Http Log Analysis

An Approach to Studying the Use of Web-Based Information Systems

Kristian Billeskov Bøving & Jesper Simonsen

Computer Science, Roskilde University, Denmark
kristian@billeskov.dk & simonsen@ruc.dk

Abstract. This article documents how log analysis can inform qualitative studies concerning the usage of web-based information systems (WIS). No prior research has used http log files as data to study collaboration between multiple users in organisational settings. We investigate how to perform http log analysis; what http log analysis says about the nature of collaborative WIS use; and how results from http log analysis may support other data collection methods such as surveys, interviews, and observation. The analysis of log files initially lends itself to research designs, which serve to test hypotheses using a quantitative methodology. We show that http log analysis can also be valuable in qualitative research such as case studies. The results from http log analysis can be triangulated with other data sources and for example serve as a means of supporting the interpretation of interview data. It can also be used to generate hypotheses, which were otherwise unthinkable. We suggest that log data be included as a main data source in the field of computer supported cooperative work, information systems, and computer-mediated communication, in order to help clarify the role of the technology related to concepts like coordination, task analysis, or communication.

Key words: http, log, log analysis, web-based information systems use, use patterns, collaboration, triangulation, QuickPlace.

1 Introduction

Web-based information systems (WIS) are often used in distributed organisations to support communication, collaboration, and coordination. Managers direct resources and set up goals for the deployment of WIS. It is however very difficult to foresee the effect of networked and distributed systems or to evaluate usage in general (Grudin 1994). Many cases report that such systems (also referred to as computer-supported cooperative work (CSCW) or groupware) are either hardly used or do not produce the intended effect (Bansler and Havn 2002; Bowers 1994; Bullen and Bennett 1990; Grudin 1988; Grudin 1994; Orlikowski 1993; 1996; 2000). The study of the use of WIS, which is distributed across time and space, challenges existing methods of observation. We have investigated how the analysis of http log files can contribute to meeting this challenge. Using data mining techniques, http log analysis can analyse usage data from a large number of spatially and temporally distributed users and it can support research designs which span over longer periods of time. Http logs have been used previously in web usage mining for the study of single users activity on web sites. However, no prior research has used log files as data to study the collaboration between multiple users in organisational settings.

In this article, we report from our experiences of experimenting with http log analysis (in short, log analysis) as an alternative approach to studying the use of the WIS application Lotus QuickPlace™, an application which today is marketed by IBM as Team Workplace™. The application offers users a web-based shared virtual workspace called a QuickPlace, here after referred to as QP. We describe how log analysis may be applied as a supplement to qualitative approaches and surveys by for example helping in the interpretation of interview data. We also demonstrate how log analysis may generate hypotheses which were otherwise unthinkable. The article presents empirical findings from a large distributed financial organisation where the usage of more than 100 QPs, comprising in total about 3000 active users and more than 20 Gb of documents, was studied over a 10 months period. We conducted our experiments with log analysis along side of other research approaches in the same organisation. These included a survey as well as a number of interviews and observations of use.

We divide our research of studying the use of WIS into three general and principally different approaches:

1. Qualitative approaches such as observations, interviews, and document studies. Such approaches usually study one or a few WIS and include

relatively few users. The result is a large and qualitatively "rich" amount of data that focus on 'why-questions' and cause explanations, for example a user's intention behind using a WIS in a particular way.

2. Surveys representing a combined quantitative and qualitative approach (by using closed vs. open questions). Surveys might reach a high number of users and WIS. The results are both quantitative data representing WIS use at a specific point in time (when answering the survey) and smaller amounts of qualitative data from the surveys' open-ended questions.
3. Http log analysis representing a quantitative approach based on querying and datamining logs from http transactions between web client and server. The result is quantified information from the log tracing actual and indisputable use. In contrast to qualitative approaches and surveys, log analysis cover *all* WIS and *all* users over a *greater and continual time interval*.

Given that a study has to cover a large number of WIS and many users, log analysis is the only approach that gives acces to a data source that includes every http transaction that actually happened over a longer period of time. The potential of log analyses is therefore obvious in respect to monitoring actual use of WIS and subsequently to inform studies interpreting and evaluating the use of WIS.

Http logs are logs of the transactions between a web-based client (the browser) and its server. All http servers, and thus all WIS, have the built-in possibility of maintaining a http log. However, log files are not problem-free when used for analysing WIS use. Log files are structured with a completely different purpose in mind than such analyses. Log files also show only very few aspects and traces of use activities. Two identical lines in a log file might therefore document use processes that would differ significantly in a direct observation of WIS use. Due to the nature of information in a http log, a number of precautions have to be taken when interpreting the log analysis data and when generalising results from settings.

These problems can be solved in very controlled settings, such as in experimental research were the WIS is custom designed so that it logs the information needed for the research. For real life studies of WIS use however, like our study of Lotus QuickPlace, using log files as a data source is not common. We have not been able to identify any case study or other empirical real-life studies using log files as a data source. Therefore the use of log files in this article must be considered explorative. The empirical basis of our study is a number of experiments with different approaches to log analyses. Our overall research question is: *How can http log analysis inform studies of WIS use?* More spe-

cifically we investigate how to perform log analysis; what log analysis says about the nature of WIS use; and how results from log analysis support other data sources.

The organisation from where we have our log files and other empirical data is a leading financial corporation in Europe, which we in the following refer to as Beta. Beta is a result of a recent merger involving several companies located in a number of countries. The merger produced an instant need for a platform independent tool to support secure communication in geographically distributed settings because Beta immediately after the merger had no secure corporate infrastructures (e.g., a LAN, or an intranet) to distribute confidential information. Lotus QuickPlace was chosen as the standard application to support communication within geographically distributed corporate organisational units, groups, projects, and teams. Lotus QuickPlace requires no integration with the existing IT infrastructure and offers a secure web-based workspace.

Lotus QuickPlace offers a workspace with facilities for sharing and co-authoring documents, exchanging files, and supporting discussions, calendar, email-notifications, etc. A QP is structured with folders that contain documents, web pages, files, etc. QP is a generic system (as defined by Bansler and Havn (1994)), which means that it needs to be configured and customised to the specific needs of the group of users. The person(s) with manager rights to a newly installed QP must start by designing an initial structure, setting up a home page, and creating and naming document folders. They also have to invite users and grant them access rights either as manager, author, or reader. When a QP is used, new needs arise for changing the initial setup, configuration, structure of information, as well as for agreeing on how to use QP.

In the following we present an overview of related research within web-usage mining (section 2). Then in section 3 we describe log file analysis in more detail, and in section 4 we explain how to prepare a http log in order to be able to perform analysis of WIS use. In section 5, we continue by giving four examples from our experiments with different approaches to log analyses. Finally we discuss how such analyses can inform studies of WIS use (section 6).

2 Related Research

No prior studies of using WIS in intraorganisational settings have included http log analysis. Related research can however be found in the extensive body

of literature within data mining techniques to the World Wide Web, referred to as web mining.

The term web mining has been used in two distinct ways (Cooley et al. 1997). The first, called web content mining is the process of automatic search of information resources across the World Wide Web, e.g., by database or agent-based approaches. Web content mining is outside the scope of this article. The second research strand, called web usage mining, concerns the discovery of user browsing and access patterns from web servers.

