CHAPTER

2

Complementary Explorative Data Analysis

The Reconciliation of Quantitative and Qualitative Principles

FAY SUDWEEKS
SIMEON J. SIMOFF

FOR MANY PEOPLE AROUND THE GLOBE, the Internet has become the place where they instinctively turn to for all kinds of information, particularly after the Coseil Europeen pour la Recherche Nucleaire (CERN) introduced the World Wide Web (WWW or the Web) in the late 1980s. The Internet has given birth to new research fields or has diversified existing research fields connected with human activities, including computer-mediated communication (CMC), computer-supported cooperative work (CSCW), electronic commerce, virtual communities, virtual architecture, various virtual environments, and information design. The Web phenomenon raised a number of research issues concerning the mechanisms and rules governing Internet activities, particularly the interaction of technology and society. Decision makers at different levels needed knowledge about the phenomenon, so politicians, corporate managers, educators, and developers turned their attention to the Internet research community.

In this chapter, we discuss various traditional methodologies and their strengths and weaknesses when applied to Internet-spawned research fields. We find that traditional methodologies need to be adapted to these new research environments in which communication technologies and sociocultural norms challenge existing research assumptions and premises. We propose a complementary explorative data analysis (CEDA) framework within an Internet research schema that integrates qualitative and quantitative procedures. CEDA was inspired by the successful collaboration of the two methodologies in the field of artificial intelligence (AI). The difference is that in AI, qualitative methods still deal with quantities. They operate over their ranges and tendencies in their behavior, not over each possible value. CEDA incorporates complementary use of both methods, depending on the particular research stage or the initial assumptions that need to be taken into consideration, thereby accommodating the unique features of Internet research.

Internet Environment and Research Methodologies

The Internet research community initially endeavored to follow the major macrosteps of classical research work: (a) problem identification and formulation, (b) research design and development of research methodology, (c) data collection, (d) data analysis, and (e) communication of results. These research activities are performed in a linear fashion. Once the problem is identified, the research design becomes crucial for the success of the whole work. Earlier research, however, identified several characteristics of the Internet phenomenon that complicated the use of the classical research schema. Because the human is the central object, participant, information generator, and collector, there was an implicit assumption that the methodology developed in social sciences would be appropriate and adequate.

The majority of social science research work is conducted within the bounds of a narrow set of assumptions, beyond which the researcher rarely deviates. Underlying any research are fundamental philosophical assumptions about ontology, epistemology, and human nature (Burrell & Morgan, 1979; Doolin, 1995; Hopper & Powell, 1985).

Assumptions of an ontological nature are concerned with the physical and social reality of research questions. When applied to Internet research, on the one hand, there is an existing physical medium that supports information communication; on the other hand, around this medium, there exists a global information ether where the social reality takes place. Between the two layers there is an almost invisible connection. However, the parameters of the physical medium, such as the capacity of the links and information storage, affect the social behavior within the information ether. For example, slow links lead to a narrower bandwidth of communication and use of different expressive techniques. An ontological research assumption in this case should make explicit connection between both "realities." *Assumptions of an epistemological nature* are concerned with knowledge. In Internet research, the issue is the distinction between information and knowledge. Is any experience on the Internet a new knowledge or just a transfer of existing knowledge into a new form? For example, should virtual architectures mimic physical architectures or develop their own laws and conventions? *Assumptions of human nature* are concerned with destiny. In Internet research, the issue is the boundary of the environment. Should we consider the Internet an environment in itself or should we consider it a complementary part or an extension of our own environment?

These philosophical assumptions influence the researcher's opinion of what constitutes an acceptable research methodology. A scientist with an objective approach searches for regularities and tangible structures existing in an external world; the researcher who focuses on subjective experience chooses to understand and interpret the individual in relation to, or "being" in, the world. The positivist (or objective) epistemological approach is sometimes labeled as "hard" scientific research. The positivists vary in their research design and methodological approach, ranging from verifying to falsifying hypotheses, but the intent in both instances is based on a belief that there are immutable structures to be discovered, explored, and analyzed. The anti-positivists' (or interpretivists') methodological approach is to be immersed in situations and allow insights to emerge during the process of investigation.

When conducting Internet research, however, there are even more factors to be taken into account. One consideration is the constant and rapid change in technology. A decade ago, most Internet users were, of necessity, skilled computer programmers, or at least, they had a relatively deep understanding of network applications. With the development of point-and-click graphic interfaces, audio and video plug-ins, cableless connections, and Web development applications, the underlying technology is more complex but is a virtually closed system. The effect of this transition is a polarization of the developers and the users in the Internet population. A second consideration is the information now available. The average Internet user is often overwhelmed by the variety and vast amount of information and has difficulty processing and selecting the relevant information. A third consideration is the notion of browsing or "surfing." In contrast to the traditional linear search along shelves of books in a library, the Internet user follows a weblike nonlinear search in which most "pages" emphasize eye-catching designs and attention-grabbing movement rather than a sequential and logical presentation of information.

These considerations complicate classical research methodologies, so increasingly, Internet researchers are turning to methods developed in the fields of information systems and data mining. In general, the research questions of interest appear at first to guide the choice of the research design and methodological tools. At the point when the methodology needs to be selected, the qualitative versus quantitative debate begins. Both methods attempt to explain the implicit concepts hidden in the bulk of data about the investigated phenomenon. However, both methods differ in their approach to the problem.

