Modeling the Dynamics of Using a Collaborative Hypertext

Chaomei Chen[1] and Roy Rada[2]

[1]*Department of Information Systems & Computing, Brunel University, Uxbridge UB8 3PH, UK;*[†]
[2]*Department of Information Systems, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250*[‡]

ABSTRACT

The dynamics of using a collaborative hypertext system is analyzed and modeled, based on time series data from the real users. First, we studied how users access a shared workspace simultaneously. Traditional time series modeling methods, notably multiple linear regression and Box-Jenkins ARIMA modeling techniques, are used. Second, we analyzed the relation between awareness-seeking events and a variety of structural properties associated with such events through a Poisson regression model. This process-oriented approach complements analytical and modeling methods concerning the dynamics of group work.

KEYWORDS

dynamic modeling, collaborative writing, interlocking behavior, time series analysis, sequential behavioral analysis, Poisson models

[†]tel: +44 1895 203080; fax: +44 1895 251686
e-mail: chaomei.chen@brunel.ac.uk
[‡]tel +1 (410) 455-2645; fax +1 (410) 455-1073
e-mail: rada@umbc.edu

1. INTRODUCTION

Protocol analysis is a central part of modeling complex processes of collaborative work. Currently, the bottleneck lies in modeling the situated nature of a collaborative process. The sheer amount of data and the diverse range of events and their contexts often lead to a time-consuming and tedious process. Automated approaches are motivated to shift the burdens from analysts, allowing them to concentrate on more critical issues.

Interest is growing in process-oriented approaches to studying the interplay between group dynamics and information technology (Delise & Schwartz, 1987; Dourish & Bellotti, 1992; Galegher & Kraut, 1994; Losada et al., 1990; Whittaker et al., 1993). Notable examples include Exploratory Sequential Data Analysis (ESDA) (Sanderson & Fisher, 1994), Automated Protocol Analysis (APA) (Smith et al., 1993), and Statistical and Grammatical Analysis of Sequential Interactive Behaviors (Olson et al., 1994). These approaches are primarily motivated to automate and ease the time-consuming process of deriving salient structures underlying complex processes. These approaches play a significant role in task analysis, protocol analysis, system design, and evaluation. They also provide an empirical basis for the development of intelligent systems.

Collaborative authoring in general is dynamic and diverse to model, usually involving a wide range of interwoven tasks and activities. The highly situated nature of collaborative authoring with computer systems has been widely acknowledged (Delise & Schwartz, 1987; Irish & Trigg, 1989; Neuwirth et al., 1990; Posner & Baecker, 1992). This situated nature presents a significant challenge for the designers and developers of collaborative authoring support systems. It is crucial to understand how collaborative authoring takes place in practice and what characteristics can be supported in a computerized environment.

In this article, we explore a process-oriented approach to modeling the dynamics of using a shared information space over time. We analyze the use of a collaborative hypertext authoring system and focus on temporal characteristics of user-driven events. We us time series analysis and modeling techniques to reveal underlying factors that influence these characteristics. Furthermore, we investigate not only the relation between the structure of a shared workspace and the distribution of hypertext nodes activated by

concurrent users but also the connection between awareness-seeking events and the structure of the underlying hyperspace. The notion of interlocking behavior is introduced to emphasize the situated nature of collaborative authoring.

1.1 Interlocking Behavior

Our approach in this study shares some principles of a generic approach called *Exploratory Sequential Data Analysis* (ESDA) (Sanderson & Fisher, 1994). ESDA emphasizes the role of exploratory analysis with respect to confirmatory analysis. ESDA helps researchers to generate hypotheses, which can then be tested with formal confirmatory procedures using inferential statistics.

User behavior evolves over time. Time series analysis aims to reveal temporal structures that are associated with various user-driven events. Time series analysis assumes a stochastic process underlying the time series. In this case, for example, collaborative authoring typically involves a large number of factors that are unknown or uncontrollable to the investigators.

Time series analysis and modeling techniques have been applied to the characterization of the behavior of dynamic systems in many disciplines, such as engineering, ecology, and applied statistics. A time series is a series of values of a single variable observed regularly over a period of time (Ostrom, 1978; Box & Jenkins, 1976). Time series analysis aims

1. to derive patterns from a time series based on its behavior in the past,
2. to predict its behavior using derived mathematical models, and
3. to estimate the impact of underlying factors on given time series.

There are three types of basic time series models: auto-regressive models, parameter models, and non-parameter models. Spectral analysis typically generates non-parameter models. In this article, we consider Box-Jenkins ARIMA time series models (Box & Jenkins, 1976; Guster & Robinson, 1993) in Part I and Poisson regression models (Ostrom, 1978; King, 1994) in Part II. Poisson regression models belong to *Generalized Linear Models* (GLIMs), which is a relatively new area of mathematical statistics.

Classic examples of time series models include second-order auto-regressive models of annual sunspot events and the first-order stationary stochastic differential equation model of turbulent flows. More recent examples include tractable sequential data analysis of air traffic control (ATC) tasks (Vortac et al., 1994), state-action transition models of human-system interaction (Cooke et al., 1996), and auto-regressive time series models of air traffic control (Edwards et al., 1985).

An interesting topic in time series modeling is the modeling of interlocking behavior or connections between wave-like structures (Losada et al., 1990). A wave-like structure is a type of self-repeating pattern that repeats itself after certain time intervals as in a periodic function as follows:

$$f(x) = f(x+kp),\ k=\{\ldots, -2, -1, 0, 1, 2, \ldots\}$$

Our approach offers a systematic method for studying the temporal characteristics of time series. This approach has the potential of bridging qualitative and quantitative studies of collaborative authoring systems. In this article, our approach is illustrated through two examples of time series analysis of a collaborative authoring system:

- What is the relation between the organizational structure of a shared workspace and areas activated by multiple users concurrently?
- In what context are people likely to invoke a group awareness function in such shared spaces?

The first issue is investigated in Sec. 2 and the second issue in Sec. 3, which focuses on an interlocking behavior. Both studies are based on the use of a collaborative authoring system, *Multiple Using Collaborative Hypermedia* (MUCH). An overview of the MUCH system is presented as follows.