The overview of related research given below shows that web usage mining has been strongly driven by the commercial interest in utilising WWW. The applications of the methods are also all focused on the interaction between a single user and a web site. It could therefore be characterised partly as a contribution to research in human-computer interaction (HCI) and partly to marketing and sales research.

Existing research on web usage mining, as well as their practical applications, is focused on analysing how single users access a web site. A typical approach is to analyse actual uses of a web site in terms of, for example, how people navigate through the web site and use this information as input for redesigning the web site. A simple example would be to discover that some specific page on level 3 of the web site hierarchy is the target of 50% of all visits to the web site. This could suggest that the page should take a more central role in the information hierarchy. Similarly, the fact that a page considered very important by the site owners but rarely used, suggests that it be moved to another place in the information hierarchy. See (Masseglia et al. 1999) and (Spiliopoulou 2000) for examples.

Web usage mining has a number of scientific as well as practical applications. The following presents a categorisation of applications of web usage mining both in research literature and in practice.

2.1 Technical Analysis

The technical analysis of http logs is used to generate simple statistics of usage, which are used, for example, for load planning and sizing of the technology, or for measuring activity levels in different areas of a web site. One of the difficult aspects of sizing web technologies is to foresee the level of activity and how it fluctuates in different time periods. Analyses of log files, which produce activity level histograms for a site, can help the planning for future investments in technology or define time periods where the site can be maintained without disturbing too many users. These analyses are supported by all commercial web site analysis tools (e.g., WebTrends

(www.webtrends.com)) of which some exist as freeware (e.g., Analog (www.analog.cx)).

2.2 Methods for Extracting Use Patterns

Different generic methods and algorithms have been developed to extract typical patterns of usage of a web site. These are mostly based on different types of sequence analysis and association rule mining adopted from the data mining discipline. For examples see (Andersen et al. 2000; Cooley et al. 1999; Hidber 1998; Pei et al. 2000; Srikant and Agrawal 1995; 1996). Sequence analysis analyses typical sequences of events in time, while association rule mining analyses associations between web pages based on users' actions. Association rule mining is also known from shopping basket analysis, providing results such as "80% of people who buy beer also buy potato chips." In the context of usage of a web site a possible result could be that "10% of people who visit the product descriptions section also visit the web-shop."

Research on methods for clustering users based on their usage of a web site has also been conducted, see e.g., (Fu et al. 1999). A practical application of user clustering could be to reflect the types of users in the way information is structured for users in the information architecture.

2.3 Personalisation and Information Architecture

Research has been conducted into the utilisation of the generic methods mentioned above for improving the user interface of a web site. The most popular approach has been to utilise the analysis of usage for providing personalisation of the user interface (Mobasher et al. 2000; 2001; Su et al. 2002; Toolan and Kushmerick 2002). The idea is that the history of usage for a single user or group of users can provide input on how to promote specific web pages on a site to specific users. A simple implementation could be to generate a list of news articles, based on news articles previously read by a user. Such an application would also require a typology for news which could be mapped to the usage. The generation of the news list would involve selections such as "if the user has previously read news of type X, and we have current news of type X, then add the current news of type X to the list of news displayed to the user."

Apart from personalisation, more generic approaches to dynamic information architectures have also been developed (Masseglia et al. 1999; Spiliopoulou et al. 1999; Su et al. 2002). They generally incorporate usage history in the automated design of the interaction with users.

2.4 On-Line Shopping Behaviour

Specific methods have been developed for studying the behaviour of on-line shoppers. Some research has been developing algorithms for applying sequence analysis to the analysis of shopping behaviour (Srikant and Agrawal 1995; 1996). Different forms of sequence analysis for analysing on-line shopping behaviour are now built into commercial data mining products and services (e.g., Clementine from SPSS, WebHound from SAS, and SurfAid from IBM). Sequence analysis is, e.g., used for optimizing design of shopping flows on a web shops.

2.5 Marketing Based on Log Analysis

Clustering of users based on usage and correlated analyses of other data such as customer segmentation models (e.g., Minerva or Kompass) or buying history have been researched as an input for marketing. The goal is to direct marketing more precisely towards potential customers. One of the buzzwords is personalised marketing (Büchner and Mulvenna 1998). While not based on the analysis of http log files but rather transactional data, the marketing e-mails sent out by the Amazon bookstores illustrate this principle.

2.6 Web Usage Mining vs. Log Analysis of WIS Use

Many web sites can be understood as collections of information or shops available for single users. The existing research in web usage mining is based on this perspective. It ignores however the fact that web technologies such as WIS function as a media for communication between users as studied within computer-mediated communication (CMC). Http logs can also be used as a source for understanding how people collaborate supported by WIS as studied within CSCW. But this requires a different analytical approach and increases the importance of combining the quantitative traces of action with actor's accounts of these actions derived from interviews or questionnaires.

The analytical unit for web usage mining is the session. A session is a sequence or collection of interactions with the system by one user limited by function and by time. In http log analysis, a session can be defined as interactions of one user preceded by 30 minutes of non-activity and followed by 30 minutes of non-activity.

In the analysis of communication between users mediated by a WIS, the session is not a relevant analytical unit. Instead, the analytical unit is typically the document (or the html page). One might intuitively think that a concept of

communication would be the analytical unit. Due to the architecture of the http-protocol specifically and WWW in general, this is not possible. The http protocol, and therefore also http logs, only record single users interaction with specific resources on a web server, e.g., a document or a html page at a specific point in time.

3 Log Files as Empirical Data

Most IT systems use log files. The purpose of log files is generally to trace the history of an application. The concept of logging is well known, for example, from maintaining a ship's log. Logging the activity of an IT system generally means to write a trace of what the system has done to a file as a part of the execution. Logging is in application design known as tracing and it is known as transactional logging in transaction monitors and database systems. In application design the purpose of the tracing is primarily to analyse the program execution in order to optimise algorithms. In transactional logging of databases (e.g., DB2, Oracle, or MySQL) and transaction monitors (e.g., CICS from IBM or Tuxedo from BEA) the purpose is to maintain consistency and to be able to recover from system breakdowns without loss of information, for example in settings where IT systems are used to automate transactions such as maintaining bank accounts.

Log files can be evaluated as empirical data related to case studies, archival records, and direct observations (Yin 1994). The qualities and deficiencies of the http log files as data for a case study might be summarised as follows:

- Indices: They provide perfect information about a very limited aspect of usage. Log lines are indices of use.
- Consistent: They provide this information consistently across time.
- Accessible: The data is relatively easy to access using multiple querying and data mining techniques.

http log files can be characterised as a special type of direct observations, which combine certain characteristics of archival records and direct observations.

Archival records are characterised by Yin (1994, p. 80) as:

- Stable—can be retrieved repeatedly.
- Unobtrusive—not created as a result of the case study.
- Exact—contains exact names and details of an event.
- Broad coverage—long span of time, many events and many settings.
- Precise and quantitative.

All of these properties are also properties of log files except for 'exact': The log file data does not link to names and details that are part of the social discourse in the organisation. This information can only be established by combining the log file data with other data sources e.g., from interviews and observations.

Log files differ in a very important way from archival records in that they are not produced intentionally by members of the organisation studied. The information present in the log files is a combined product of both the de-facto standard http log format and of the technical design of the WIS technology.

Yin (1994) characterises direct observations as:

- Reality—covers events in real time.
- Contextual—covers context of the event.