Quantitative methodologies assume that collected data are measurable, or if they are not, it is necessary to design an experiment or computer simulation in a way that respective measurements can be taken. Once the measurements are done, the problem is to fit (in a broad sense) the data adequately. Derived dependencies are then interpreted in the context of the initial problem formulation with a possible test of the hypothesis about the nature of the data and the errors in the measurements. In qualitative methods, the interest is centered on the qualitative characteristics of the phenomenon. Rather than trying to quantify every detail, these methods try to grasp the form, the content, and some constraints of the investigated phenomenon and analyze its qualities (Lindlof, 1995).

We question, however, this neat qualitative and quantitative dichotomy. We argue that each methodology has its own set of costs and benefits, particularly when applied to Internet research, and that it is possible to tease out and match the strengths of each with particular variables of interest.

Recently, protagonists of both sides have been encroaching cautiously onto rival territory. Thus, researchers may quantify qualitative data—for example, coding concepts from interviews and surveys in a manner suitable for statistical analysis. Researchers may also qualify quantitative data—for example, using quotes from complementary dialogue to support a statistical pattern derived from data collection. Adding a little of one methodology to the other adds flavor and aesthetic appeal, but it is not essential. This is the major drawback in current attempts to develop a research schema that benefits from both methods.

Quantity and Quality: Two Approaches to a Common Phenomenon

Quantitative and qualitative methods are quite distinct in the emphasis they place on each (Stake, 1995). In quantitative analyses, argumentation is based on a representation of the phenomenon as a finite set of variables. There, we seek systematic statistical or other functional relations between these variables. In qualitative analyses, argumentation is based on a description of the research objects or observation units rather than on approximation of a limited number of variables. In other words, in qualitative analyses, references to excerpts or cases in the data are used as clues.

In the next sections, we define distinctive steps in quantitative and qualitative research and compare the methodologies with respect to the major dimensions associated with scholarly inquiry: (a) the purpose of the inquiry, (b) the role of the researcher, (c) the acquisition of knowledge, and (d) presentation of the research.

QUANTITATIVE RESEARCH

Purpose of the Inquiry

The purpose of quantitative research is to explain observed phenomena. It was developed to provide the ability to predict and control examined concepts. Consequently, these concepts need to be quantified. To do this, the researcher needs to know the form, type, and range of the content of the data before the commencement of an experiment. The methodology is based on the model of hypothesis testing. The idea was introduced and developed in the late 1920s and early 1930s. Although in practice there are some variations, ideally, the path of quantitative research is traversed from observation to generation of theoretical explanation to further testing of the theory. Recently, the overall schema has been extended with exploratory data analysis, when hypotheses are formulated and reformulated during the analysis.

The initial step in quantitative research is the design of the experiment. The researcher specifies the goals of the research, the initial hypothesis, and the respective ranges of person responses for measuring quantified concepts.

Each range defines the structure for the data collected. The basic assumption in quantitative methodology is that observations and experiments can be *replicated.* The overall experimental schema needs to be designed in a way that ensures a higher accuracy of the estimation of these quantified values.

Role of the Investigator

The next step is to observe groups of people (study participants) and to record data. The role of the investigator is an *objective* one. The investigator acts just as an observer. In the case of a passive experiment, the researcher only records the observations without setting values to "measured variables." In the case of an active experiment, the researcher may need to intrude and set up some of the variables.

Acquisition of Knowledge

The next step is data analysis. The selection of the appropriate data analysis method depends on the initial assumptions, the nature of experimental observations, and the errors in these observations. On the basis of the numerical results of this analysis, the scientist has to provide some explanations for the observed behaviors and to *construct knowledge.* These explanations are usually in the form of an approximating model. Furthermore, either with or without refining experiments, the researcher might *generalize* these observations and propose a *theory.* Consequently, instead of trying to explain a unique event or phenomenon, the results of the research should apply to a class of cases as well. This theory could be used for building *predictive models* and become the basis for a specific research question, tested in a controlled manner to verify or falsify.

Presentation of Research

The research results are then visualized using a variety of graphing techniques designed to condense the vast amount of raw data. These presentation techniques usually expose some particular characteristics of the data structure and relationships between variables. The researcher has some degree of freedom to tweak the representation of the data to enhance the perception of the results. Usually, each technique has one or more parameters that are sensitive to noise and smoothing. For instance, the appearance of a histogram is largely controlled by the number of bars used to depict the data. When many bars are used, the pattern of the data may look complex with fine-grained details. The reader may wonder if a simpler underlying form exists. On the other hand, the use of too few bars may obscure patterns in the data that are important to the viewer. In this case, the data may look simple with course-grained details, and the reader may wonder if important details are missing.

QUALITATIVE RESEARCH

Often, the researcher is faced with data in the form of loosely structured descriptive texts or dialogues, images, and other illustrations rather than in the form of well-structured records. This, and similar problems, has led to the development of the relatively new method of qualitative research, in which the results are obtained by other than quantification analyses.

Purpose of the Inquiry

The purpose of qualitative inquiry is to understand observed phenomena. Quantitative research begins with a theory formulated as a set of hypotheses, and the purpose of a study is to find support for or to disprove the theory. Qualitative research begins with an area of interest or a research question, and a theory emerges through systematic data collection and analysis.

The object of inquiry for the qualitative researcher is typically a case. A case is a social practice, an integrated bounded system (Smith, 1979) that may or may not be functioning well. Case study is the study of a social practice in the field of activity in which it takes place. Case research is defined as research in which the researcher has direct contact with the participants and the participants are the primary source of the data. It follows, then, that the primary methods used in case research are interviews and direct observations. Other methods, such as experiments and surveys, separate the phenomenon from its context (Yin, 1989).