1.2 The MUCH System

The MUCH system was designed for collaborative writing among a group of distributed users (Chen et al., 1994; Zheng & Rada., 1994). Asynchronous collaborative writing was the major mode of collaboration. The architecture of the system is shown in Fig. 1.

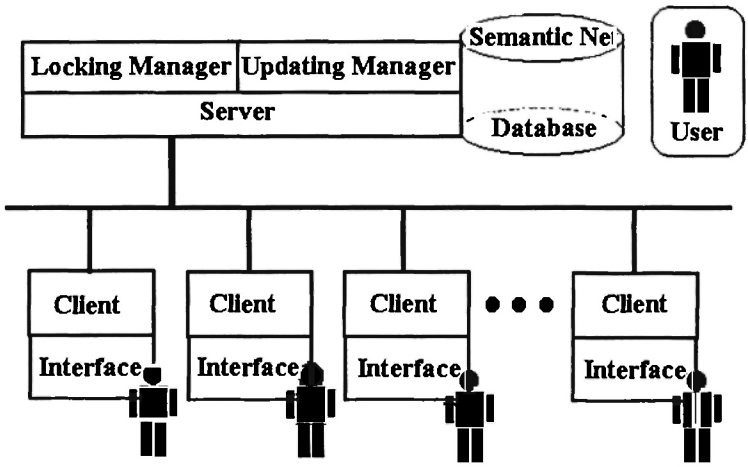

Fig. 1: The client-sever architecture of the MUCH system

1.3 Hypermedia-Based Writing Model

Figure 2 shows the user interface of the MUCH system. It includes an outline window and a content window. The outline window displays a hierarchical view of the underlying shared information space. The content window displays the content of an activated node.

The MUCH system is based on the Dexter Hypertext Reference Model (Halasz & Schwartz., 1994.). The Dexter model consists of three layers: the presentation layer on the top, the storage layer in the middle, and the within-component layer at the bottom. In the MUCH system, the top-level view of the underlying semantic network is a hierarchy generated by graph traversal algorithms, especially based on depth-first search (DFS). Figure 3 shows traversal options and the outcome of the traverse.

1.4 Collaborative Writing with Hypertext

In collaborative writing, co-authors have to communicate and coordinate their work (Whittaker et al., 1993). The MUCH system maintains the history

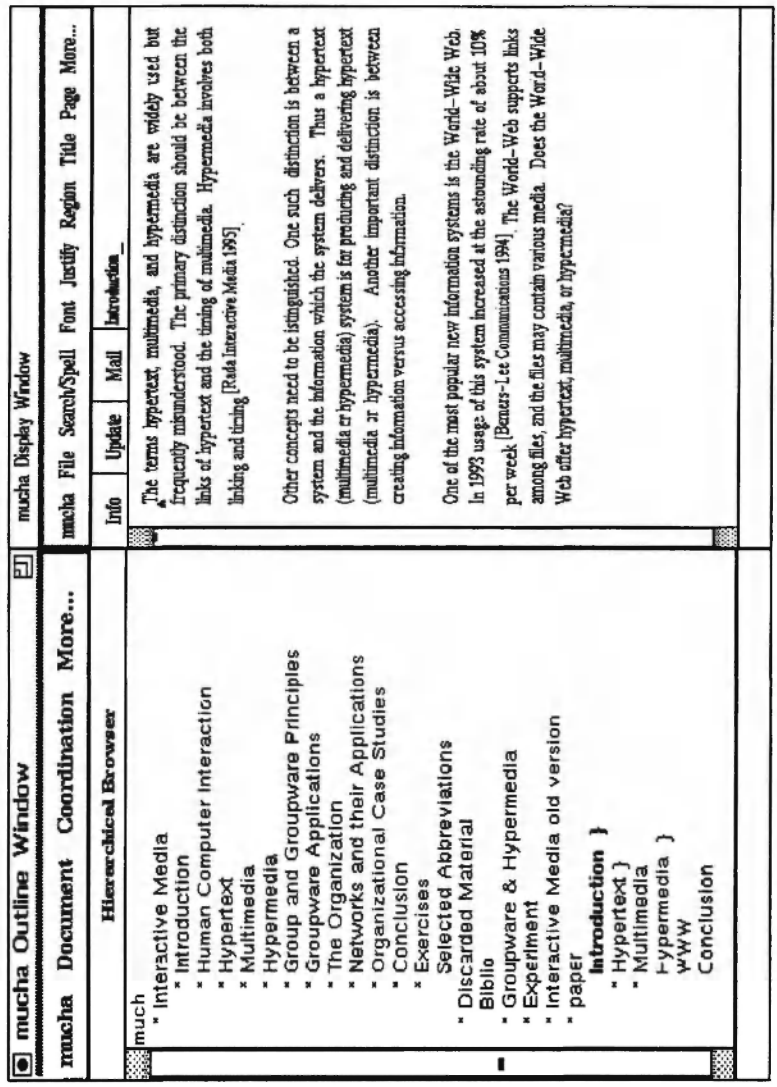

Fig. 1: The user interface of the MUCH system.

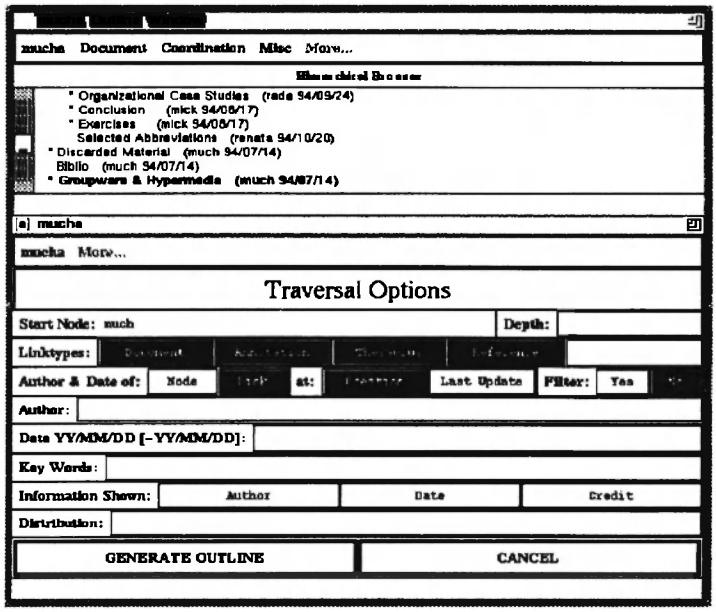

Fig. 2: A traversal algorithm and the resultant hierarchy of the traversal.