Clearly log files only capture a very limited aspect of events and it does not capture what Yin refers the context of the event. Instead log files have another important characteristic: The data consist of structured computer records. This means that log files are directly available for computational analysis – which is also the only way to handle such huge amounts of data.

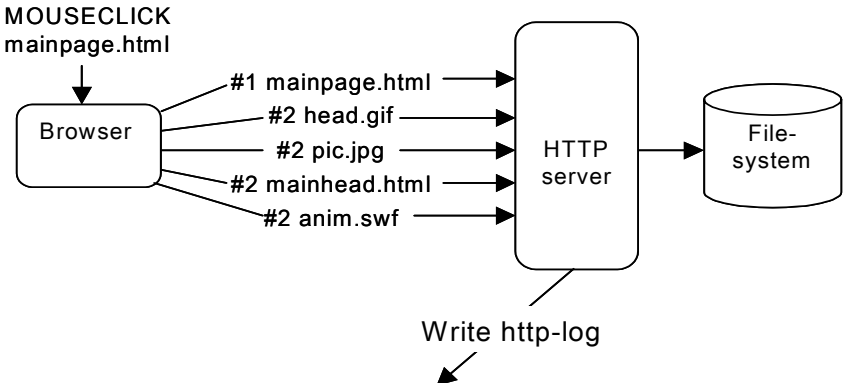
4 Preparing a Log for Analysis

A http log made on the server contains log lines with data describing each transaction between client (browser) and server stored in chronological order in a text file. Preparing the original http log in order to make analysis of WIS use involves a process comprising a number of steps:

1. Defining the goal of the analysis, and therefore also the user actions relevant for the analysis. This is an iterative process of defining goals and investigating the possibilities available with the server architecture, which produces the log file. One of our goals was to analyse to which extent users collaborated on writing documents. Needed for this analysis is a log where you can identify documents and users as well as an action of type "edit an existing document" (see table 1 below).
2. Cleansing and preparation of the data. Most of the data in a raw log file is irrelevant (see figure 2 below), and it is preferred to extract data from the URL and store them as separate data. We extracted the relevant data from the log file to a database using Perl scripts and standard SQL.

3. Breaking the code of the relationship between user actions and properties of the URL in a log line. In our case this included making a series of controlled test including different user actions (e.g., download, edit, move, etc.) to see what was produced in the log, and how user actions could be identified.
4. Generation of data matrixes based on the cleansed data. Most analyses of the log data are made on matrixes containing aggregated information. A matrix of aggregated information could be storing a document ID and the number of times it was accessed by a user instead of storing each user access with all of its properties such as time, user, etc. In our case this was done in the form of tables in a relational database.
5. Analysis and visualisation of the data matrices. In our case this was achieved using a number of tools: Clementine from SPSS, SQL, Excel spreadsheets, and SPSS statistical package.

One complication in analysing the log file is the problem that one user action (e.g., using the mouse to press a button on the screen) typically causes a number of lines to be written to the log. This is due to the architecture of html and http. When a user request a plain html document through a browser from a http-server using a URL, the http-server will serve the document. However, html documents usually contain elements, which are placed on the http-server as individual resources (e.g., images, java-applets, flash movies, etc.). The browser then analyses the html document and requests the elements, which were linked from the html document in the same way it requested the original document. Figure 1 illustrates a simple example where a user enters the URL address for a html page (#1 in figure 1) resulting in automatic subsequent request for four additional resources (#2 in figure 1). The mouse click, or typing of the URL in the browser address field and pressing return, eventually produces 5 loglines in the server http log. The problem is then to locate the resource relevant for the analysis. Typically, the relevant resource to analyse is the one that uniquely identifies that the user is in fact looking at the contents of a specific document.




```
123.90.183.23 Host Username Time "GET /mainpage.html" HTTP-return Bytes
123.90.183.23 Host Username Time "GET /head.gif" HTTP-return Bytes
123.90.183.23 Host Username Time "GET /pic.jpg" HTTP-return Bytes
123.90.183.23 Host Username Time "GET /mainhead.html" HTTP-return Bytes
123.90.183.23 Host Username Time "GET /anim.swf" HTTP-return Bytes
```

Figure 1. The resulting http log (5 loglines) after executing a request for a html page in the address field of a browser.

A log file only contain data from the http transactions and does not contain information on the content of a document. It is however possible to reveal data related to a user action by analysing the properties of the URL in a log line.

In our case we did a controlled test on a test server, where we performed and documented on paper a series of user actions in a QP. We then afterwards analyzed the lines written to the log as a result of our actions. This test was used as the basis for uniquely relating properties of the URL (e.g., the presence of the string "?Download") to a user action (e.g., that the user has downloaded a specific document).

```
192.168.73.200 RUCQuickPlaceServer CN=frank/OU=alpha-QP
[05/May/2001:14:58:42 -0100] "GET /QuickPlace/alpha-
QP/Main.nsf/$defaultview/505538f1eb2b9dbf0525670800167214?"
?DownloadHTTP/1.1"200 4789
```

Figure 2. Example of a single log line from the http log from the Lotus QuickPlace server. The presence of the string "?Download" shows that this URL requested an action of type Download (see table 1 below)

From our analysis of the Lotus QuickPlace log files we have been able to identify the data in the log lines as shown in table 1. The data is typical for http based client server architectures and most of this data would be possible to identify for any WIS (also WIS based on other platforms than Lotus QuickPlace).

The identified data in the log files were extracted and stored in tables in a relational database. Further analysis were made by means of multiple SQL queries and datamining techniques.

Action types identified in URLs	*Attributes logged for each action type*
Read (html page)	Server ID
Upload	QP ID
Download	Login (user) ID
Edit	Document ID
Move	IP address
Delete	Time stamp
Search	http return code

Table 1. Data identified in the Lotus QuickPlace log file shown as different action types and the attributes related to an action type. Each log line that contain an action (e.g., Download) also contain all the attributes related to that action. Thus a log line as the one shown in figure 2 reveals that user 'frank' downloads a Document with the encrypted id '505538f1eb2b9dbf0525670800167214' from QP 'alpha-QP' located on Server 'RUCQuickPlaceServer' at time '05/May/2001:14:58:42 –0100' using a PC that has IP address '192.168.73.200' and the result of the download gave http return code '200 4789' (successful download of 4789 bytes).

5 Examples of Log Analysis

In the following we present four examples of using the analysis of log-files for studying the use of WIS, and we evaluate how each example can contribute to the understanding of the role of a technology in a work context. The examples do not attempt to cover the whole spectrum of uses, but should serve as an inspiration illustrating how log analysis can be used. The first example includes some basic statistics available through log analysis. In the second example such statistics are demonstrated covering longer time spans to investigate the lifecycles of QP's. The third example describes the potential of global models investigating lifecycles on a document level. The fourth and final example gives an account of a more detailed analysis of specific QP use related to a collaborative process of translating documents.

5.1 Simple Statistics on WIS Use

Simple statistics on the use of WIS might be used to provide background data for various analyses, as well as for testing very specific hypotheses.

We analysed the distribution of the number of users in the QPs in Beta. As the histograms in figure 3 show, the number of users vary from 1 to more than 248. The histogram shows the distribution of frequency of number of users per QP.

