities, and a collection of tools, all tied together through a configurable governance structure. A simplified ontology of the LiveNet CVE is shown in Figure 9. In terms of the ontology, workspaces contain roles, occupied by participants (i.e. actual people), who perform actions.

Some actions may operate on document artefacts, others may be interactions with other workspace participants through discussions. Most workspace elements such as documents, discussions and participants, may be shared between workspaces. Thus workspaces are not just stand-alone entities but nodes in a network of inter-connected collaboration spaces. Neither are structures of workspaces in LiveNet static—once created, a workspace can be dynamically adapted to evolve together with the collaboration carried out in it, while likewise entire "ecologies" of inter-connected workspaces can co-evolve.

4.2 Integrating Collaboration Data

Early versions of LiveNet had not been developed with the support of knowledge extraction in mind. Consequently, only a limited amount of collaboration data was available, namely only data which was necessary for the internal operation of the system. While this allowed some structural-level features to be extracted, there was no data to support the extraction of collaboration-level or conceptual-level features.

Subsequently, the provision of suitable collaboration data was "retrofitted" onto an earlier version of LiveNet. Building on the existing domain understanding, conceptual data requirements were developed, followed by data modeling. These were integrated into LiveNet by appropriately adapting its design (and implementation) – corresponding to the flow from data modeling to environment design in our framework. Finally, collection of the new collaboration data and knowledge discovery followed.

The first iteration of this cycle lead to some knowledge extraction, both on the collaboration and the structural levels. However, analysis of the collaboration also revealed that certain data elements which were not captured at the time would be needed to provide a more complete picture of the collaboration. This had not been accounted for in the first cycle of integration of collaboration data. Consequently, a second cycle was initiated in which data understanding and data modeling were refined, and environment design was brought up-to-date with the new data model. The subsequent data collection and data mining lead to a more comprehensive analysis of collaboration and a richer knowledge discovery. Following this second cycle, new data requirements are already emerging which, once implemented, will lead to a yet richer body of collaboration data. This confirms to us the validity of our framework in feeding discovered knowledge back into the ongoing development of the collaborative environment. It also highlights the fact that this is likely not achieved in a single effort, but is an iterative process, with insight from each iteration triggering a new iteration.

4.3 Knowledge Discovery in LiveNet

Collaboration data in LiveNet consists of two parts: a database contains the internal data of the CVE, maintaining the current state of all workspace elements (documents, roles, participants, etc.). The second part is a set of log files that are external to the system itself and which record all user actions carried out in the system over time. Although the vast majority of users interact with LiveNet through a web interface, the log records captured by the LiveNet server are on a semantically much higher level than those in the corresponding web access log. While a web log includes IP addresses, document names, timestamps and http request types, the LiveNet log records information in terms of the LiveNet CVE's conceptual model. Thus every record includes the name of the workspace and its owner, the name of the participant carrying out the action, his/her role name, the LiveNet server command requested, etc. This allows analysis to exploit metadata available in the application and to capture higher-level actions than a mere web log does (this corresponds to the approach of [2]).

The analysis we carried out focused primarily on the log of collaboration actions, and to a lesser extent on the workspace database. It involved pre-processing of the log, visualization of workspace data, and actual data mining. The pre-processing step normalizes session numbers, aggregates lower-level events into higher-level actions, and calculates session summaries. In this context, a session is the sequence of actions carried out by a user from login to logout time. Data pre-processing is considered part of collaboration data collection and is usually automatically performed.

The data used originated from students and instructors of a number of courses at the University of Technology, Sydney, who used the LiveNet system both to coordinate their work, and to set up workspaces as part of the students' assignments. The data covers a three month period, with a total of 571,319 log records, They were aggregated into 178,488 higher-level actions in a total of 24,628 sessions involving 721 workspaces and 513 users.

4.4 Space Structuring

During knowledge discovery, using visualization certain of the relationships existing within and between workspaces can be discovered. This particularly aids exploratory analysis, when the purpose is to get an understanding of the structure of, and patterns in, the data. We selected data originating from students of one course who used LiveNet during the mentioned period. There were a total of 187 student users, organized into 50 mostly 3-5 person groups, whose use accounted for about 20% of the above-mentioned log data.

Initial visualization focused on networks of workspaces, to discover how individual student groups partitioned their work in terms of distinct workspaces, and to what extent these workspaces were linked to one another. This exploratory analysis revealed two distinct patterns: the majority of users preferred to use just one workspace to organize all their course work (such as posting drafts of assignment documents, discussing work distribution and problems, etc.). This workspace tended to contain many objects—or have a high *absolute workspace density* [6]. We term such groups *centralizers*. To a certain extent, this mode corresponds to the single-task collaboration mentioned earlier. On the other hand, a few groups tended to partition their work across a collection of connected workspaces, usually with a separate workspace for each major course assignment. These workspaces tended to contain fewer objects (having a lower absolute workspace density) than the ones of the centralizers. We term these groups *partitioners*. Their collaboration style corresponds to the multi-task collaboration.