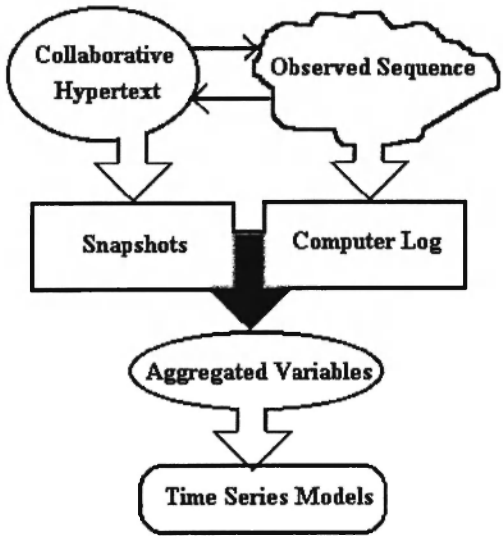

Fig. 3: Structural snapshots and interactive events are matched to derive the time series model.

of each shared workspace, which is usually a hypermedia network. Historical information is available to all the users; for example, who created a node in the first place, who recently edited the content of the node, and how many times it has been visited by others. Users can view the information by invoking a function called INFO.

The MUCH system keeps track of the information, which may improve the mutual intelligibility among co-authors of a collaborative hypertext. The system maintains the name of the user who created a node, the names of the users who have made changes to the content of the node, and how many times users other than the authors (namely, the creator and all the subsequent writers) have visited the node. The information is attached to each node in the information space. Users can view the evolutionary information with a function called INFO.

Concurrent accesses to the shared workspace are controlled by a locking manager in the MUCH system. The content of a hypertext node can be changed only by one user at a time. A locked node will become read-only to other users. Other users cannot modify the node until the node is unlocked. The MUCH system maintains a relaxed *What-You-See-Is-What-I-See* (WYSIWIS) interface for its concurrent users. The content of the current node will be refreshed on the screen if users explicitly invoke the UPDATE function. The MUCH system also provides several functions for users to handle the underlying semantic-network and its contents. These functions largely influence how users interact with the collaborative hypertext. The main functions include the following:

- *UNFOLD*—to expand the existing hierarchical view from a node entry by a one-step forward breadth-first traversal. All the nodes associated with the focal node will appear in the view;
- *READ*—to load and display the content of a node double-clicked in the hierarchical view;
- *LOCK*—to lock the content of a selected node for editing, but concurrent users still can read the content with the READ function;
- *INFO*—to display the history of a currently visited node;
- *UPDATE*—to save changes since the last UPDATE or since the latest LOCK and to release the updated content to the update manager in the

server, which will subsequently update the screens of all the concurrent users who are on the same node.

2. LINEAR REGRESSION MODELS

The MUCH system has been used by more than 300 users over the past few years on various collaborative writing projects. For example, a group of researchers at the University of Liverpool have used the MUCH system extensively in various research projects involving collaborative writing. The users are located in their own offices within the same building. Many of them work on their own timetable. The MUCH system enabled these researchers to maintain a persistent and shared workspace. The MUCH system was also used by undergraduate students as part of their courses. In a typical course, students would use the MUCH system to write essays collaboratively in their three-person groups. Experiences with the previous versions of the MUCH system are summarized as follows to provide a context of our modeling work.

2.1 Experiences with Previous Versions

Previous experience with the MUCH system has highlighted several issues concerning interacting with a hypertext document space (Chen et al., 1994; Chen, 1997). For example:

- What collaboration models should the users adopt?
- How do they communicate and co-ordinate their work?
- How often does a user need to check the status of the shared workspace?
- What are the patterns of changing in terms of the organization and the content of a shared workspace in the MUCH system?

The MUCH system maintains a structural overview of all active workspaces. The experiences in the past have suggested some limitations of the overview:

- The structural overview alone does not reveal the relevance of different parts of the workspace.
- Users need a facility that can inform them the changes in the shared workspace, especially by whom and when.

The MUCH system was essentially designed to support collaboration in an independent or a loosely coupled mode rather than in a tightly coupled mode. It is important for users to keep track of the changes in the shared workspace, not only with finer granularity but also with why particular changes were made. The MUCH system provides a facility called the INFO function to help users discover the status of the current hypertext node in the shared workspace.

To identify which are the key functions of the MUCH system on which users rely in their sessions, we first constructed multiple linear regression models of time series in this section. Conventional linear regression models are subsequently improved by using generalized linear models in Sec. 3.

Time series analysis can be found in a variety of scientific disciplines. For example, Losada et al (1990) studied time series of interactive behavior in meetings. Guster and Robinson (1993) used time series models to determine the dynamic characteristics of the workload of an IBM 4381 in a multi-user environment. In the following study, we focus on the access locality issue in terms of how many unique nodes are activated in a time series of 1-h intervals. Auto-Regressive Integrated Moving Average (ARIMA) models are derived from the time series by using SPSS *Trends*.™ ARIMA models are also known as Box-Jenkins models. The Box-Jenkins (1976) method is the most popular time series modeling method.

2.2 Data Collection and Analysis

The time series data came from two sources: usage data and snapshots of shared workspaces. Usage data recorded the occurrences of user-driven events. Snapshots of the organizational structure of each shared workspace were taken monthly. The snapshots set the context in which a particular event occurred (see Fig. 4). Three months of usage data were collected from the MUCH system. Each user-driven event was recorded, including a time stamp, the name of the user, the title of the activated hypertext node. Sample data were drawn an 8-h window from Monday to Friday, from 9:00 through 17:00. Transactions outside this window were excluded from the subsequent analysis.