The histograms show a large diversity of the number of users. This suggests quite diverse uses of the technology. One may assume, that the QP with more than 248 users is used in different ways for different purposes (as a local intranet) than the ones with, e.g., 15 users. Such an assumption can then be investigated by means of further more detailed log analysis or supplemental data collection methods like surveys and interviews.

The distribution of the number of users might also be used to test very specific hypotheses. One such hypothesis for the number of users in the QPs is that there would be an "ideal" number of users, which most of the QPs would have, and that few QPs would have much fewer or many more users than the average. This could be formulated in statistical terms as a hypothesis stating that the number of users pr. QP is distributed according to the normal distribu-

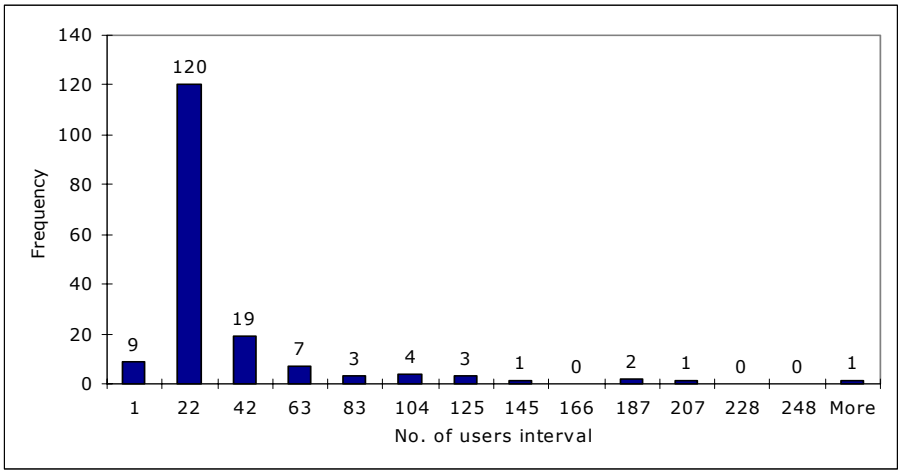


Figure 3. Histogram of users per QP counting the number of unique active users in each QP during the whole period of logging. The histogram should be read so that, for example, 19 QPs have between 23 and 42 active users and that 120 QPs have between 2 and 22 active users.

tion. As it turned out, this hypothesis was refuted by the data. Instead we found that the number of users pr. QP were distributed according to a power law distribution with a calculated Pearson correlation of 0.971. See e.g., (Huberman 2001) for an overview of the analysis of power law distributions on e.g., web site visits.

Simple statistics on usage might also be combined with survey data to give examples of how QPs are used. By extracting the number of users and the number of documents in selected QPs and combine these with answers from a survey conducted in Beta, we were able to provide the following short characteristics of use in two different QPs.

> A QP called 'Alpha-QP' is a QP, which exemplifies one of the QPs with many users. Within 10 months, 203 users had been active using 5297 different documents. The QP supports an organisational unit called "Alpha", which is an organisational unit spanning four Nordic countries as well as all other countries where Beta is present. According to our survey conducted within Beta about QP use, the QP is used to support credit projects, distribute credit limits and related information on issuing credits to large customers, for marketing materials, and for meeting agendas and minutes, as well as distributing internal information, e.g., holiday lists. (Bøving 2003, p. 109)

> The QP 'Gamma-QP' had 35 active users who within 10 months were working on 1896 documents. The QP was used to collect daily risk reports in the form of spreadsheets from different parts of the Beta and to consolidate them into one spreadsheet, providing a daily snapshot of the overall risk situation. In contrast to 'Alpha-QP', this QP was used to support one very specific task. (Bøving 2003, p. 109)

Simple statistics might be used for a number of purposes in the research process:

- Selecting QPs for detailed study either by selecting a typical QP or by selecting extremes to cover the range of usage.
- Test specific hypotheses, e.g., on the distribution of QP sizes.
- Help the interpretation of survey results by providing simple facts like number of users or number of documents.

5.2 QP Life-Cycles

The following analysis illustrates the potential of log analysis for providing analyses that span a longer period of time.

In order to look closely into how new QPs evolve, we made an analysis of QPs, which have been started during the log period.

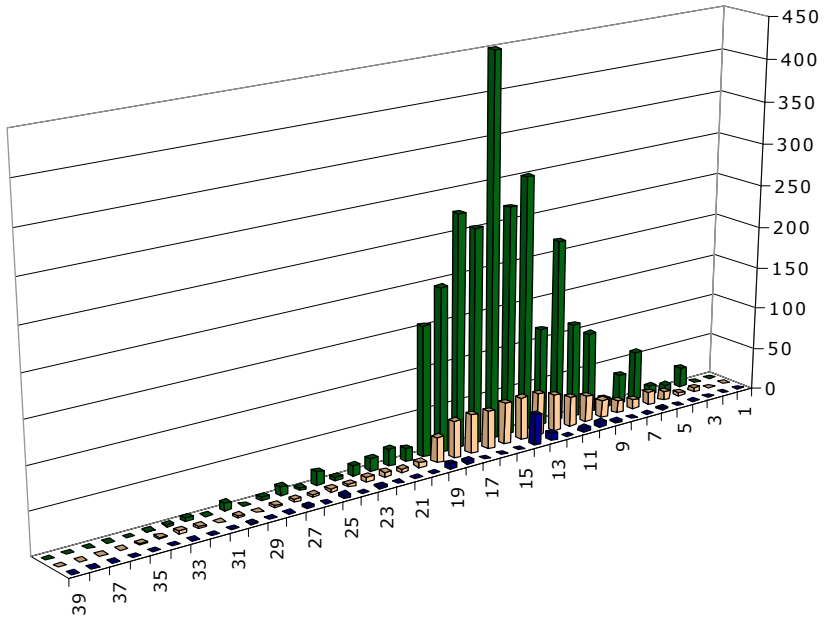


Figure 4. QP used to support a project completed within the log period. Y-axis shows number of reads (high column in back), number of users (middle column), and number of edits (low column in front). X-axis shows the weeks of use (a period of 39 weeks). The QP shows initial use after week 1 and no use is logged after week 35.

Our log represented in total 170 different QPs. Some of these were active before we started logging, some died out during the period of logging and some started up during the log period. For some we had the full life-cycle represented in our log (from a QP was initially used until no one used it any more).

In order to analyse the life-cycle of a QP, we identified and selected 37 QPs which were all started in the period of logging. On a per-week basis, the number of active users, number of document reads and the number of document edits was plotted for each of the 37 QPs. These graphs give an overall idea of the historical development of activity or, indeed, whether there was activity at all.

The sample graph in figure 4 shows the historical development of activity for a QP, which had been used to support a project. The project, we learned

from the survey, had the purpose of implementing a new corporate name and logo throughout the organisation.

One should be cautious about making conclusions from these graphs. It is impossible, simply from a plot of the number of users and the activity, to deduce anything about what a QP was actually used for. One observation we made was that 38% (14 out of 37) of the QPs never really got started. They showed little activity for only a few weeks, or showed very fragmented activity over a longer period. The graphs tell us nothing about the reason for this, but we may conclude that quite a few new QPs never reach an activity level which indicates that they actually succeed in getting used. The historical graphs can therefore be used as a starting point for investigating into the reasons for the large percentage of such failed attempts to use QP. Potential reasons and hypotheses to be investigated could be related to models of technology adoptions (Gallivan 2001; Rogers 2003) or experiences from other studies, for example Winograd's studies of the Communicator system (Winograd 1987), Orlikowski's studies of the Notes system (Orlikowski 1992), or Parnas and Clements' studies of the process of documentation (Parnas and Clements 1986).

