CONTINUOUS ANALYSIS OF INTERNET TEXT BY ARTIFICIAL NEURAL NETWORK

by

Peter Jörgensen
September 1, 2002

A dissertation submitted to the
Faculty of the Graduate School of State
University of New York at Buffalo
in partial fulfillment of the requirements for the
degree of

Doctor of Philosophy

Department of Communication

Dedication

This dissertation is dedicated to my parents, both of whom earned advanced degrees late in life and especially to my wife, Corinne, whose love and support have always been greater than I deserve.

Contents

Table of Tables

Table of Figures

Abstract

Many segments of modern society desire the ability to find relationships, thus meaning, in public discourse. Notable segments include marketing, politics, government, activism and public safety. The increasing use of the Internet for public dialog has made Internet communication a potentially rich source of information in this regard. This study explores the use of an Interactive Activation with Competition (IAC) artificial neural network to aid in processing text to find relationships. A fully recurrent IAC network was modified to process text in a continuos fashion, adding nodes as new terms were encountered. Forty-nine email messages from two threads in the Open Library/Information Science Education Forum were processed using two variations in the self-organizing phase of network formation. The messages were processed with and without a linear decay function applied to the external activation of nodes between sentences. With the function self-organization includes reduced external activation levels of terms in sentences that have already been processed in the current message. Without it self-organization externally activates only terms in the sentence currently being processed. This could be considered a type of context control. The use of the linear decay function produced three effects. When the function was used, roughly half the number of noise strings were highly associated with key terms; the entire network was more differentiated from key terms and; the key terms were more highly associated with each other. These effects could reduce or eliminate the need for stop word filtering as well as improve system performance. Future research should explore refinements in this concept, as well as its ability to disambiguate terms. Combining this approach with models of discourse structure is another area for research.

Chapter 1
Introduction

Individuals and organizations are interested in accurate and robust means to track opinions and attitudes for marketing, political, and social reasons (Lasswell, 1931 p317). Attitudes and opinions are expressed and shared through a variety of communication modalities. One very widely used form of interpersonal communication is email (Georgia Tech Research Corp., 1998). Email messages in public forums, such as listserv lists, can be a steady and sometimes vast source of personal expression in machine readable text form. List email is created at an estimated rate of 36.5 billion messages or 675 terabytes per year (1.8 gigabytes per day) (Lyman, 2000). This is clearly too much for existing manual methods to index or categorize. Analyzing this material requires tools and techniques that can deal with large amounts of heterogeneous text. One such technique is text mining which is the discovery of useful information by automated analysis of non-homogeneous collections of text (described in more detail below).

Artificial neural network (ANN) software is able to discover relationships between terms in large bodies of natural language text (Woelfel, 1993; Merkl, 2000) in much the same way that the human brain does – by developing associations between words based on exposure to those words in context. Although ANN tools are very powerful when applied correctly to appropriate texts they suffer from the limitation of operating in batch mode and requiring considerable human intervention to optimize their performance in a particular domain and with a particular type of text. This is not an issue when the text is being studied in a historical context, such as the reasons behind shifts in public attitudes during a war, because the possibility exists to analyze and reanalyze the

text without regard to the time that passes during the analysis. Some social scientists, however, are interested in real-time events, such as shifting public opinion or the changing mood of investors, consumers or other groups. Other factors that make ANN techniques less than fully ideal for real time text analysis are their general dependency on the creation of stop word lists (Pasquale, 2001; Sanderson, 1994) and, their lack of use of context (Elman, 1990) and the structure of discourse (Liddy, 1991). ANN tools would, therefore, be greatly more useful if they could be made to depend less on operator intervention.

This dissertation deals with texts, specifically text streams carried on the Internet, and their automatic processing by artificial neural network software. To that end, this dissertation investigates modifications of an interactive activation and competition (IAC) artificial neural network algorithm to: include the ability to add terms to the network as they are encountered in new texts; handle all terms encountered (i.e. without ignoring stop words) and; incorporate the notion of context. Two approaches to external activation of terms will be investigated. The first approach sets the external activations of all terms to zero after each Hebbian learning cycle takes place, i.e. after each sentence is processed. The second approach, which is an attempt to incorporate context, reduces the external activations of the terms (without going below zero) by a reduction factor after the Hebbian learning cycle has been run. Only when the beginning of a new message is encountered are the external activations of all terms set to zero. The following review of the literature will provide an overview of the state of our knowledge about text, artificial neural networks, and Internet communications.