User-driven events were mapped into the original context recorded in these snapshots. The following structural variables, derived from the structural

snapshot data, were regarded as random variables that characterize the context of a user-driven event:

- LEVEL(n)—the level of the node *n* in the overview of a given workspace;
- LINK(n)—the number of children that the node *n* has in the hierarchical structure;
- CREDIT(n)—the number of times that the node *n* has been accessed by collaborative users rather than its creator;
- UPDATE(n)—the number of times that the node *n* has been updated so far;
- SIZE(n)—the size of a textual node *n*, such as the total number of lines, words, and characters contained in the node *n*.

Within each one-hour time interval, the frequency count of a particular type of event was the total number of events recorded within the interval. The time series of this type event was defined as the sequence of frequency counts for all the consecutive time intervals.

ARIMA models explain the characteristics of a time series based on its past behavior. The Box-Jenkins (1976) method has three major stages: identification, estimation, and diagnostic checking. As a standard procedure, before the analysis, non-stationary time series were stabilized by transformations and differences. An ARIMA model can be identified based on a corresponding auto-correlation function (ACF) and partial auto-correlation function (PACF). AIC and SBC were used as the criteria in model selection.

First, the time series of unique nodes activated, the NODE time series, was modeled through the conventional multiple linear regression method. An ordinary least-squares regression model (OLS) assumes, among other things, that the residuals of the model must be independent, or at least essentially independent and have a zero mean. This assumption is often violated in time series data because of auto-correlation in the time series. As a result, the regression model still has the correct coefficient estimates; the significance level of these estimates is biased and appears more significant than it really is (Ostrom, 1978; King, 1994). It is necessary to incorporate this type of characteristic into the regression model. ARIMA time series regression models provide a solution to restore the validity of regression models, and the

Durbin-Watson test indicates whether first-order autocorrelation is present in the residuals.

Another requirement for a multiple linear regression model is the absence of perfect multicollinearity. None of the independent variables should be perfectly correlated with another independent variable or linear combination of other independent variables. The set of chosen independent variables should not include the redundant variables. We adopt a commonly used method of assessing multicollinarity. Each independent variable in the initial regression model was regressed on all the other independent variables in the model. The independent variable with the highest R^2 was dropped from the model, and this step was repeated until none of the independent variables in the remaining model significantly overlapped. Finally, a fundamental assumption was that the NODE time series was approximated by normal distribution; therefore, it is appropriate to use the OLS method.

2.3 Results

The results include regression models of users working with a shared workspace, known as the *class* workspace, in the MUCH system. The shared workspace *class* in the MUCH system was created and used by M.Sc. students. Fig. 4 shows the number of concurrent users in 1-h time intervals (9:00 to 17:00, Monday through Friday). High spikes indicate the peak use of the database during lectures given in the lab. Toward the end of the course, the class actually had re-scheduled to have two or three consecutive lectures in the same day as shown in the figure.

Medium-sized spikes, in particular, those between Case Number 60 and 80, corresponded to group essay writing. In contrast, fewer such spikes were observed outside the regular lecture spikes in the first stage of the course. Fig. 5 is the time series plot for the NODE variable along time. The NODE time series showed a similar pattern to the size of the group, but the magnitude of the peaks decreased after the first lecture. Compared with the medium spikes in Fig. 4, the group essay writing stage involved a relatively small number of unique nodes.

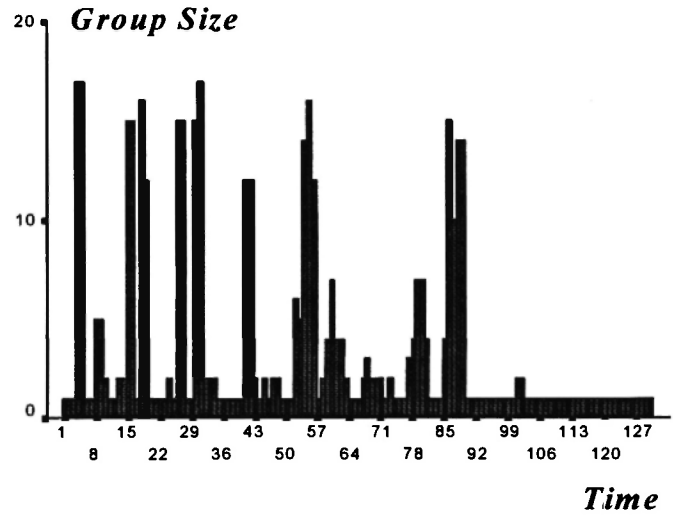

Fig. 4: The number of concurrent student users in one-hour time intervals.

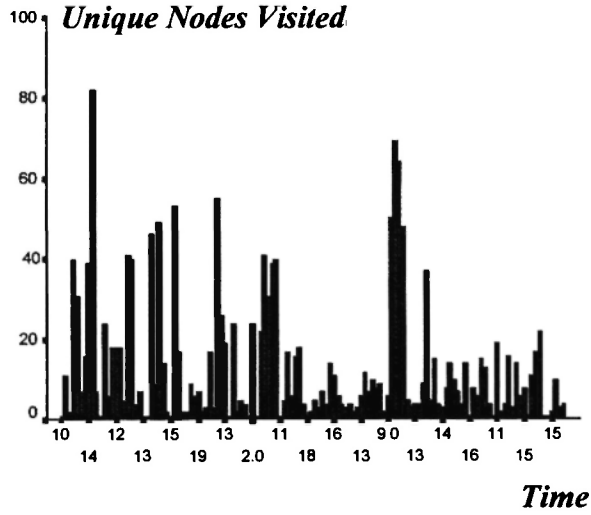

Fig. 5: The number of unique nodes visited in one-hour intervals.

An OLS regression model was built by selecting (stepwise) the most representative explanatory variables into the model from both usage variables and structural variables. This regression model explains 63% of the variance in the NODE time series (Adj. R2 = .63; Durbin-Watson D = 1.45).