The graphs might also be used as a means for interpreting more detailed observations (as described in section 5.4), by providing data on, whether the activity observed in detail is extraordinary or typical or whether it is in the beginning or end of the QP lifecycle.

5.3 Statistical Generalisations of Document Life-Cycles

This section and the following illustrate work on the use of http log analysis for the study of communication or collaboration. This analysis might be used in conjunction with theories of CMC like e.g., genres of communication (Orlikowski and Yates 1994; Yates and Orlikowski 1992; Yates et al. 1997; see Bøving (2003) for examples from Beta) or theories on CSCW like, e.g., articulation work and coordination mechanisms (Schmidt et al. 1992; Schmidt et al. 1996; see Pors and Simonsen (2003) for examples from Beta).

The analysis developed in the course of our case study was centred on documents. It analyses the life cycle of a document from its creation until it is no longer in use (e.g., if it is deleted or not read during a specified period of time).

We have experimented with attempts to create global models of the document life cycles. The strength of a global model is that it explains all the data gathered in a simple way. Our attempts to produce global models for the data have taken two approaches. Firstly, we used a bottom-up approach using the

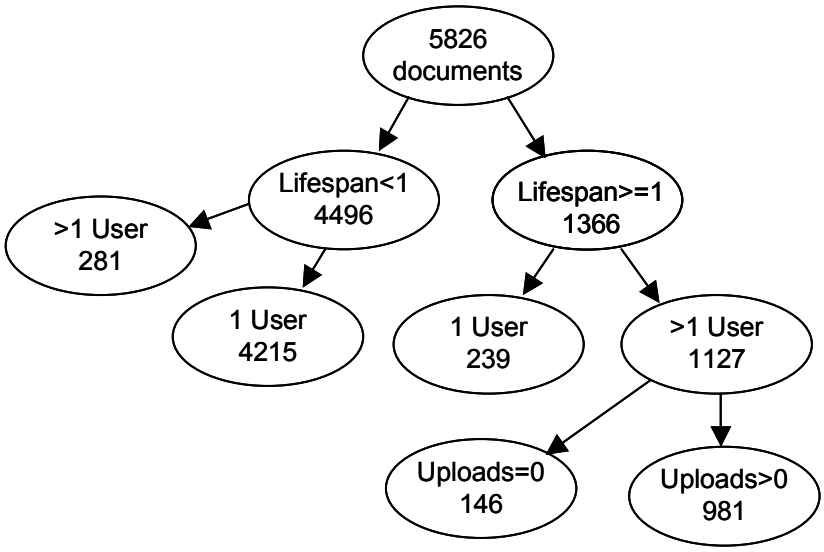


Figure 5. A typology of document life-cycles.

K-means clustering algorithm, see, e.g., (Hand et al. (2001), which produced 6 clusters of document life cycles (for a full account of the process and results see (Bøving 2003)). Secondly, we used a top-down approach. We present two possible top-down typologies of document lifecycles (see figures 5 and 6). They are hierarchical typologies made from a stepwise division of the documents according to a number of criteria. The typologies were produced from a matrix, which aggregated a number of properties, which characterises the life-cycle of each document in the sample. The properties were lifespan, number of reads, number of edits, number of unique readers, etc.

The first typology (figure 5) divides a sample of 5826 documents first according to lifespan, then number of users, and finally according to the presence of attached files.

For each type of documents, the average value of some of the properties was calculated as a way of characterising the documents of each type. These averages are exhibited in Table 2.

Document type	*% of document sample*	*Average values for properties*
Dead documents	72,3	Uploads = 2.58, Downloads=0.24, Edits=0.12
Short term coordination	4,8	Users=2.48, Uploads = 1.50, Reads=4.45, Downloads=1.67, Edits=0.30
Personal archive	4,1	Lifespan= 12.95, Uploads=2.64, Reads=4.80, Downloads=2.12, Edits=0.90
Publish document	2,5	Lifespan= 29.41, Users=3.86, Reads=10.44, Edits=0.99
Publish files	16,8	Lifespan= 25.83, Users=5.58, Uploads=2.69, Reads=13.88, Downloads=8.17, Edits=0.81

Table 2. Characteristics of document types.

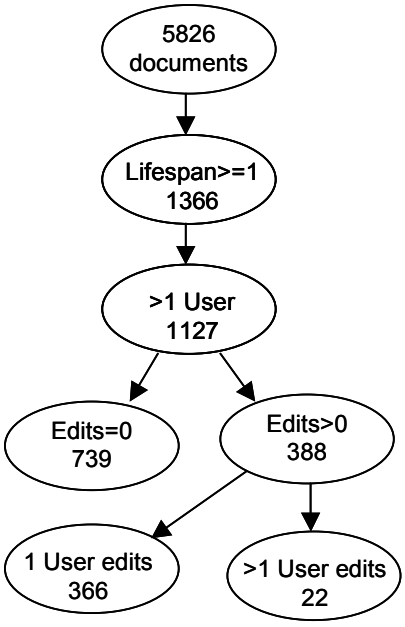


Figure 6. Typology of 'living' documents

This kind of typology provides a global model, which captures some basic properties of the use of the QPs in Beta for communication or collaboration. It provides for example one interesting observation: 73% of the documents in the sample may be characterised as a kind of dead documents: We define a dead document as a document where only one user (the one that initially uploaded it) ever accessed it. Clearly, this observation would not have been possible by other means than log analysis.

The second typology (figure 6) serves to clarify how the documents are divided according to how they are edited. It only includes documents with a lifespan of at least one day and more than one user, and provides thus a more fine-grained view of the documents that (supported by QP) are actually used for communication or collaboration.

The typology in figure 6 shows that most documents are not edited at all. While representing 12,7% of the total document sample they represent 66% of the documents with a lifespan of at least one day and accessed by more than one user.

Document characteristics	*% of document sample*	*Average values for selected properties*
No edits	12,7	Lifespan= 24.60, Users=5.10, Uploads=2.73, Reads=9.96, Downloads=7.67
Edited by one user	6,3	Lifespan= 29.31, Users=5.77, Uploads=2.22, Reads=19.58, Downloads=13.69, Edits=2.33
Edited by several users	0,3	Lifespan= 32.95, Users=6.95, Uploads=8.26, Reads=27.77, Downloads=16.95, Edits=4.05

Table 3. Characteristics of document types.

A striking observation of this analysis is the low number of documents edited by several users. One of the features of the QP technology is that it supports collaboration on documents, where several users can edit the same document without the risk of overwriting each other's changes. It can be used to co-author documents without e-mailing the document back and forth. In the log period this happens only in 22 occasions or 1,9% of the documents with a life cycle of at least one day and more than one user. It may be concluded that this feature of QP is only very rarely used in Beta. The log data does not provide a specific explanation for this, but one explanation can be ruled out. The

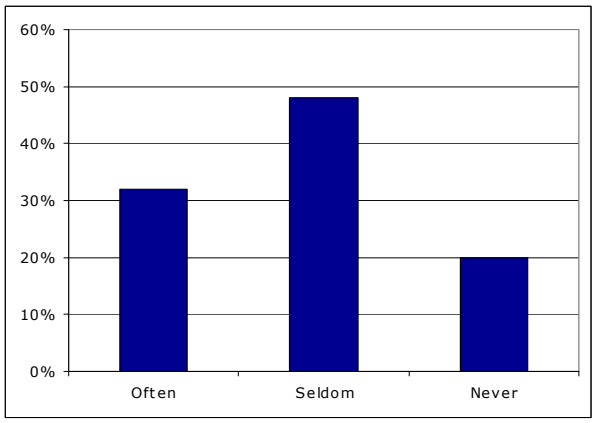


Figure 7. Responds from the survey-question: Do you write collaboratively with other people?

respondents in the survey were asked whether they wrote collaboratively with other people, see figure 7.