Role of the Investigator

The starting point for the researcher can be either the case or the question (Stake, 1995). In the former, the case presents itself as a problem, and there is a need or a curiosity to learn more. Because there is a personal interest in the case, it is referred to as intrinsic case study. In the latter, a general problem arouses interest, and a particular case is chosen as a possible source for explanation. Because the case is an instrument to a general inquiry, it is referred to as instrumental case study.

Thus, the *role* of the investigator is *participatory* and personal. The issue on which both approaches differ most is the priority placed on the role of interpretation during this step. All research, of course, requires some form of interpretation, but whereas quantitative research advocates the suspension of interpretation during the value-free period of experimentation, qualitative research advocates actively interpreting phenomena throughout the observation period.

Acquisition of Knowledge

The next step is data interpretation. During this step, the typical qualitative researcher conceptualizes the data and *discovers knowledge.* The conceptualization process ranges from merely presenting the data as they were collected to avoid researcher bias to building a theory grounded in the phenomenon under study. These intuitive and interpretive processes are not regarded as less empirical than quantitative research. Observations and data collection are rigorously systematic, occurring in natural rather than contrived contexts.

Qualitative research is not so much generalization as extrapolation. In certain explicated respects, the results are related to broader entities. The aim is to find out what is specific and particular about the solutions adopted by these people that can be *related* to the broader population. Although the solutions adopted by the people in the case study may be regarded as isolated individual cases and as such as exceptional, some factors are very much the same for a larger population. This means it is possible to conclude indirectly (e.g., referring to other research) in which respects and to what extent the data are really exceptions, in which respects they are comparable to other solutions or population groups, and what sorts of different solutions exist.

Presentation of Research

Qualitative researchers include a great deal of the collected data to present their interpretation of the results. Research reports usually include supporting data fragments in the form of quotes from the raw data. In this case, the researcher can slant the results toward a specific interpretation by exposing particular quotes and omitting others.

Rationale for Integrated Research

Numerous attempts at integrated research over the past two decades have resulted in labels such as *triangulation, micro-macro link,* or *mixed methods* (Bryman, 1988; see also Ragin, 1987; Tschudi, 1989). The idea is to employ a combination of research methods typically used to analyze empirical results or interpretations. The rationale is that the weakness of any single method—qualitative or quantitative—is balanced by the strengths of other methods. In reality, however, the qualitative and quantitative analyses are usually distinct, mutually exclusive components of the research. One component is unstructured textual data of a phenomenon being investigated (e.g., transcripts of interviews or verbal reports from protocol studies), analyzed with an interpretive or hermeneutic method (Prein & Kuckartz, 1995). The other component is numerical data of the same phenomenon (e.g., from a content analysis or a survey questionnaire), analyzed with some statistical procedure. The result is an integrated view that narrowly focuses on a particular social phenomenon.

There is such a variety of social norms that to understand them it is necessary to identify some regularities from observations. Regular patterns are grouped together and form typologies (or categories) of human processes and behavior. The process of typification is a fundamental anthropological technique that enables us to understand our everyday world as well as to conduct scientific inquiries. It is an integral aspect of human thought in that representations of unique experiences or stimuli are encoded into an organized system that economizes and simplifies cognitive processing (Rosenman & Sudweeks, 1995).

Typologies are distinct, discrete classifications of information that help to give order to a confusing, continuous mass of heterogeneous information. In some way, this continuum of information has been divided into discrete regions where points within each such region bear qualitative similarities to each other, whereas points in different regions bear qualitative differences to each other. The construction of meaningful typologies, therefore, is the foundation of scientific inquiry.

Typification as a combined scientific methodology has its foundations in Weber's and Schutz's works (cited in Kuckartz, 1995), who were concerned with linking hermeneutic regularities in texts and standardization of information. Developing this methodology further, Kuckartz (1995) uses a case-oriented quantification model whereby typologies are developed from data rather than predefined. In terms of data analysis, this methodology corresponds to data-driven exploration in which we do not specify what we are looking for before starting to examine the case data. For example, we may parse the text in a sample of e-mail messages looking for concepts that can become the basis for the development of formal models.

Issues Specific to Internet Research

The majority of Internet CMC research is conducted in laboratories under controlled experimental conditions. These studies may not present an accurate picture of the reality of virtuality. The external validity is problematic for three reasons: (a) Study participants are an atypically captive audience; (b) groups studied in experiments tend to be unrealistically small; and (c) an almost natural inclination of experimental design is to contrast with a face-to-face standard of comparison (Rafaeli & Sudweeks, 1997, 1998). This contrast may be misleading.

The replicability of CMC field research is difficult, if not impossible, for two main reasons. On a *technological level,* the Internet is permanently changing its configuration and supporting technology. The underlying networking protocols cannot guarantee the same conditions when replicating experiments simply because each time the path of information communication is unique; thus, the time delay and consequences connected with it are different. On a *communication level,* the difficulties in replication come from the creative aspect of language use. Although the rules of grammar are finite, they are recursive and capable of producing infinite language (Chomsky, 1980). Novel sentences are constructed freely and unbounded, in whatever contingencies our thought processes can understand. Apart from standard cliques, sentences are rarely duplicated exactly, yet each variation is generally comprehended. It follows, then, that experiments involving text generation can rarely be repeated. This lack of replication is a violation of the initial assumptions for the application of statistical analysis.