The model includes two explanatory variables: READ and DELETE, but structural variables such as LEVEL, LINK, and GSIZE failed to enter the model. Usage events MODIFY and CREATE were also excluded from the model. Durbin-Watson statistic of d=1.45 is outside the 5 percent significance interval for k=2:

$$d = 1.45 \notin [dL = 1.63, dU = 1.72]$$

The null hypothesis—that the residuals are not auto-correlated at the p=.05 level—was rejected according to the Durbin-Watson test. The residuals are first-order auto-correlated. In SPSS *Trends*™, the AREG procedure for auto-regression was used for incorporating this characteristic of auto-correlation into the regression model.

Table 1

Coefficient estimates for the OLS regression model

Variable	B	SE B	Beta	T	Sig T
READ	.020005	.003358	.720186	5.958	.0000
DELETE	-.657813	.289186	-.274946	-2.275	.0321
(Constant)	3.240313	.422020		7.678	.0000

The square root of the NODE time series was auto-correlated at a small-to-medium level (ϕ_1= .2046). In other words, about 20% of the unique-node counts in the previous time interval were carried over into the current time interval. The time series regression model of the number of unique nodes activated is given by Eq. (1). This is a first-order auto-regressive model with two independent time series. This model explained nearly 70% of the variance in the node selection time series.

$$SQRT(Y_t) = 2.79 + 0.20 \times SQRT(Y_{t-1}) + 0.02 \times READ_t - 0.45 \times DELETE_t \quad (1)$$

2.4 Discussions

Regression models of the unique-node time series revealed the overall interrelationship between usage variables and structural variables. The moderate first-order auto-correlation coefficient (ϕ_1 = .205) indicated that 20% of the unique node counts in one interval is likely to be carried over into the next interval.

READ and DELETE appear to be the best predictors for the number of unique nodes activated, explaining 63% of the variance. READ was one of the browsing techniques frequently used with the MUCH system. DELETE events give the lower bound of the number of unique nodes activated in a given time interval.

Structural variables such as GSIZE, LEVEL, and LINK were expected to explain a considerable amount of variance of the time series of unique nodes. Nevertheless, the resultant model does not include them. To understand the behavioral patterns of the interactive process, one needs to study collaborative writing within an integrated environment. The MUCH system provides limited features of an integrated environment. In addition, people's social interaction, meetings, and discussions off-line are not captured to the same extent as are the actions that take place within the collaborative hypertext system. Some practical and less technical issues also arise in the use of a shared information space.

3. MODELING INTERLOCKING BEHAVIOR

Researchers in collaborative writing have classified the nature of interaction into three modes: the independent mode, the loosely coupled mode, and the tightly coupled mode. The engagement between collaborators is minimum in the independent mode and maximum in the tightly coupled mode. The MUCH system provides users with an INFO function on each hypertext node. The user can use the INFO function to find out the history of the currently visited node, including who created the node and who has subsequently modified the node (see Fig. 6). The INFO function, therefore, is a key concept for investigating characteristics that are associated with transitions from one mode to another.

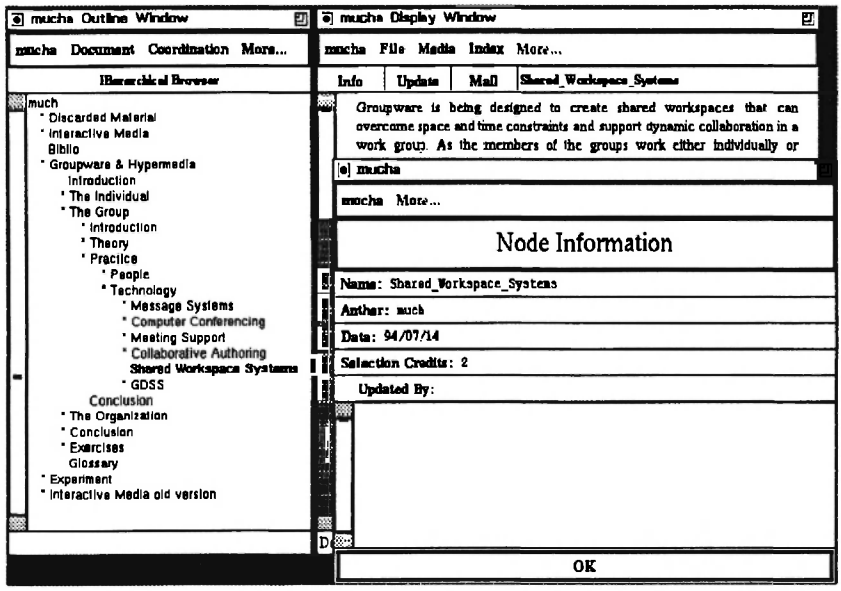

Fig. 6: The display of the Node Information returned by the INFO function

In the following study, we focus on time series analysis of usage events of the INFO function. Compared with other events in the MUCH system such as READ, the occurrences of INFO events were rare. The validity of regression models requires that variables concerned must have a normal distribution. Nevertheless, in general this assumption does not hold for frequency counts of rare events because the true Poisson distribution of such events cannot be adequately approximated by normal distribution (Ostrom, 1978.; King, 1994). Therefore, Poisson regression models are used to characterize interrelationships between the INFO time series and the context of such events. We characterize the occurrences of INFO events in the context of the hierarchical structure of an underlying workspace. Such relationships would be valuable for users to keep track of the evolution of a shared workspace and to coordinate more effectively with collaborators.

We focused on the most influential factors that could explain the occurrences of INFO events. For example:

- At what levels of the hierarchical structure are users more likely to use the INFO function, higher levels or lower levels?
- Is there a connection between the frequency of INFO events and the number of children nodes the current node has?
- Is the INFO function more frequently used on large-sized nodes?

INFO events at lower levels of the hierarchical structure may correspond to more detailed and substantial information, whereas INFO events at higher levels of the structure may correspond to broader issues that are involved in collaborative writing. In the following analysis, we will address these questions from a situated-action point of view.