Chapter 2
Literature Review

Text analysis

Introduction

Part of science is the discovery of associations and patterns in the world around us. According to John Barrow: "the goal of science is to make sense of the diversity of Nature..." [through] "the transformation of lists of observational data into abbreviated form by the recognition of patterns," all with the goal in mind of "algorithmic compression." (Barrow, 1991, p10-11) Of particular interest to scholars of communication and informatics are the patterns and associations that occur in texts, be they newspaper articles, works of literature, or conversations. These patterns are studied to learn about social structures and interactions, individual thought processes and the entire range of cognitive and communicative processes between. The study of patterns in text has been applied to a variety of domains including literature (Pasquale, 2001), mass media (Danielson, 1997), political science (Franzosi, 1997), philosophy (Pasquale, 2001) and others.

Definition of "text"

From an historical review of the field of text or content analysis Popping (2000) adopts the following definition of text: "every *linguistic* means of expression (thus music, for example, is excluded…)" (emphasis his) (p. 8). Other definitions include "every *semiotic* structure of meaning" and simply "*written* language" (Popping, 2000, p. 8). For the purpose of this study I will be adhering to the

simple definition of *text as written linguistic expression*. Text, as defined here, is produced in a variety of situations from the newsroom to the bedroom. Increasingly, text production and communication is mediated by technology. This dissertation will examine the analysis of texts created during technology mediated communication (TMC).

Content analysis

Analysis of text is usually referred to as *content analysis*. Content analysis has been used by public opinion researchers especially to study the effects of the media on public opinion (Rubenstein, 1995; Breen; McMillan, 2000; Schneider, 1997; Shi-xu, 2000, and others). It has found widespread use in the study of literary and philosophical corpora (McKinnon, 1989), the fields of rhetoric, political science, and others (Palmquist, 2002). After a brief introduction to the history of content analysis I will discuss three categories of content analysis, namely thematic text analysis, semantic text analysis, and network text analysis. I will also discuss some of the computer tools developed to assist in textual content analysis.

History of content analysis

Content analysis encompasses a broad range of types of communication including text, music and art. Drawing on a number of sources, Shapiro and Markoff (1997) define content analysis as "any systematic reduction of a flow of text (or other symbols) to a standard set of statistically manipulatable symbols representing the presence, the intensity, or the frequency of some characteristics relevant to social science" (p. 14). Their specification of the object of this methodology as "a flow of text" is particularly applicable to this study's analysis of real-time TMC texts. It should be noted that the end result of content analysis

is measurement of the presence, the intensity, or the frequency of some characteristics. Shapiro et al., assume that measurement requires the reduction of the text to a set of symbols which can be statistically manipulated (Shapiro, 1997).

Early content analysis was a laborious manual process. In the 18th century an analysis consisting of counts of religious symbols in songs was done in Sweden (Krippendorff, 1980, as cited in Popping, 2000, p. 2). This represents the first quantitative text analysis on record (Popping, 2000). More recently Lasswell et al. analyzed text from newspapers relying primarily on hand coding and counting (1931).

The three main types of content analysis

Thematic text analysis

Thematic text analysis attempts to identify concepts or themes, especially of a recurring or changing nature in texts (Stone, 1997). The concepts of interest are counted and compared, for example, at different times or in different sources (e.g. newspapers). The recording unit in thematic text analysis might be column-inches, headlines, or particular words or concepts. A form of thematic text analysis known as contingency analysis looks at co-occurances of themes. Thematic text analysis has been used by Naisbitt (1982, cited in Stone, 1997) to determine that certain newspapers can serve as "bellwether" indicators of public opinion, thus augmenting standard public opinion polling techniques. The increasing role of the Internet as a medium for the dissemination of news and opinion may indicate that the same holds true for it.

Semantic Text Analysis

Semantic text analysis investigates the relationships between the themes in a text, thus going beyond the simple descriptive counting of themes to a level of analysis that maintains "the narrative flavor of the original text." (Popping, 2000, p. 27) "[S]emantic text analysis must begin by isolating a population of texts that exhibit the structure assumed in the research at hand." (Roberts, 1997, p. 59) Semantic text analysis makes use of templates called semantic grammars into which the text is coded. A typical four part semantic grammar might include *Agency, Position, Action* and, *Object* (Popping, 2000, p. 28). The text is analyzed one clause at a time, that is in blocks containing one inflected verb. The words in the text block are coded as to their syntactic function (e.g. agent) and their valence. Valence is usually positive or negative, or some other evaluative relationship.

Network Text Analysis

Taking the analysis one step further, network text analysis looks at all of the words as they relate to each other rather than just within separate blocks. A network of the words is created by directionally connecting words to those other words that have, for example, a causal relationship. The connections also have strengths associated with them, some words thus being more strongly connected than others. Any term in the map may be viewed as a focal term, in which case its connections form the basis of the analysis. The cognitive mapping model views the resulting network as representative of the mental map that the text's author had in mind when the text was produced (Carley, 1997). For a review of the progression of early, non-computer assisted text analytical methodologies including *frequency analysis* (until the 1950s), through *valence-analysis* (middle 1950s), *intensity analysis* (1950s and '60s), to *contingency analysis* (from 1960) see

(Popping, 2000, p. 2-14).