Another aspect of Internet research is that it has to deal with heterogeneous sociocultural structures. The Internet is, of course, populated with people of many cultures. Culture has been defined as a complex set of behaviors and artifacts with three major dimensions: *ideas* (traditional values and beliefs); *norms* (behaviors that adjust to the environment of traditional values and beliefs); and *material culture* (artifacts produced in the environment of traditional values and beliefs; Bierstedt, 1963). On the Internet, cultural complexity appears to be an intractable problem. Global communication technologies bring together cultures that differ dramatically on each of the three dimensions.

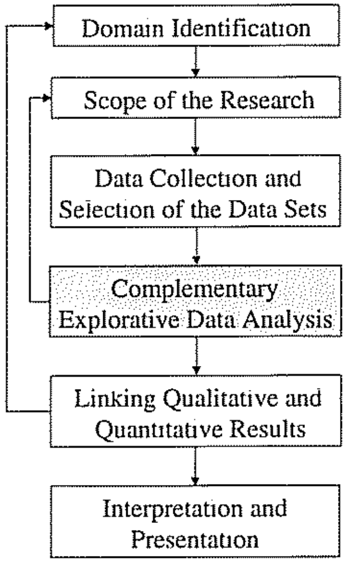
Domain Identification
Scope of the Research
Data Collection and Selection of the Data Sets
Complementary Explorative Data Analysis
Linking Qualitative and Quantitative Results
Interpretation and Presentation

Figure 2.1. Stages of Internet Research

The Internet Research Schema

Although at first glance it seems that quantitative and qualitative research are radically different, they share an important common thread. Both methods make interpretations of the phenomenon they want to examine. Both traditions create a framework for their analysis based on those interpretations. In reality, the difference between these two methods is a discursive one.

To overcome the difficulties outlined in the previous sections, we have developed an integrated methodology for Internet research (Figure 2.1). Internet research incorporates a number of separate research domains, including electronic commerce and business systems, CMC, CSCW, and distributed information systems. Therefore, the first stage is devoted to the identification of domain specifics. These specifics influence the selection of

TABLE 2.1a Quality/Quantity Matrix From a Data Point of View

	Data	
Methods	Qualitative	Quantitative
Qualitative	Survey analysis, interviews, speech acts analysis, participant observation	Qualitative reasoning, constraint reasoning
Quantitative	Data mining, cluster analysis, fuzzy data analysis, neural nets	Statistics, regression and correlation analyses, numerical simulation

TABLE 2.1b Quality/Quantity Matrix From a Methods Point of View

	Data	
Methods	Qualitative	Quantitative
Qualitative	Metaphors, ontologies, categories	Survey data
Quantitative	Text data, vocabulary, categories hierarchy	Numerical samples, coded categorical data, measurements

the appropriate research methods and the possible scope of the research. Once the scope is specified, the schema follows the traditional line of data collection.

The data collected in any of the Internet research domains are a heterogeneous combination of quantitative measurements and qualitative observations. Before the complementary explorative data analysis stage, the researcher defines the combination of methods that need to be used. Table 2.1 illustrates this heterogeneous picture in a "quality/quantity" matrix from both a data (Table 2.1a) and a methods (Table 2.1b) point of view.

Thus, CEDA can be viewed as a dynamic framework that provides valid integration of both methods. CEDA employs quantitative methods to extract

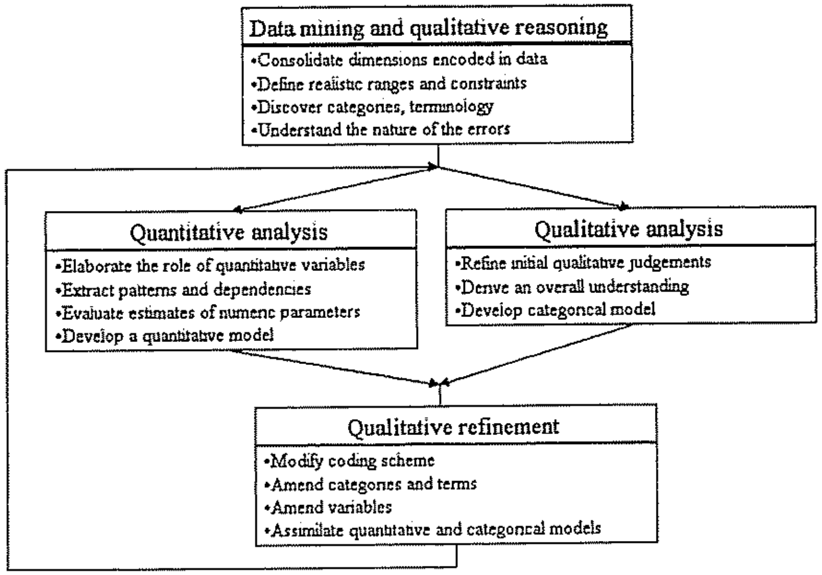
Data mining and qualitative reasoning
•Consolidate dimensions encoded in data
•Define realistic ranges and constraints
•Discover categories, terminology
•Understand the nature of the errors
Quantitative analysis
•Elaborate the role of quantitative variables
•Extract patterns and dependencies
•Evaluate estimates of numeric parameters
•Develop a quantitative model
Qualitative analysis
•Refine initial qualitative judgements
•Derive an overall understanding
•Develop categorical model
Qualitative refinement
•Modify coding scheme
•Amend categories and terms
•Amend variables
•Assimilate quantitative and categorical models

Figure 2.2. Processes in the Complementary Explorative Data Analysis (CEDA) Framework of the Internet Research Schema

reliable patterns, whereas qualitative methods are incorporated to ensure capturing of the essence of phenomena. Figure 2.2 gives a breakdown of the processes in the CEDA framework of the Internet research schema. The frameworks allows the use of different data sets in a common research cycle rather than the traditional approach of applying different analyses to the same data set.