The reason is not that people do not write collaboratively (more than 30% do this often), they just do it without the support that QP offers for this purpose.

The two global models of the document life-cycles illustrate how log analysis can provide a perspective on how WIS are used to support communication and collaboration.

The analysis also gives an example on how the findings from the log analysis can be data-triangulated with survey data to provide a more thorough description of how WIS is used.

5.4 Detailed Analysis of a Specific QP Use

One drawback with the global models of document lifecycles is that they ignore the context of the use of QP. The document types described in the global models aggregates communication made in different work settings for different purposes.

The example provided in this section shows how log analysis can be triangulated with descriptions captured by interviews to provide a detailed description of one incident of use. In data-mining terms the analysis of the single instantiation of use is denoted a search for local patterns in the data (Hand et al. 2001).

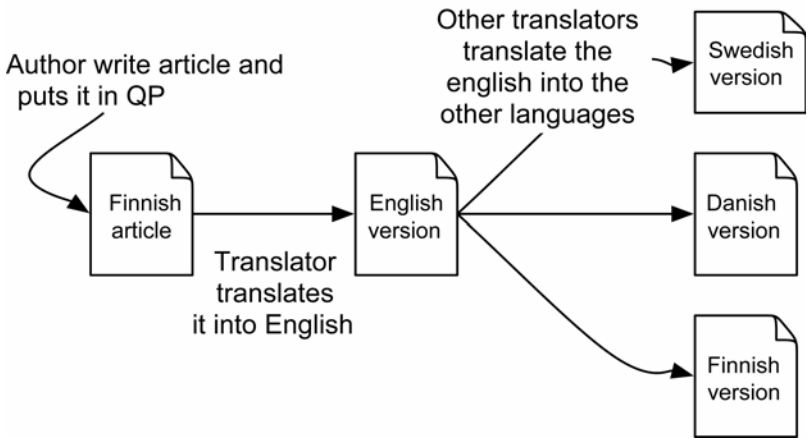


Figure 8. The translation process

The purpose of the analysis is to give a detailed account of how the QP is used to support a specific task in order to provide a more precise understanding of the role of the technology in a specific setting

One of the QPs in the Beta was used to support the translation of corporate annual and interim reports, press releases and a magazine for employees. A description of the translation of the internal magazine can be summarised in the following diagram of the process as depicted in figure 8:

1. The authors of articles for the internal magazine upload their articles in the QP in the folder named after the language in which it is written.
2. The translators download the articles and translate them into English (typically one translator per source language). Once they have completed the draft translation, they e-mail it back to the author and when he/she has accepted the draft it is sent to a proof-reader. The translator then places the English master version in a QP folder named "English".
3. A deadline is agreed on when all the articles should be available in English. Then the translators translate the English master versions into the other languages. When they have completed the translations they upload them to the QP in the respective language folders.

If an article is changed after the translation process has started, the author must place the new version in QP and notify the involved translators by the QP notification function, e-mail, or by telephone. By the deadline the result

should be that all the articles are available collected in the respective folders by language.

The above description of the translation process is made based on the interview data. The interviews alone leads to an interpretation that QP is the primary (and, except for the authors accept, also the only) tool used to support the coordination in handling different versions of documents in the translation process.

In the following, we present an analysis of the translation of a press release during August 2001 based solely on data from the log file analysis. This allows us to focus on the specific role of the WIS in the translation process and on the potential for triangulating the interpretations made from the log data with the interview data.

The first indication of the translation process in the log file is that Peter on 15/8 12:00 uploaded a document which we shall name "Danish". This file was named 16aug-dk.doc. The "16" probably refers to the deadline of the translation process.

The log analysis shows that the "Danish" document was downloaded by five other users later that afternoon. A sixth user downloaded the "Danish" document on the following day (16th August) at 9:00. Hereafter the document was not touched again in the log period, and neither was it edited or deleted, as this would otherwise have shown in the log. This means that no new versions of the attached "Danish" document were uploaded in the document lifecycle.

Peter who uploaded the "Danish" document also uploaded the Norwegian and the English versions on the same day. On the next day (16th August) at 11:00 he created a new folder (which we name "collection-1") and uploaded all language versions of the two press releases to that folder, which amounts to ten attached documents. By the time he did this, all language versions were available as documents in the QP. Mary had uploaded the Swedish version and Karen had uploaded the Finnish version during the afternoon of 15th August.

If we assume that QP was used as the coordination tool for the translation process, we would predict that Peter had downloaded the Swedish and Finnish version from the QP in order to be able to collect them in "collection-1." However, this was not the case. Some other medium must have been used to send the documents to him, and probably the e-mail system was used. The translators of the Swedish and Finnish versions probably sent the documents to Peter by e-mail instead of uploading them directly in the QP. The reason might be that the two translators stuck to the way the translation process had worked before introducing the QP.

If we turn to the other people using the "Danish" document, all of them either simply looked at the document or downloaded the file, and all these

actions happened in the period from the creation of the "Danish" document to the creation of the "collection-1" folder.

In addition to the observation made on Peter's actions, another observation points to an interpretation that the QP was not the primary (nor only) medium supporting coordination (as interpreted from the interviews). Approximately five hours after Peter had created the "collection-1" folder, Mary, the manager of the translation unit created another folder ("collection-2") and uploaded all ten documents. Mary had herself uploaded the Swedish versions and downloaded the Finnish versions. In order for her to upload the ten documents, she must therefore have had some of them sent, probably by e-mail.

From the first analysis of the detailed process we can make two interesting observations:

1. While Mary described QP as the medium used to coordinate the translation process, the log file analysis tells us that at least e-mail was used also. Actually it seems that e-mail was used as the primary way of routing the documents from the translators to the people responsible for publishing them.
2. Peter and Mary seems to have acted as proxies (routing documents on behalf of others) for some translators, while other translators seemed to have uploaded documents themselves.

Six people accessed the "Danish" document during its short lifecycle. All these accesses took place before all the five language versions (ten documents) were collected in "collection-1." Karen and Chris accessed the "Danish" document with the action type 'read' which means that they were checking to see whether the document was there or not (an action of type 'download' is needed to open the document and read the contents). Karen uploaded the two Finnish documents and immediately after that she read the "Danish" document as if she was checking that the Danish version was also there. This indicates that the QP was used by Karen as a way of checking the status of the translation process. The remaining four user's access of the "Danish" document were of action type download, which indicate, that they have actually read the contents of the attached word document.