3.1 Related Work

Two different types of awareness arise in the use of CSCW systems: awareness concerning to the activity of other users in general, and awareness that focuses on the changes made by others in particular artifacts. According to Markova (1987), human awareness can be characterized as follows:

- people's knowledge of their own agency and of that of others;
- people's ability to monitor events in their own lives and to make decisions about their own future on the basis of that knowledge; and
- people's ability to communicate their awareness of themselves and others to other human beings.

Awareness issues arise naturally in collaborative writing with a shared hypertext database. The context in which an INFO event occurs reflects the interest of the user in awareness concerning who has contributed to a shared workspace. Interrelationships between the INFO time series and structural characteristics of the underlying workspace can lead to insights into the modeling of situations where users have to find out the status of a shared workspace.

Generalized Linear Models (GLM) comprise a set of statistical methods for analyzing and modeling discrete dependent variables (McCullagh & Nelder, 1989). These methods model how likely an event is to occur in terms of the probability of an event. As the INFO time series followed a Poisson distribution, we investigated whether the occurrences of INFO events are related to the structure of shared workspaces with a Poisson regression model.

3.2 Data Collection and Analysis

First, we established that the occurrences of INFO events followed a Poisson distribution. A Poisson probabilistic regression model was fitted using general log-linear models (GENLOG). The occurrences of INFO events were matched to users' positions in the workspace when such events occurred. The position of a user in the workspace at a particular time was determined by a number of measures, including the depth of the current hypertext node in the hierarchical structure, how many structural links are available from the node, the latest selection credits, and the size of textual nodes.

Similar to the regression study described earlier, time series data were sampled from two sources: observed INFO events and the snapshots of the structure of the corresponding workspace. Each INFO event was recorded with a time stamp, the name of the user who activated the function, and the hypertext node on which the INFO event took place. The occurrences of INFO events were matched to the structural snapshots such that each INFO event was associated with the context of the event as well as with the event itself.

Let n denote a hypertext node in a given workspace S. Let $n.attr$ denote the value of an attribute $attr$ of node n. The depth of node n in the hierarchical structure is denoted as $n.depth$. The depth of the root node of a hierarchy is 0. The depth of a child node of the root is 1. The time series of INFO events in the i^{th} interval corresponds to a sequence of nodes $\{n_{i1}, n_{i2}, \ldots n_{ik}\}$, on which these events occurred. The depths of these nodes were aggregated within each interval. Equation (2) represents the aggregation procedure.

$$s_i(depth) = \sum_{j=1}^{k} n_{ij}.depth \tag{2}$$

Several structural time series were derived from the combined time series, including Zdepth, Zcredit, and Zword. The prefix Z in these variable names indicates that they were standardized scores. For example, $Zdepth_i$ was defined in Eq. (3) by the average depth of the observed path in the i^{th} time interval.

$$Zdepth_i = \frac{s_i(depth)}{s.d.} \tag{3}$$

where s.d is the standard deviation of the sequence $\{s_i(depth)\}$. Other explanatory variables were calculated similarly.

The analysis of the INFO time series was based on a Poisson regression model in the context of corresponding structural properties. A Poisson process is frequently used to model the occurrences of rare events in the world. A Poisson regression model can be specified via a log-linear model, in which the dependent variable follows a Poisson distribution. In this case, the INFO time series must follow a Poisson distribution. A K-S test on the INFO time series indicated that these INFO events indeed followed a Poisson distribution. The general log-linear model analysis procedure GENLOG in SPSS for Windows was used. The cell values of the contingency table that are associated with the log-linear model represent the occurrences of INFO events.

Given a vector of explanatory variables $X_t = (X_{1t}, X_{2t},\ldots,X_{kt})$ and a vector of partial correlation coefficient $\beta = (\beta_1, \beta_2,\ldots,\beta_k)$, a formal expression of a Poisson regression model is given in Eq. (4).

$$P(Y_t) = \frac{\lambda_t e^{-\lambda_t}}{Y_t!} \quad Y_t = 0,1,2,\ldots; \quad t = 1,\ldots n. \tag{4}$$

where $\lambda_t = \exp(X_t\beta s)$. Let Y(t) denote the occurrences of INFO events in a one-hour time interval t, $X_i(t)$ the value of the i^{th} explanatory variable with its coefficient β_i in the same time interval t, the Poisson regression model is specified in Eq. (5).

$$ln(Y_t) = \sum_{i=1}^{k} \beta_i X_i(t) + c \tag{5}$$

The explanatory variables were transformed to standard scores. An adequately fitted Poisson regression model was subsequently selected to represent the context-behavior relationship.

3.3 Results

The INFO time series was studied as situated-actions via the relationship between the occurrences of INFO events and several contextual, explanatory

variables of the corresponding nodes where INFO events took place. An INFO event can be seen as an indication of awareness needs.

3.3.1. *Overall usage profile of shared workspaces*: The time series analyzed involved a total of five shared workspaces in the MUCH system, consisting of thousands of transactions during a 3-mo period of a longitudinal observation. The *class* workspace was accessed by the greatest number of users (User = 30), with the widest spectrum of functions from the presentation layer—the use of 25 distinct functions was observed. Table 2 summarizes the overall usage profile of each shared workspace, including the total number of transactions observed, the total number of unique nodes involved, the size of the user group, and the number of distinct functions used.

Table 2

Usage of the 5 shared workspaces over 3 consecutive months.

Workspaces	Transactions	Nodes	Users	Functions	Started	Ended
Groupware	12,089	750	11	22	Apr	July
Class	9,247	474	30	25	Apr	July
Management	7,068	698	16	22	Apr	July
OSCAR	3,148	315	7	18	Apr	July
General	1,590	96	10	21	Apr	July

Table 3 lists the use of each function. Some heavily used functions appear to entail some cognitive and behavioral overheads, due to particular design decisions for the presentation layer of the MUCH system.

The UNFOLD function was heavily used in all but one workspaces. On the other hand, an excessive use of the UNFOLD function in a particular workspace may indicate that users were largely browsing the workspace rather than carrying out authoring tasks.