CEDA has the potential to conduct parallel and interconnected research. This complementary analysis requires the linking of the results obtained by each of its components. The final result may lead to revision of the identified domain specifics and changes in the combination of analysis methods within the Internet research schema.

Application of the Internet Research Schema

We now provide an example of the proposed schema applied to CMC research.

DOMAIN IDENTIFICATION

A global society (or cybersociety; Jones, 1995, 1997, 1998) created by the Internet is no longer a projected vision of technocrats; it is becoming a reality. However, the global society may not be the "global village" as envisioned by McLuhan and Powers (1986), but more like virtual neighborhoods (or cybervillages). Before the Web explosion, cybervillages were defined not by geopolitical boundaries but by listserv subscriptions or chat channels. Today, even those loosely defined boundaries are blurred as cybervillages connect to a web of hyperlinks.

As the technology changes at a pace never before experienced, Internet CMC research is engaged in a catch-up situation. A modern Internet research methodology should take into account rapidly changing technology, social norms, and communication behaviors. To be able to specify and develop such a methodology, we need to identify the features specific to Internet communication research.

Communication is computer mediated. First, and obviously, Internet CMC differs from traditional face-to-face communication because the computer provides an interface between interlocutors. A common practice in Internet research is to regard face-to-face conversation as the ideal communication environment (Schudson, 1978), whereas CMC is rated as less than ideal. Experimental work has discovered a number of dysfunctional attributes of computer mediation, including flaming (Mabry, 1998; Siegel, Dubrovsky, Kiesler, & Maguire, 1986; Sproull & Kiesler, 1991) and unsociable behavior (Hiltz, Johnson, & Turoff, 1986; Matheson & Zanna, 1990), disinhibition and deindividuation effects (Hiltz & Johnson, 1989), and a lean environment (Short, Williams, & Christie, 1976; Walther, 1992). Somewhat more optimistic experimental work introduced findings on status leveling (Dubrovsky, Kiesler, & Sethna, 1991), socioemotional connections (Rice & Love, 1987), consensus formation (Dennis & Valacich, 1993), brainstorming creativity (Osborn, 1953), and collaborative productivity (Sanderson, 1996).

Communication requires technical knowledge. Each communication environment requires specific knowledge. In a face-to-face environment, we learn at a very early age not only the phonetics and grammar of the language but also, for example, the management of taking turns in conversations (Sacks, Schegloff, & Jefferson, 1978). In written communication, we add knowledge of orthography and a more formal use of language. In telephone communication, we learn how to search for telephone numbers, to press the right sequence of keys, and to engage in preliminary phatic conversation. Every Internet communicator, however, needs at least minimal technical knowledge of computers. To communicate, even with the simplest graphic mailer, the user needs to know enough of the operating system to launch the application; to compose, reply, and send a message; and to quit the application. As computer technology is being introduced more and more into elementary educational institutions, computer literacy will develop in parallel with linguistic literacy. In the meantime, however, computer literacy is a problem for the majority of current and potential Internet users and affects individual levels of interactivity.

Communication is affected by information and processing overload. Mass communication is ubiquitous, whether active or passive. We all absorb mass communication, whether it is active (television, theater, newspapers) or passive (roadside billboards, newsstand headlines, advertising on public transport). In most instances, we are able to be selective and control the amount of information absorbed. Internet communication places enormous pressures on cognitive processing. Discussion lists often generate hundreds of messages a day, and to contribute to a conversation means responding immediately before the topic shifts and the sequence is lost. On the Web, designers endeavor to engage the browser's attention by manipulating font type and size, text spacing, graphics, colors, backgrounds, video clips, sound bits, animation, and interactive gimmicks. Research has indicated that although minimal levels of novelty can stimulate and demand attention, extreme novelty leads to overstimulation, cognitive overload, distraction, and ultimately, impaired information processing.

Communication has a sense of virtual presence. Communicating with strangers on a regular basis is not new. There have been many examples of "pen pal" relationships that have lasted for many years. The sense of virtual presence in these instances, however, is not strong, because there are long delays between communication exchanges. The message exchange process on the Internet, on the other hand, can be almost instantaneous. The effect is a written correspondence that is like a conversation. Formalities, phatic introductions, signatures, and many other features of written communication are eliminated (Ong, 1982). In such a communication environment, indirect social cues are transmitted, and the virtual presence takes on qualities of a real presence. In fact, quite often, the mental distance between regular participants in discussion groups is less than with colleagues working in the same office.

SCOPE OF THE RESEARCH

Only recently are communication and cultural problems associated with a global community being investigated (e.g., Ess, 1996; Jones, 1995, 1997, 1998; Smith, McLaughlin, & Osborne, 1998; Voiskounsky, 1998). Global norms about privacy, freedom of speech, intellectual property, and standards of conduct are being developed. To understand new global communities, we address two broad aspects of mediated discussions: First, we explore communication patterns of texts, which form part of an ongoing conversation; and second, we explore the process of cohesiveness in a cross-cultural group. Specific questions of interest include the following: How does mediated communication compare with traditional interpersonal relationships? How does the mass-mediated group process work? What features of mediated communication enhance interaction and contribute to the cohesiveness of a virtual community?

DATA COLLECTION AND SELECTION OF THE DATA SETS

On the Internet, the web of computer networks provides a medium for a convergence of communication and social interaction. People congregate in global virtual neighborhoods such as discussion groups and chat rooms to engage in topics ranging from entertaining trivia to philosophical issues. In this chapter, we use qualitative data from publicly archived mediated discussions within these virtual communities. Both data sets consist of e-mail messages. Data Set A includes 3,000 e-mail messages, randomly sampled from network discussion groups between March and September 1993. Data Set B consists of 1,016 messages exchanged among a collaborative group of researchers between May 1992 and April 1994.