Content analysis assisted by computer

Early work in machine (or computer) analysis of texts was done in the field of information retrieval (IR), particularly in indexing and query refinement. Indexing of texts has traditionally been performed manually by trained indexers using specific rules and controlled vocabulary to facilitate the retrieval process (which is also done by trained individuals.) In an effort to find ways to speed up the indexing process, as well as reduce costs and improve retrieval, many experiments have been carried out and systems built based on computerized analysis and indexing of texts. Among the earliest of these were the SMART (System for the Mechanical Analysis and Retrieval of Text) system created at Cornell University (Salton, 1968). Salton documents the ability of the SMART system to match the performance of a traditional manual system (Medlars) in recall and precision by utilizing discriminator dictionaries and thesauri. Stone (1997) points out that computers can enhance thematic text analysis by overcoming the bounds of coders' rationality and attention spans in analysis of lengthy texts and many themes.

Natural language processing

Natural language processing (NLP) is a class of computer techniques that attempts to map the words and phrases in a text to concepts and relationships between these concepts using a rules based approach. The syntactic function of the words in the text, usually one sentence at a time, is determined; the meanings of the words are established, usually taking the context into account; and finally the relationships between these are derived. Due to the inherent ambiguity of most natural languages, NLP has had limited success "understanding" texts. The

best systems have been restricted to a particular domain such as insurance claim processing (Liddy, 1989).

Cluster Analysis

Like multivariate analysis, cluster analysis looks for factors that contribute to an overall phenomenon. In addition it attempts to group the factors into related clusters. It requires that first a similarity matrix is constructed and assumes that the data are valid, that is, it does not provide tests to determine whether the data is internally consistent or not (Thapalia, 2001). Among the methods of calculating the similarity matrix are the nearest neighbor, farthest neighbor, cosine method and various artificial neural networks. A common form of output for the hierarchical cluster analysis is the dendrogram (Fielding, 2000). The *k-means* type of cluster analysis does not produce a hierarchy of clusters.

N-gram Grammars

N-gram grammars utilize probabilities to disambiguate words in a text. The probabilities are based on prior analysis of existing texts in the language and domain of the text being analyzed. All words that are included in the system are assigned probabilities of occurrence given the co-occurances of the previous n-1 words in the text. Most n-gram grammars are bi-gram or tri-gram, that is they are limited to using the previous one or two words in the sentence to calculate the probability of the word in question. Even with this limitation (i.e. n=2) the number of parameters is quite large on the order of 25,000,000 for a vocabulary of 5000 (Stolcke, 1994). Of particular significance to this dissertation is the fact, as Stolcke (1994) points out, that n-gram grammars are not extensible, that is, the n-gram is not of any use in predicting the probability of new terms. This dissertation will investigate characteristics of text without explicitly identifying

the symbols.

Content analysis techniques based on text all share the drawbacks implied in limiting the analysis to the textual component ignoring other aspects of the communication (e.g. visual or music) In some cases, however, this is not a serious limitation because the communication being analyzed (e.g. newspaper editorials, internet chat) is carried out entirely through text as defined here.

Text Mining

Text mining is a relatively new form of text analysis which has been made possible largely by digital computer technology and the increase of texts available in machine-readable, or digital form. It is related to data mining or knowledge discovery in databases (KDD) in that it attempts to synthesize knowledge by "identifying valid, novel, potentially useful, and ultimately understandable patterns in data."(Fayyad, 1996, p.51)

Swanson (1988) demonstrated the utility of searching collections of documents for patterns of information that no one document contains. His research involved the literature regarding the role of magnesium in epilepsy and the relationship between epilepsy and migraine headaches. His work resulted in two subsequently borne out hypotheses regarding the role of magnesium in migraines. None of the medical literature of the time, however, associated these two concepts. Only by studying a collection of (somewhat unrelated) documents was Swanson able to make his discovery. Other examples of the value of looking at document collections are reviewed elsewhere – see, for instance Dworman (1996).

Text mining "may be defined as the process of analyzing text to extract

information that is useful for particular purposes" (Witten, 2001, p. 1). To date, most of the attention of researchers in text mining has focused on static texts such as web pages (Hearst, 1999) or writings of a single author (Dworman, 1996). Text mining differs, however in that while KDD deals with information residing in databases, and which is, therefore, structured, text mining deals with unstructured text.

Text mining has been likened to "extracting precious nuggets of ore from otherwise worthless rock" (Hearst, 1999). In this analogy one searches not for a fact or the answer to a specific question, but instead looks for valuable information or knowledge that exists only in a large corpus of text. One of the earliest examples of the power of text mining is the well-known research by Swanson (1993), discussed above. Various approaches to text mining have been developed including information extraction (Nahm, 2001), hybrid use of computational linguistics and user-guided analysis (Hearst, 1999), the use of the z-score (Bradley, 1992), cluster analysis, multidimensional analysis (Shapiro, 1997), and artificial neural networks (Woelfel, 1993). This last tool, the artificial neural network, is central to this dissertation and will therefore be discussed in detail.