The frequencies of INFO events are strongly correlated with those of MODIFY, followed by READ, UNFOLD, and DELETE, which is no surprise—the more modification involved within a workspace, the more likely

Table 3

The usage of various functions with 7 workspaces

wkspace	total	unfold	fold	read	create	delete	modify	rename	lock	update	info	users
gware	8033	2119	890	2404	53	108	115	71	1033	935	62	2
reuse	4172	990	362	1243	6	2	2	24	738	578	16	6
man	6695	2681	627	1893	116	15	34	14	491	390	44	5
general	1585	237	67	388	55	10	3	2	343	334	7	7
oscar	2746	913	307	688	16	29	10	27	325	292	24	3
class93	1680	721	134	441	13	11	7	28	141	118	4	0
class94	1170	462	155	269	10	20	12	4	72	62	26	5

Table 4

Correlations between INFO frequencies and other functions.

Functions	Correlation
modify	0.887
read	0.861
unfold	0.817
delete	0.802
update	0.679
lock	0.663
rename	0.639
create	0.513
users	-0.135

that one would also look at the history of various nodes in the workspace. A deeper interrelationship between INFO and structural properties, however, can be explained to some extent with our Poisson model.

3.3.2. *Awareness in context.* The GENLOG procedure in *SPSS for Windows* revealed that the model in Eq. (6) adequately represented the data (Pearson χ^2 = .0960, DF = 1, P = .76). The explanatory variables, or independent variables, were specified as the covariant variables in the model, except the hour variable, which is categorical.

$$\ln(\text{count}) = \beta_0 + \beta_1 \times \text{Zcredit} + \beta_2 \times \text{Zdepth} + \beta_3 \times \text{Zlink} + \beta_4 \times \text{Zword} + \beta_5 \times \text{Zsize} + \beta_6 \times \text{hour} \quad (6)$$

The Poisson distribution of the count of the INFO event was specified in the model. Table 5 shows the results of the estimates for the model and associated significance tests based on computer logs from the class workspace (April to July).

Liao (1994) discussed five ways of interpretation generalized linear models, including the two ways used in this study. Positive estimates in the table mean that an INFO event is more likely to occur as the levels of corresponding explanatory variable increase, if other variables held constant.

Table 5

A Poisson regression model of the relationship between INFO events and contextual characters of the path activated in the same time intervals.

Parameter x_k	Estimate β_k	SE	Z-value * $p < 0.05$	Marginal Effect
Zdepth	19.2786	7.5654	2.55*	0.5938
Hour	.2060	.3023	.68	0.0063
Zword	.9397	9.7437	.10	0.0289
Zlink	-.6285	1.4958	-.42	-0.0194
Zcredit	-11.2120	4.1423	-2.71*	-0.3453
Zsize	-15.2472	2.6603	-5.73*	-0.4696
Constant	1.4017	3.1006	.45	0.0436

* Significant at $p = 0.05$

Negative estimates mean that an INFO event is less likely to occur as the levels of these variable increase. A significant test at a conventional α level of 0.05 was used to test whether the effect was statistically different from zero.

Three estimates in this model were statistically different from zero. In particular, Zdepth had a positive estimate of $\beta_{level} = 19.28$ ($p < 0.05$). This difference means that INFO events are more likely to occur as the aggregated depth increases. In other words, users are more likely to invoke the INFO function at lower levels of the hierarchical structure of a shared workspace. On the other hand, both Zcredit and Zsize had negative estimates. Our explanation is that INFO events are more likely to occur on hypertext nodes that have been rarely visited, hence with a lower Zcredit level. INFO events are more likely to happen on hypertext nodes that are relatively small in size, Zsize. Because most work-in-progress nodes tend to be small in size, it is possible that INFO events are more likely to be related to the early stage of collaborative writing.

The Poisson regression model can be interpreted in terms of the marginal effect of an explanatory variable on the expected frequency of INFO events. The marginal effect of x_k on expected y is given by $\theta\beta_k$, where $\theta = 0.0308$ is the expected occurrences of INFO events in a given time interval.

The marginal effect of Zdepth on the INFO count is (0.0308)(19.2786) = 0.5938. In other words, if the value of Zdepth increases by one, the expected INFO count will increase by 0.5938, other things being equal. In contrast, the marginal effect of Zcredit is (0.0308)(-11.2120) = –0.3453, suggesting that the expected INFO count will decrease by 0.3453 as the Zcredit increases by one standard deviation.

In fact, Zdepth was negatively correlated with the frequency of UNFOLD and was positively correlated with the frequency of LOCK (Pearson's r= –0.529, p <0.001 and r= 0.187, p = 0.001, respectively). Therefore, the Poisson model suggests that INFO events are most likely to occur in areas under construction within a workspace. Furthermore, when we included a dummy variable Phase in the model to specify the two phases of reading and authoring, the Phase variable would have the most predominant effect on INFO events. The writing phase corresponded to the increased probability of the occurrence of an INFO event. This is consistent with our expectation that users would need the INFO function more frequently if the underlying shared workspace is under construction.

4. DISCUSSION

Maintaining the group awareness among collaborators is an essential part of a successful collaboration (Dourish & Bellotti, 1992; Whittaker et al., 1993). Our analysis, based on the Poisson model, has indicated that users tend to need the group awareness more frequently in areas of a shared workspace that involve a lot of changes. The effect of word counts on INFO events was not significantly different from zero (β_{word}= 0.9397, p >0.05). On the other hand, the effects of the overall depth, the overall credit, and the total number of concurrent users were all statistically significant from zero.

The Poisson regression model has revealed that INFO events are less likely to occur on nodes with higher selection credits. Our explanation is that as users become familiar with the organization of the workspace, the need for information provided by the INFO function will be reduced. Overall, INFO events are more likely to be found in a workspace where users engage in writing as opposed to reading.

5. CONCLUSIONS

The use of quantitative methods is not to replace qualitative methods. Rather, the intention is to explore a wider spectrum of combined methods to deal with the complexity of understanding the phenomenon and encoding this understanding into practical CSCW systems. We have studied the time series of INFO events and some characteristics of the context where INFO events were likely to occur. INFO events can be seen as an indicator of how users shift their focus back and forth between individual work and more social, organizational aspects of the work.