COMPLEMENTARY EXPLORATIVE DATA ANALYSIS

Having identified the domain, defined the scope, and collected the data, we now apply the CEDA framework (see Figure 2.3).

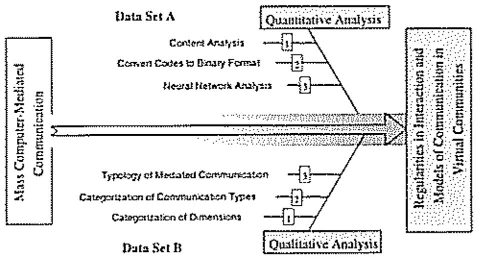
Data Set A
Quantitative Analysis
Content Analysis
1
Convert Codes to Binary Format
2
Neural Network Analysis
3
Mass Computer-Mediated Communication
Regularities in Interaction and Models of Communication in Virtual Communities
Typology of Mediated Communication
3
Categorization of Communication Types
2
Categorization of Dimensions
1
Data Set B
Qualitative Analysis

Figure 2.3. Application of the Complementary Explorative Data Analysis (CEDA) Framework

Quantitative Analysis

To understand more about global cultural norms, we focused on communication patterns in virtual communities. In particular, we were interested in features of messages that form part of an ongoing conversation—that is, messages that engage group members sufficiently to participate and respond and thus contribute to the development of group cohesiveness and consciousness.

Step 1: Content Analysis. Because the situational conditions are unknown prior to the study, variables are experientially rather than operationally defined, and some of the variables develop throughout the study. Texts of Data Set A were coded on 46 variables. Each message was described in terms of features and content, such as relevance, time, tone, purpose, and so forth. The codes were a mixture of objective and subjective ratings (see Sudweeks & Rafaeli, 1996, and Rafaeli, Sudweeks, Konstans, & Mabry, 1998, for a detailed description of the content analysis).

Step 2: Converting Codes to Binary Format. For a quantitative analysis, we chose to use a neural network because it allows a typology of features to emerge. Data analyses, such as a Euclidean cluster analysis, provide techniques for identifying correlations between particular features in a given data set, a useful indication of where the aggregation (boundaries) within a data set might appear. This form of analysis is widely recognized as providing a static view of data (a "snapshot" of typical and atypical instances) because the clusterings are based entirely on pairwise correlations. An alternative to the cluster analysis is the autoassociative neural network (ANN) in which clusterings are more dynamically created across all features synchronously. This quantitative method is modeled on human cognition. Features are drawn into particular groupings and form dynamic allegiances that can effectively overrule the original cohesion based on a simple pairwise correlation. The pattern of network activation captures complex information about dependencies between combinations of features.

ANNs are special kinds of neural networks used to simulate (and explore) associative processes. Association in these types of neural networks is achieved through the interaction of a set of simple processing elements (called *units*) connected through *weighted connections*. These connections can be positive (or *excitatory*), zero (no correlation between the connected units), or negative (*inhibitory*). The value of these connections is learned during a Hebbian training procedure (see Berthold, Sudweeks, Newton, & Coyne, 1998, for a detailed description).

To prepare the data for the ANN, codes identifying author, coder, and message number were deleted and the remaining variables converted into a binary format for processing. Each entry was split into as many mutually exclusive "features" as the entry had options. Because the main focus of interest was conversation threads in group discussions, three new entries were extracted from the original database to explore interactive threads:

1. *Reference height:* how many references were found in a sequence before this message
2. *Reference width:* how many references were found that referred to this message
3. *Reference depth:* how many references were found in a sequence after this message

Thus, as a preliminary result of the recoding of the data, we obtained a formal model of a thread in CMC (Figure 2.4). In addition to the threefold split proposed by Berthold et al. (1997, 1998) we included explicitly the time variable. Each message is completely identified by two indexes—one for its level and one for its position in time in the sequence of messages at this level. Such a model allows the comparison of the structure of discussion threads both in a static mode (e.g., their length and width at corresponding levels)

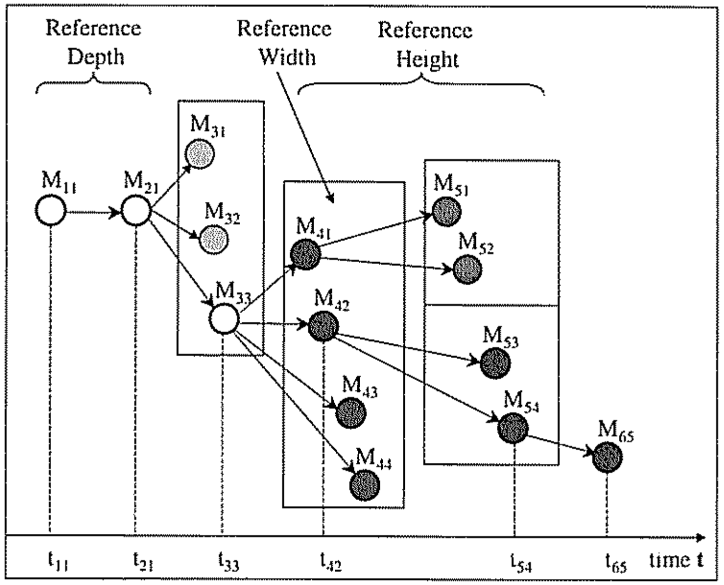

Figure 2.4. Formal Model of a Thread in Computer-Mediated Communication

and in a dynamic mode (e.g., detecting moments of time when one thread dominates another in multithread discussions, such as those that occur on bulletin boards or in MOO [multi-user domains object oriented]-based educational environments).