Figure 10 shows a map of LiveNet workspaces with colours highlighting absolute workspace density—lighter colour indicating lower density, darker colour indicating higher density. Branching out from the central node at the top are networks of workspaces for three groups. Nodes represent workspaces, edges represent hierarchical relationships between workspaces. What the map reveals is that the group on the right, Team40, has a very high density in the workspace used for facilitating its work (the

workspace Team40_Master). Moreover, it uses only one workspace for this purpose. Thus the right group is a typical example of a centralizer. On the other hand, workspaces in the group at the centre have a much lower density. Out of the eight workspaces in this group, six are used for facilitating aspects of the group's work. This is indicative of a partitioner group.

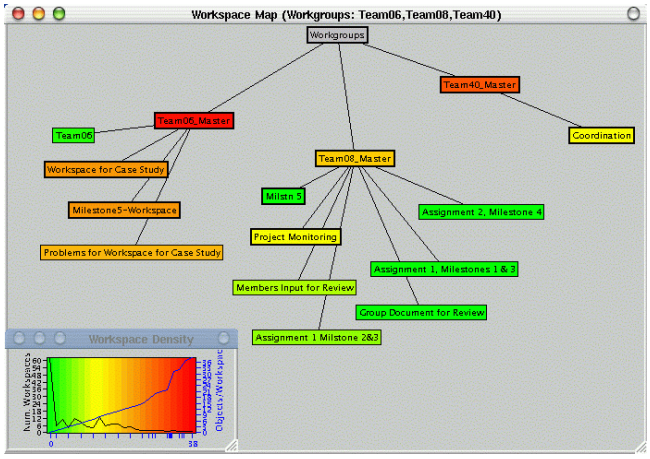


Figure 10. Workspace densities of three different groups

There are plausible explanations for both the centralizer and partitioner cases. Both approaches have their own advantages: in the centralizer case, it is convenience in not having to create multiple workspaces, to switch between them, and in addition to have everything available to all participants in a single location. In the partitioner case, the advantage is increased clarity, structuring according to task, and consequently reduced cognitive load in the case of multi-task collaboration. Furthermore, some groups may bring certain preferences as to the way to organize their work into workspaces and enact these preferences in the way they structure their virtual working environment. When such preferences are recognized during knowledge discovery, and deposited in the organizational memory, they can feed back into the design of new virtual collaboration environments, thus helping to offer more adequate support to cooperative groups with diverse working styles.

4.5 Feasible Actions

A further area we investigated was focused on identifying which actions different groups mainly carried out within LiveNet. All in all, 80 different actions are available in LiveNet. The majority of student groups used only about half of these. The major actions carried out are related to the main LiveNet conceptual elements: workspaces, roles, participants, documents, and discussions. A taxonomy of these actions is presented in Figure 11.

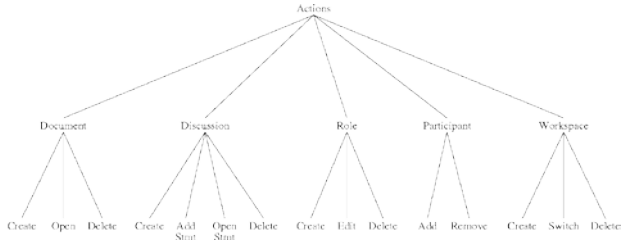


Figure 11. Taxonomy of major high-level LiveNet actions

While all groups had been given the same task—to prepare a number of assignments and to set up a collection of workspaces to support a given process—the way they implemented this task varied markedly. This was evident in a number of aspects of their use of the LiveNet system, such as intensity of use, number of workspaces created, number and length of sessions, number of actions per session, etc. One area of our analysis focused on the proportional distribution of main actions. This revealed that strong differences existed among different groups. To illustrate two examples, Figure 12 shows action distributions among the major high-level actions of the taxonomy of Figure 11 for one group whose distribution of actions was fairly even across categories (with the exception of the participant category): the five major action categories did not vary greatly, none of them exceeding 0.29 of the total (circle size signifies proportion out of the total).