The six people who read the "Danish" document must all have been notified that it had been published probably by means of the notification function in QP. The log analysis does not provide any information about their reason for reading the document, but it is likely that at least some of them read them for proof or approval. If they had provided input for the Danish press releases to the translator Peter, they must either have e-mailed the comments, phoned him or provided the input face-to-face (all Danish members allocated to the

translation process were located in one single open office space). A number of people read the “collection-1” and after a break in activity in the QP of three hours, Mary created a new folder “collection-2” and uploaded all ten documents. Our interpretation of this is that the readers of “collection-1” had provided input for the different language versions and the updated versions had been routed via e-mail to Mary. The finding, as in the analysis of the actions of Peter and Mary, is that QP was not used as the only tool to support the coordination of the translation process. The use of QP was mixed with the use of e-mail, phone calls and face-to-face conversations.

What is then the role of QP in this translation process?

Firstly, the primary routing of documents from the translators was only partially done using QP. E-mail must have been a central part of it. The log analysis show that for all published documents in the translation process, a number of people downloaded (and probably read) the attached documents. Either they did it only to inform themselves of the contents of the press release, or they needed to provide corrections or to provide approvals for the press releases. But in the cases where the readers fed comments back into the translation process, they did it by other means than the QP.

Secondly, the use of QP was limited to actions of upload, download, and read. No documents in the QP were edited after the initial upload. QP provides facilities for locking documents in and out and thus supporting that documents can be edited several times by different people without the risk that versions get out of synchronisation. These specific functions of QP supporting coordination of document editing were not used. The exchanges back-and-forth of new versions between the translators and readers must have been accomplished using e-mail.

The interpretation of how QP is used based on both interviews and log analysis illustrates, how log analysis can provide detailed input for an investigation into the use of a computer medium in a specific work setting. It also exemplifies how log analysis can be used as data for a qualitative study of a specific work practice.

In this case log analysis has provided an additional perspective, which can be triangulated with interviews, to provide a richer understanding of how the medium is actually used for specific communications.

6 Discussion

The primary quality of the log file is its rigorous and accurate detail. Like other quantitative data it captures only few and limited aspects of WIS use but

in a systematic manner that allows analyses spanning a large number of users and WIS in long continual periods of time.

In the analysis of QP use for the translation process, triangulation by means of comparing interpretations of use, which were provided through interviews and accounts of use through log analysis, improved the researchers' interpretation of the interview statements. Log files might serve as a way of challenging a description made on the basis of interview data. The log analysis changed quite radically the interpretation of the interviews. The interviews give the impression that QP is used as the primary tool to coordinate the translation process. The result of the log-analysis gives us a rather different picture of this pointing to the fact (among others) that e-mail also must have been an important medium for coordinating the translation.

Log analysis turned out to be useful for another kind of triangulation. It can provide insight into the use that cannot be captured through interviews alone, because the interviewee's information on how the WIS is used stems from his own personal experiences. The document life cycle is a perspective on use that it is practically impossible to extract from interviews or direct observations of use. A specific example was the finding that 72% of documents were dead, i.e., never touched since they were initially uploaded. Another example is that although WIS in general and QP in particular are characterised as tools for collaboration and coordination, most documents had a very simple life cycle. They were created and after that a number of users would read it. Only in 22 cases (in total out of all activities performed by 3000 users over 10 months) were documents edited by more than one person. Statistics like these can inform evaluations of an organisation's overall use of WIS beyond the possibilities given by observations, interviews, and surveys. Such statistics might also be automated in order to provide a status that frequently can be (re-)monitored, for example in the case that researchers or management want to monitor changes in WIS use.

Http log files are, in contrast to archival records, not produced intentionally by the organisation. Log files are designed for technical purposes and using log files to understand social practices is not an intended part of the design of the log file. This means that the analysis of log files is sometimes cumbersome and sometimes impossible due to the design of the log file. One consequence of this was that in most cases there was no link between the contents of what was communicated and the information in the log file. Only in the case of attached documents that could be identified and analysed by qualitative techniques, was it possible to relate the information in the log file to a social meaning of the information.

The problem of relating social meaning of information to information in the log file puts severe constraints on the interpretation of the results of the log analysis. This is, e.g., observable in the global models produced for document usage. The types of document life cycles generated from the cluster analysis should not be interpreted as patterns that are directly related to concepts of social practice. A typified document life cycle can cover very diverse uses in terms of the social setting in which it is integrated. As an example, the documents involved in the translation and the documents involved in for example meetings (agenda, minutes, etc.) are all characterised as the same document type in the cluster analysis.

Apart from the methodological challenge of using log analysis, a more practical challenge is evident. Using log analysis in case studies requires knowledge of statistical analysis, data mining, and relational databases. There are many pitfalls in the process of preparing data and analysing data, which can render the results misleading or useless. Paradoxically it requires data analysis techniques in this situation even to discover that errors have been made. The necessary skills are in many cases not present by case researchers who are used to performing qualitative analyses. In the process of research leading up to this article, a great deal of new knowledge concerning computing and statistical analysis had to be acquired during the process, which has been time consuming. Denzin and Lincoln (2000) refer to this as investigator triangulation: Different researchers with different knowledge and background collaborate on collecting and analysing data. Using log analysis in case studies calls for investigator triangulation simply because experienced interview researchers are seldom experienced data miners at the same time.

The research presented in this article has shown that http log analysis is in fact an approach for studying WIS use and should be taken seriously. It not only lends itself to the testing of quantitative hypotheses, but can also be applied in qualitative studies as a means of triangulation. Log analysis can help solve the problem of studying the use of WIS, which is evident when WIS are used in temporally and geographically distributed settings.

7 References

Andersen, J., Larsen, R. S., Giversen, A., Pedersen, T. B., Jensen, A. H., and Skyt, J. “Analyzing Clickstreams Using Subsessions,” *Proceedings of the ACM Third International Workshop on Data Warehousing and OLAP (DOLAP'00)*, R. Missaoui and I. Song (eds.), Washington, DC, 2000, pp. 25-32.

Bansler, J., and Havn, E. "Information Systems Development with Generic Systems," *Proceedings of the of the Second Conference on Information Systems*, Walter R.J. Baets (ed.), Nijenrode University, 30-31 May, Breukelen, The Netherlands, Nijenrode University Press, 1994, pp. 707-715.

Bansler, J. P., and Havn, E. C. "Knowledge Sharing in Heterogeneous Groups: An Empirical Study of IT-Support for Sharing Better Practices," *Proceedings of the Third European Conference on Organizational Knowledge, Learning and Capabilities*, paper no. 282, H. Tsoukas and N. Mylonopoulo (eds.), 5-6 April 2002, Athens, Greece, 2002, pp.1-14.

Bøving, K. B. "Mine the gap—a multi-method investigation of web-based groupware use," unpublished Ph.D., Institute for Film & Media Studies. Copenhagen, University of Copenhagen, 2003.

Bowers, J. "The work to Make a Network Work: Studying CSCW in Action," *Proceedings of the of the ACM Conference on Computer-Supported Cooperative Work (CSCW94)*, R. Furuta and C. Neuwirth (eds.), ACM-press, 1994, pp. 287-298.

Büchner, A.G., and Mulvenna, M. D. "Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining," *SIGMOD Record* (27:4), 1998, pp. 54-61.

Bullen, C. V., and Bennett, J. L. "Learning from User Experience with Groupware," *Proceedings of the of the Conference on Computer-Supported Cooperative Work, October 7-10, 1990 Los Angeles, California*, F. Halasz (ed.), ACM-press, 1990, pp. 291-302.