Step 3: Neural Network Analysis. After processing, the data consisted of 149 binary features; that is, each feature had a value of "1" (present) or "0" (not present). To identify typical features present in messages that stimulate conversation, one feature is *clamped* (forced to be present with a value of "1") to restrict the feature space of solutions. After training, the network settles on a pattern of features typically associated with the clamped feature. For example, Table 2.2 shows the frequency of features present in messages that contain "humor."

The sensitivity of associative features was then calculated. Distinguishing features of interactive messages (referenced to or referenced by another message) and noninteractive messages with their sensitivity scores are given in Tables 2.3 and 2.4 respectively.

TABLE 2.2 Frequency of Feature Activations When "Humor" Is Present

Feature description	*Frequency*
Medium-length message (11-25 lines)	50%
Appropriate subject line	90%
Does not contain question or request	70%
Appropriately formatted	70%
Male author	80%
Contains no abusive language	100%

TABLE 2.3 Distinguishing Features of a Typical Interactive Message

Feature Description	*Sensitivity Score*
Medium length (11-25 lines of text)	1
Appropriate subject line	1
Contains statement of fact	1
No question or request	1
No emoticons	2
No punctuation device to express emotion	3
Male author	2
Identifies gender by name and/or signature	3
Does not include quoted text	2
Addresses another person	1

TABLE 2.4 Distinguishing Features of a Typical Noninteractive Message

Feature description	*Sensitivity Score*
Does not refer to previous message	4
New topic, not referring to previous discussion	18
Does not use first person plural	1
Is not referred to by later messages	9

Qualitative Analysis

The second aspect of cultural norms in virtual communities is to identify communication features that contribute to an ongoing conversation and interaction in an environment in which many traditional features of interpersonal relationships are not present. Data Set B provides the data for qualitative analysis.

Step 1: Categorization of Dimensions. First, the messages were reviewed to identify and categorize major dimensions or regularities that occurred throughout the data. Five salient dimensions were identified:

1. *Issues:* the topics to be discussed and resolved
2. *Leadership:* the inclination to conform or reject leadership and authority
3. *Debate:* argumentativeness, criticism, or aggression among participants
4. *Relationships:* expressions or avoidance of friendship or intimacy among participants
5. *Action:* goal-directed or task-directed activity (Sudweeks & Allbritton, 1996).

Following a technique developed by Romm and Pliskin (1995), each occurrence of a dimension was highlighted and labeled. The dimensions provided the means for observing the emergence of "turning points." Turning points, or changing patterns of regularities, indicate the development of group communication norms and standards. For example, early in the period, the focus of group discussions was on methodological issues to be resolved and on who should be responsible for coordinating the project. At a later point, the discussions became volatile, thereby introducing another dimension. The group dynamics therefore evolved to a different phase at this point.

Step 2: Categorization of Communication. The texts were reviewed again to identify not only how communication behaviors were managed but also the types of communication content. The content fell into three broad categories: (a) conceptual, (b) socioemotional, and (c) action (task oriented). The communication was managed in both a formal and informal manner.

Management of Communication. *Informal management* is the collective informal creation and enforcement of communication norms. Norms are mutually acceptable definitions of communication behaviors among individuals so that interactions can be organized into an agreed-on state:

Example

We seem to be getting semiserious about this. Maybe one tentative and fairly easy way to proceed is to appoint Basil and Cyril the "leaders" (not because they talk the most, but because this is already their research interest and they have some experience in it). (May 28, 1992)

Formal management is connected with the enforcement of rules. Formally, management is needed to generate information, process knowledge, and disseminate the products of knowledge. Whereas informal management is generally performed on a collective level, formal management of communication occurs on an individual or small-group level.

Example

As this project begins to take on the prospects of developing a real finished product (i.e., the coded database), I think it might be appropriate for us to discuss the future "ownership" of that data.

Content of Communication. *Socioemotional communication* is content that deals with interpersonal relationships among the communicators. Socioemotional communication addresses the creation of relationship norms among communicators.

Example

First . . . I waded in here over the weekend, got into a barroom fight or two (there IS a certain amount of Dodge Citydom in the current situation), left, and was persuaded by Frank that I was not dealing with a crew of ogres, unemployed CIA operatives, and voyeurs. (June 26, 1992)

Conceptual communication involves the creation and prescription of shared rules to follow and involves a medium to high level of interactivity. Conceptual communication often requires that implicit communication be made explicit. Realistically, it is not always possible to have complete or full shared creation of mutual understanding of meaning, but this is what conceptual communication strives for.

Example

My reading of [this] question is not so much how CMC changes communication, but whether one can predetermine the cognitive approach by pre-selecting the form of communication. The assumption there is that various disciplines think in very specific ways, and that each way can be matched to communication forms. (June 13, 1992)

TABLE 2.5 Communication Management and Content in the Development of the Virtual Community

Time Period	*Dimensions*	*Communication Management*	*Communication Content*
1	Issues, leadership	Informal	Conceptual
2	Issues, leadership, debate	Formal	Conceptual, socioemotional
3	Leadership, relationships	Informal	Socioemotional
4	Issues, leadership, action	Formal	Task oriented
5	Issues, leadership, debate, action	Formal	Task oriented
6	Relationships	Informal	Socioemotional

Task communication deals with the explicit work to be accomplished. Task communication focuses on information content of communication, whereas conceptual communication focuses on the creation of meaning preceding the processing of information. Task communication deals with specific activities to be completed by members and often has to be conducted independent of other group members. Task communication can be defined as information exchange rather than communication.