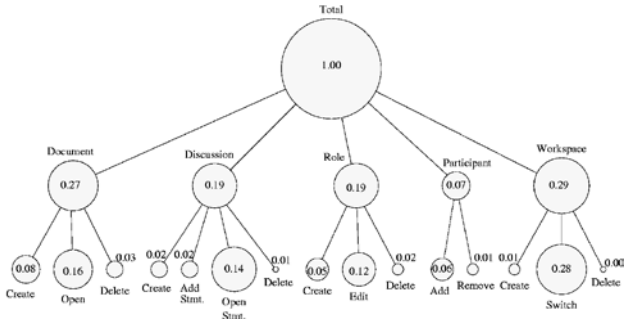


Figure 12. Relatively even distribution of actions in group 1

Figure 13, on the other hand, shows a highly uneven distribution of actions in another group, where one action category (role) strongly dominates with 0.56 of the total, and two other action categories (document and discussion) barely register.

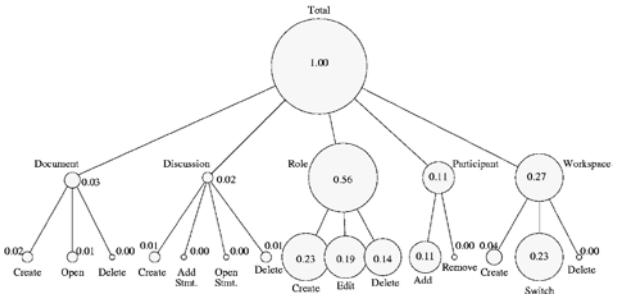


Figure 13. Highly uneven distribution of actions in group 50

This difference may be explained when considering that group 1 (Figure 12) had a total of 627 sessions consisting of a total of 7446 actions, while group 50 (Figure 13) had only 36 sessions and 633 actions. Not only did group 1 use LiveNet much more intensively, but they also made much greater use of the system to facilitate their own work (as manifested in the solid proportion of actions in the document and discussion categories). Thus the skew in action distribution towards role-related actions on the part of group 50 is caused by the under-utilization of other LiveNet features, not by an absolute high number of actions related to roles (in absolute terms, group 1 carried out 431 role-related actions, while group 50 carried out only 142 such actions). It should be noted that the choice of these two groups for illustration was not coincidental: group 1 was the best-performing group in

the course, while group 50 was the worst-performing group, as measured in the marks obtained for their assignments in the course, one of which involved heavy use of LiveNet. The situation was comparable in other similarly scoring groups.

When such cases are identified and included in the organizational memory as part of a record of collaboration, they can be of use in evaluating virtual work. This can be particularly useful with fully virtual teams that never meet face-to-face, where conventional management methods for project monitoring and control are severely limited or absent. The organizational memory thus takes on the additional role of a management instrument.

5. CONCLUSIONS

Collaborative virtual environments have the potential to change the way we work. Unfortunately, earlier observations of human activities in them uncovered very few aspects of computer-mediated collaboration. The new generation of environments has the potential to produce vast amounts of data about collaboration. Data mining technologies offer instrumentation capable of extracting semantic information with the potential to turn collected data into valuable assets. The integration of collaborative environments with data mining technologies provides unique opportunities to unveil some secrets in the art of human collaboration.

The framework presented in the paper looks at the integration of data mining technologies in collaborative virtual environments at the early design stages of the virtual environment. A key issue at the design stage is the selection of the data that should be recorded. These records are complementary to the standard logs of the web server. Careful design and analysis of this log have the potential to lead to improvements of the structure of the space and tuning the set of feasible actions with respect to the purpose of the environment. The applicability of the framework has been tested and demonstrated on a real environment.

An important part of the framework is the way knowledge is returned back to the environment. The framework allows also a feedback from the organizational memory towards modification of the knowledge representation schema, used for representation and incorporation of discovered knowledge. The detailed discussion of the issues related to the modification of the knowledge representation schema, however, are beyond the scope of this paper.

6. ACKNOWLEDGMENTS

This research has been supported by the Australian Research Council, the University of Technology, Sydney and the University of Macau.

7. REFERENCES

[1] Ackerman, M.S. Augmenting the organizational memory: A field study of Answer Garden. In *Proceedings of the Conference on Computer Supported Cooperative Work*, 1994, 243-252.

[2] Ansari, S., Kohavi, R., Mason, L., and Zheng, Z. Integrating e-commerce and data mining: Architecture and challenges. In *WEBKDD 2000 Workshop: Web Mining for E-Commerce – Challenges and Opportunities*, 2000.

[3] Bannon, L.J., and Kuutti, K. Shifting perspectives on organizational memory: From storage to active remembering. In *Proceedings of the Twenty-Ninth Hawaii International Conference on System Sciences*, 1996, vol. 3, 156-167.