Cooley, R., Mobasher, M., and Srivastava, J. "Web Mining: Information and Pattern Discovery on the World Wide Web", *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, IEEE, 1997.

Cooley, R., Tan, P.-N., and Srivastava, J. "WebSIFT: The Web Site Information Filter System," *Proceedings of the of the 1999 KDD Workshop on Web Mining*, San Diego, CA, Springer-Verlag, 1999.

Denzin, N. K., and Lincoln, Y. S. *Handbook of Qualitative Research*, Sage Publications, Thousand Oaks, Calif., 2000.

Fu, Y., Sandhu, K., and Shih, M.-Y. "Clustering of Web Users Based on Access Patterns," *Proceedings of the of the 1999 KDD Workshop on Web Mining*, San Diego, CA, Springer-Verlag, 1999.

Gallivan, M. J. "Organizational Adoption and Assimilation of Complex Technological Innovations: Development and Application of a New Framework," *The DATABASE for Advances in Information Systems* (32:3), 2001, pp. 51-85.

Grudin, J. “Why CSCW applications fail: problems in the design and evaluation of organization of organizational interfaces,” *Proceedings of the of the 1988 ACM conference on Computer-supported cooperative work*, I. Greif (ed.), Portland, Oregon, ACM Press, 1988, pp. 85-93.

Grudin, J. “Groupware and social dynamics: Eight challenges for developers,” *Communications of the ACM* (37:1), 1994, pp. 92-105.

Hand, D.J., Mannila, H., and Smyth, P. *Principles of Data Mining*, MIT Press, Cambridge, Mass., 2001.

Hidber, C. *Online Association Rule Mining*, Technical Report TR-98-033, International Computer Science Institute, University of California at Berkeley, September 1998.

Huberman, B. A. *The Laws of the Web: Patterns in the Ecology of Information*. Cambridge, Mass., MIT Press, 2001.

Masseglia, F., Poncelet, P., and Teisseire, M. “Using Data Mining Techniques on Web Access Logs to Dynamically Improve Hypertext Structure,” *ACM SIGWEB letters* (8:3), 1999, pp. 1-19.

Mobasher, B., Cooley, R., and Srivastava, J. “Automatic personalization based on Web usage mining—Web usage mining can help improve the scalability, accuracy, and flexibility of recommender systems.,” *Communications of the ACM* (43:8), 2000, pp. 142-151.

Mobasher, B., Dai, H., Luo, T., and Nakagawa, M. “Effective Personalization Based on Association Rule Discovery from Web Usage Data,” *Proceedings of the WIDM'01 3rd ACM Workshop on Web Information and Data Management*, IEEE, Atlanta, Georgia, 2001, pp. 9-15.

Orlikowski, W. J. “Learning from Notes: Organizational Issues in Groupware Implementation,” *Proceedings of the of the Conference on Computer-Supported Cooperative Work, October 31 to November 4, 1992, Toronto, Canada*, J. Turner and R. Kraut (eds.), ACM-press, 1992, pp. 362-369.

Orlikowski, W. J. “Learning from Notes: Organizational Issues in Groupware Implementation,” *Information Society* (9:3), 1993, pp. 237-250.

Orlikowski, W. J. “Evolving with Notes: Organizational Change around Groupware Technology,” In *Groupware and Teamwork. Invisible Aid or Technical Hindrance?*, C. U. Ciborra (ed.) John Wiley & Sons, Chichester, 1996, pp. 23-59.

Orlikowski, W. J. “Using technology and Constituting Structures: A Practice Lens for Studying Technology in Organizations,” *Organizational Science* (11:4), 2000, pp. 404-428.

Orlikowski, W. J., and Yates, J. "Genre Repertoire—the Structuring of Communicative Practices in Organizations," *Administrative Science Quarterly* (39:4), 1994, pp. 541-574.

Parnas, D.L., and Clements, P.C. "A Rational Design Process: How and Why to Fake it," *IEEE Transactions on Software Engineering* (12), 1986, pp. 251-257.

Pei, J., Han, J., Mortazaviasl, B., and Zhu, H. "Mining Access Patterns Efficiently from Web Logs," *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications (PAKDD'00)*, Lecture Notes In Computer Science, Springer-Verlag, 2000, pp. 396–407.

Pors, J. K., and Simonsen, J. "Coordinating Work with Groupware: The Challenge of Integrating Protocol and Artefact," in *Organizational Information Systems in the Context of Globalization*, M. Korpela et al. (eds.), Kluwer Academic Publishers, 2003, pp. 53-68.

Rogers, E. M. *Diffusion of Innovations, 5th Edition*, Free Press, 2003.

Schmidt, K., and Bannon, L. "Taking CSCW Seriously: Supporting Articulation Work," *Computer Supported Cooperative Work (CSCW): An International Journal* (1:1-2), 1992, pp. 7-40.

Schmidt, K., and Simone, C. "Coordination mechanisms: Towards a Conceptual Foundation of CSCW Systems Design," *Computer Supported Cooperative Work. The Journal of Collaborative Computing* (5:2-3), 1996, pp. 155-200.

Spiliopoulou, M. "Web Usage Mining for Site Evaluation: Making a Site Better Fit its Users," *Communications of the ACM* (43:8), 2000, pp. 127-134.

Spiliopoulou, M., Pohle, C., and Faulstich, L. C. "Improving the Effectiveness of a Web Site with Web Usage Mining," *Revised Papers from the International Workshop on Web Usage Analysis and User Profiling (WebKKD99), San Diego, August*, Lecture Notes In Computer Science, Springer-Verlag, 1999, 142-162.

Srikant, R., and Agrawal, R. "Mining sequential patterns," *Proceedings of the Eleventh International Conference on Data Engineering (ICDE)*, P.S. Yu and A.L.P. Chen (eds.), IEEE, Taipei, Taiwan, 1995, pp. 3-14.

Srikant, R., and Agrawal, R. "Mining sequential patterns: generalizations and performance improvements," *Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology (EDBT)*, Avignon, France, Lecture Notes In Computer Science, Springer-Verlag, 1996, pp. 3-17.

Su, Z., Yang, Q., Zhang, H., Xu, X., Hu, Y.-H., and Ma, S. "Correlation-Based Web Document Clustering for Adaptive Web Interface Design," *Journal of Knowledge and Information Systems* (4:2), 2002, 151-167.

Toolan, F., and Kushmerick, N. "Mining web logs for personalized site maps," *Proceedings of the Third International Conference on Web Information Systems Engineering (Workshops)—(WISEw'02),* IEEE, December 11, Singapore, 2002, pp. 232-237.

Winograd, T. "A Language/Action Perspective on the Design of Cooperative Work," *Human-Computer Interaction* (3:1), 1987, pp. 3-30.

Yates, J., Orlikowski, W., and Rennecker, J. "Collaborative Genres for Collaboration: Genre Systems in Digital Media," *Proceedings of the 30th Hawaii International Conference on System Sciences*, Hawaii, IEEE Computer Society Press, 1997, pp. 50-59.

Yates, J., and Orlikowski, W. J. "Genres of Organizational Communication—a Structurational Approach to Studying Communication and Media," *Academy of Management Review* (17:2), 1992, pp. 299-326.

Yin, R.K. *Case study research: design and methods*, Sage Publications, Thousand Oaks, 1994.