Example

As a consequence of being the only person to answer [the] call for volunteers to act as "Oracles" during project coding . . . my first task is to recruit others . . . (April 16, 1993)

Step 3: Typology of Dimensions and Communication. In a third review of the texts, the texts were divided into time periods, delineated by the turning points identified in the first review, and the frequency of communication types in each period was calculated. Table 2.5 shows the importance of communication styles in the development of an interactive virtual community.

Evaluation and Future Applications

We have examined the collection of data and the theoretical aspects and applicability of quantitative and qualitative analyses in Internet research. On the basis of these outcomes, we proposed an adaptive Internet research schema that combines consistently both methods. The CEDA framework can be applied to a variety of Internet research fields, including the following:

1. *Virtual communities.* Labeled initially as "virtual" to stress the absence of face-to-face physical presence, these are CMC communities that are real. These communities are established either on the basis of asynchronous e-mail message exchange or on a synchronous presence in text-based virtual environments—for example, MOOs and MUDs (multi-user domains).

2. *Internet-based distance education and on-line learning.* The communication between students and between students and educators is an essential part of current distance education. With the introduction of Internet and Web-based course delivery, communication becomes an essential part of course support. Technological support is a necessary condition for conducting successful collaborative studies, but a sufficient condition is the use of appropriate methodology. Experience from Web-mediated courses (Simoff & Maher, 1997) suggests that learning approaches taken from face-to-face courses need to be reconceptualized to take into account the unique opportunities offered by distributed computer media. The Internet research schema presented here is useful for elaborating and improving student communication in these new course environments. Applying the methodological schema to conduct research in this field will lead to the evaluation of practical specifications.

3. *Virtual organizations and intranet corporate research.* The methodological schema could be used for the analysis of the content of e-mail and multimedia communication, styles, efficiency, and productivity in traditional practices and emerging new business units—virtual organizations.

4. *Business information systems.* The methodological schema could provide results for improving the content-based information retrieval—the kernel of multimedia business systems. The research schema includes analysis of corporate e-mail, extraction of descriptive categories, compiling ontological representations of the results, and incorporation of these ontologies in the intranet search engines, thus shifting retrieval from simple keyword matching to categorical identification and category-based retrieval.

References

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

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

Bierstedt, R. (1963). *The social order.* New York: McGraw-Hill.

Bryman, A. (1988). *Quantity and quality in social research.* London: Routledge.

Burrell, G., & Morgan, G. (1979). *Sociological paradigms and organisational analysis: Elements of the sociology of corporate life.* London: Heinemann.

Chomsky, N. (1980). *Rules and representations.* Oxford, UK: Basil Blackwell.

Dennis, A. R., & Valacich, J. S. (1993). Computer brainstorms: More heads are better than one. *Journal of Applied Psychology, 78*(4), 531-537.

Doolin, B. (1995). Alternative views of case research in information systems. In G. Pervan & M. Newby (Eds.), *Proceedings of the 6th Australasian Conference on Information Systems (ACIS'95)* (pp. 767-777). Perth, Australia: Curtin University of Technology.

Dubrovsky, V. J., Kiesler, S., & Sethna, B. N. (1991). The equalization phenomenon: Status effects in computer-mediated and face-to-face decision-making groups. *Human-Computer Interaction, 6,* 119-146.

Ess, C. (Ed.). (1996). *Philosophical perspectives on computer-mediated communication.* New York: SUNY Press.

Hiltz, S. R., & Johnson, K. (1989). Experiments in group decision making, 3: Disinhibition, deindividuation, and group process in pen name and real name computer conferences. *Decision Support Systems, 5,* 217-232.

Hiltz, S. R., Johnson, K., & Turoff, M. (1986). Experiments in group decision making: Communication process and outcome in face-to-face versus computerized conferences. *Human Communication Research, 13,* 225-252.

Hopper, T., & Powell, A. (1985). Making sense of research into the organizational and social aspects of management accounting: A review of its underlying assumptions. *Journal of Management Studies,* 22(5), 429-465.

Jones, S. G. (1995). *CyberSociety: Computer-mediated communication and community.* Thousand Oaks, CA: Sage.

Jones, S. G. (1997). *Virtual culture.* London: Sage.

Jones, S. G. (1998). *CyberSociety 2.0: Revisiting CMC and community.* Thousand Oaks, CA: Sage.

Kuckartz, U. (1995). Case-oriented quantification. In U. Kelle (Ed.), *Computer-aided qualitative data analysis: Theory, methods and practice.* Thousand Oaks, CA: Sage.

Lindlof, T. R. (1995). *Qualitative communication research methods.* Thousand Oaks, CA: Sage.

Mabry, E. A. (1998). Frames and flames: The structure of argumentative messages on the net. In F. Sudweeks, M. McLaughlin, & S. Rafaeli (Eds.), *Network and Netplay: Virtual groups on the Internet* (pp. 13-26). Menlo Park, CA: AAAI/MIT Press.

Matheson, K., & Zanna, M. P. (1990). Computer-mediated communications: The focus is on me. *Social Science Computer Review, 8*(1), 1-12.

McLuhan, M., & Powers, B. R. (1986). *The global village: Transformations in world life and media in the 21st century.* New York: Oxford University Press.

Ong, W. J. (1982). *Orality and literacy: The technologizing of the word.* London: Routledge.