[4] Berthold, M.R., Sudweeks, F., Newton, S., and Coyne, R. Clustering on the net: Applying an autoassociative neural network to computer-mediated discussions. *Journal of Computer Mediated Communication 2*, 4 (1997). (Available at: http://www.ascusc.org/jcmc/vol2/issue4/berthold.html).

[5] Berthold, M.R., Sudweeks, F., Newton, S., and Coyne, R. It makes sense: Using an autoassociative neural network to explore typicality in computer mediated discussions. In Sudweeks, F., McLaughlin, M., and Rafaeli, S., eds. *Network and Netplay: Virtual Groups on the Internet*. AAAI/MIT Press, Menlo Park, CA, USA, 1998, 191-220.

[6] Biuk-Aghai, R.P., and Hawryszkiewycz, I.T. Analysis of virtual workspaces. In Kambayashi, Y., and Takakura, H., eds. *Proceedings of the 1999 International Symposium on Database Applications in Non-Traditional Environments*, 1999, 325-332.

[7] Capin, T.K., Pandzic, I.S., Magnenat-Thalman, N., and Thalman, D. *Avatars in Networked Virtual Environments*. John Wiley and Sons, Chichester, 1999.

[8] Chen, C. *Information Visualization and Virtual Environments*. Springer-Verlag, London, UK, 1999.

[9] Conklin, E.J. Capturing organizational memory. In Baecker, R.M., ed. *Readings in Groupware and Computer-Supported Cooperative Work: Assisting Human-Human Collaboration*. Morgan Kaufmann Publishers, 1993, 561-565.

[10] Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. From data mining to knowledge discovery: An overview. In Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., eds. *Advances in Knowledge Discovery and Data Mining*. AAAI Press/MIT Press, Menlo Park, California, USA, 1996.

[11] Furst, S., Blackburn, R., and Rosen, B. Virtual team effectiveness: A proposed research agenda. *Information Systems Journal 9*, 4 (1999), 249–269.

[12] Greenhalgh, C. *Large Scale Collaborative Virtual Environments*. Springer-Verlag, London, UK, 1999.

[13] Gutwin, C., and Greenberg, S. Design for individuals, design for groups: Tradeoffs between power and workspace awareness. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, 1998, 207-216.

[14] Hawryszkiewycz, I.T. Workspace networks for knowledge sharing. In Debrency, R., and Ellis, A., eds. *Proceedings of AusWeb99, The Fifth Australian World Wide Web Conference*, 1999, 219-227. (Available at: http://ausweb.scu.edu.au/aw99/papers/hawryszkiewycz/paper.html).

[15] Ishii, H., Kobayashi, M., and Grudin, J. Integration of interpersonal space and shared workspace: Clearboard design and experiments. In Greenberg, S., Hayne, S., and Rada, R., eds. *Groupware for Real-Time Drawing: A Designer's Guide*. McGraw-Hill, Berkshire, England, 1995, 96-125.

[16] Knuth, D.E. *The Art of Computer Programming, Vol 1: Fundamental Algorithms*. Addison-Wesley, Reading, MA, USA, 1973.

[17] Maher, M.L., Simoff, S.J., and Cicognani, A. Potentials and limitations of virtual design studio, *Interactive Construction On-Line*, a1 (1997).

[18] Maher, M.L., Simoff, S.J., and Cicognani, A. *Understanding Virtual Design Studios*. Springer-Verlag, London, UK, 2000.

[19] Maher, M.L., Simoff, S.J., Gu, N., and Lau, K.H. Designing virtual architecture. In *Proceedings of CAADRIA2000*, 2000, 481-490. (Available at: http://www.arch.usyd.edu.au/ ~chris_a/MaherPubs/2000pdf/caadria2000. pdf)

[20] Simoff, S.J. Monitoring and evaluation in collaborative learning environments. In *Computer Supported Collaborative Learning*, 1999. (Available at: http://kn.cilt.org/cscl99/A83/A83. html).

[21] Simoff, S.J., and Maher, M.L. Analyzing participation in collaborative design environments. *Design Studies 21*, (2000), 119-144.

[22] Simoff, S.J., and Maher, M.L. Loosely integrated open virtual environments as places. *Learning Technology 3*, 1 (2001). (Available at: http://lttf.ieee.org/learn_tech/issues/ january2001/index.html#3).

[23] Spiliopoulou, M., and Pohle, C. Data mining for measuring and improving the success of web sites. *Data Mining and Knowledge Discovery 5*, 1/2 (2001), 85-114.

[24] Sudweeks, F., and Simoff, S.J. Complementary explorative data analysis: The reconciliation of quantitative and qualitative principles. In Jones, S., ed. *Doing Internet Research*. Sage Publications, 2000, 29-5.