We have presented our analysis based on a Poisson regression model, aiming to identify a number of interrelationships between awareness-seeking events and structural properties in the context of an underlying workspace. We concluded that users are more likely to need the INFO function in an evolving workspace or when a substantial amount of writing takes place in the workspace.

The significance of this study is that it has established a framework for studying the relations between the occurrences of awareness seeking action and the structure of the collaborative hypertext being used on the basis of empirical evidence. An investigation of the interrelationship between task processes and the organization of a shared workspace is generally applicable to a range of domains, from intranets within organizations to the Internet at a large scale. For designers and system engineers, statistical and probabilistic models offer the opportunity of monitoring and measuring users' behavioral patterns in the context of an evolving workspace such that one can device intelligent agents to help users in areas where they will need the information the most.

The work provides a useful basis for the development of a framework for computer-supported collaborative writing. The process-oriented perspective used in this study has led us to some insights into how users interact with a shared workspace. Future work should aim to improve the understanding of collaborative writing achieved through this investigation. For example, our investigation has focused on a subset of co-authoring support mechanisms in the MUCH system; in future studies, one may focus on a smaller and more

manageable collaborative writing practices on the World-Wide Web, as well as on intranets within organizations.

The dynamic modeling approach has the advantage of being flexible and sensitive to significant changes in existing patterns of collaborative work as the computer supporting techniques change and improve. From a user-centered design perspective, a computer system should be adaptable to user behavioral patterns in such a way that the use of the system increasingly conforms to the work process that has been adopted by a particular group of users. Moreover, users will not be constrained by particular implementation decisions made at early stages of software development. With historical and empirical data, computer systems can build a profile of a user in a collaborative context so as to provide the mechanisms needed.

The investigation shows some promising properties of an empirical and statistic approach to transcend the dynamics and complexity for a better understanding of situated actions. The knowledge obtained from this approach complements the findings from approaches, such as ethnomethodologies. Theoretical and empirical approaches must be integrated to improve our understanding of what support is needed for people in collaborative writing and what an intelligent system can offer.

REFERENCES

Box, G. and G. Jenkins. 1976. *Time series analysis, forecasting and control.* Holden-Day, San Francisco, California, USA.

Chen, C. 1997. Writing with collaborative hypertext: Analysis and modeling. *Journal of the American Society for Information Science*, **48**, 1049–1066.

Chen, C., Rada, R. and Zeb, A. 1994. An extended fisheye view browser for collaborative writing. *International Journal of Human-Computer Studies*, **40**, 859–878.

Cooke, N.J., Neville, K.J. and Rowe, A.L. 1996. Procedural network representations of sequential data. *Human–Computer Interaction*, **11**.

Delise, N. and Schwartz, M. 1987. Contexts—a partitioning concept for hypertext. *ACM Transactions on Office Information Systems*, **5**: 168–186.

STARTDourish, and V. Bellotti. 1992. Awareness and coordination in shared work-spaces. *ACM. Conference on Computer–Supported Cooperative Work (CSCW'92)*, ACM.

Edwards, M.B., et al. 1985. The role of flight progress strips in en route air traffic control: A time–series analysis. *International Journal of Human-Computer Studies*, **43**, 1–14.

Galegher, J. and Kraut, R.E. 1994. Computer–mediated communication for intellectual work – an experiment in group writing. *Information Systems Research*, **5**, 110–138.

Guster, D. and Robinson, D. 1993. The application of Box–Jenkins techniques in forecasting computer resource needs. in: *1993 Information Resources Management Association (IRMA) International Conference.* Salt Lake City, UT.

Halasz, F. and Schwartz, M. 1994. The Dexter Hypertext Reference Model. *Communications of the ACM*, **37**, 30–39.

Irish, P. and Trigg, R. 1989. Supporting collaboration in hypermedia: Issues and experiences. *Journal of the American Society of Information Science,* **40**, 192–199.

King, G. 1994. Statistical models for political science event counts: Bias in conventional procedures and evidence for the exponential Poisson regression model. *American Journal of Political Science*, **32**, 838–863.

Liao, T.F. 1994. *Interpreting probability models: Logit, probit, and other generalized linear models*, Sage, Thousand Oaks, California, USA.

Losada, M., Sanchez, P. and Noble, E. 1990. Collaborative technology and group process feedback: their impact on interactive sequences in meetings. in: *CSCW'90*, ACM Press, Los Angeles, California, USA.

Markova, I., *Human awareness: Its social development.* 1987, Hutchinson, London, UK.

McCullagh, P. and Nelder J.A. 1989. *Generalized linear models*, 2nd edition, Chapman & Hall, London, UK.

Neuwirth, C., et al. 1990. Issues in the design of computer support for co–authoring and commenting. in: *CSCW'90*, ACM Press, Los Angeles, California, USA.

Olson, G.M., Herbsleb, J.D. and Rueter, H.H. 1994. Characterizing the sequential structure of interactive behaviors through statistical and grammatical techniques. *Human–Computer Interaction*, **9**, 427–472.

Ostrom, C. 1978. *Time series analysis: Regression techniques*, Sage Publications, Beverly Hills, California, USA.

Posner, I. and Baecker, R. 1992. How people write together. in: *Twenty–Fifth Hawaii International Conference on System Sciences*, Kauai, Hawaii.

Sanderson, P.M. and Fisher, C. 1994. Exploratory sequential data analysis: Foundations. *Human-Computer Interaction*, 9, 251–317.

Smith, J., Smith, D. and Kupstas E. 1993. Automated protocol analysis. *Human-Computer Interaction*, **8**, 101–145.

Vortac, O.U., Edwards, M.B. and Manning, C.A. 1994. *Sequences of actions for individual and terms of air traffic controllers. Human-Computer Interaction*, 9. 319–344.

Whittaker, S., Geelhoed, E. and Robinson, E. 1993. Shared workspaces: How do they work and when are they useful? *International Journal of Man–Machine Studies*, **39**, 813–842.

Zheng, M. and R. Rada. 1994. MUCH electronic publishing environment: Principles and practices. *Journal of the American Society for Information Science*, **45**, 300–309.