Artificial neural networks

This dissertation explores the use of using a type of artificial neural network (ANN) known as the interactive activation with competition (IAC) model to discover term associations in text streams. This section will review the history of ANN methodology as well as some of its key limitations and strengths.

Theory and history of artificial neural networks

Artificial neural networks implement algorithms inspired by how the neurons of the brain (and in some cases sensory organs) are thought to perceive, learn, and recall patterns. The fundamental concept, traceable to McCulloch and Pitts (1943 as cited in Garson, 1998) is that neurons (variously called units or nodes in ANN implementations) are highly interconnected by links whose strengths (or weights) have been dynamically adjusted on the basis of algorithms when stimuli were applied. Thus when stimuli (words, concepts, pixels) are presented to the system the associated neurons (nodes) are activated and strengthen their connections with each other. Depending on the exact model and algorithms used, connections to inactivated nodes may also be weakened, a process mimicking forgetting. In this way ANNs become sensitized to patterns of stimuli which, when presented at a later time can then elicit a response, even when the patterns are incomplete or inexactly match a "learned" pattern. ANNs used in this way are a type of associative memory.

After an initial flurry of excitement and progress, early work in ANNs was limited by the misconception (caused by a misreading of Minsky and Papert's 1969 *Perceptrons*) that they could not solve an XOR problem (Garson, 1998). In other words, it was believed that they could only solve linearly separable problems (Abdi, 1999, p. 18). A linearly separable problem is one in which a (hyper)plane can be drawn such that all members of one solution set lie on one side of the plane and all members of the other solution set lie on the other. A simple two dimensional example of a linearly separable problem can be seen in Figure 1. A linearly inseparable XOR problem is illustrated in Figure 2. In these examples the points shown as Xs belong to one group, based on some criteria, and the points shown as Os belong to another. The Xs and Os in Figure 1 lie on opposite sides of a straight line whose slope and intercept define the boundary between the groups. In Figure 2, the non-linearly separable example, there is no

straight line that can be drawn which will separate the Xs from the Os. This assumed limitation of ANNs was shown by Rumelhart and McClelland to apply only to simple one layer networks (Rumelhart, 1986). The development of the multilayer network and other techniques, especially refinements in learning rules, revived interest in neural network research in the late 1980s (Garson, 1998). Since then neural network techniques have found widespread use in many fields which require robust techniques for prediction. These fields include economics, meteorology, and sociology. Much ANN work centers around pattern recognition leading to applications in psychiatry, criminology, and classification. This is due to the network's ability to recognize patterns that are incomplete and/or contain extraneous data.

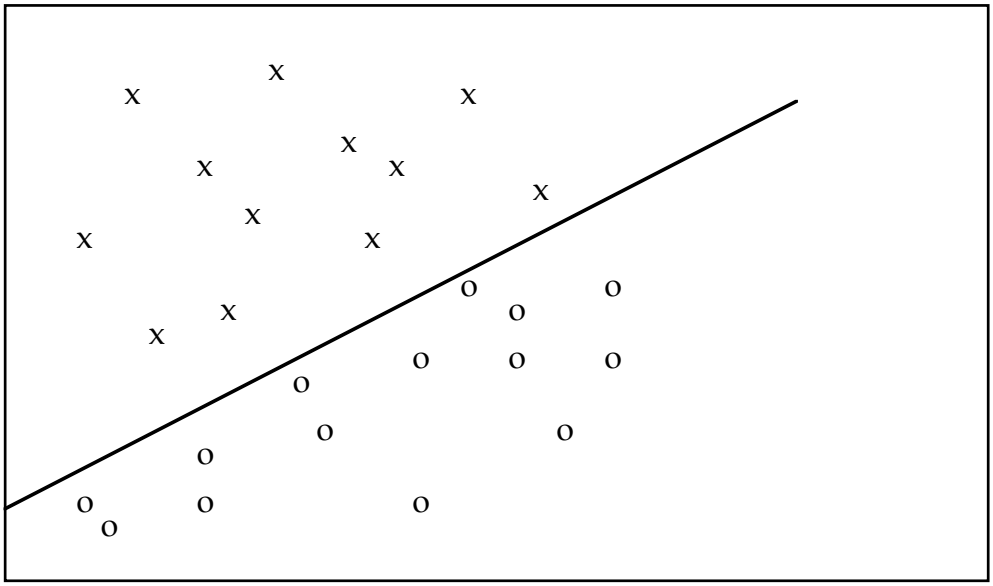


Figure 1 – The Xs and Os are linearly separable

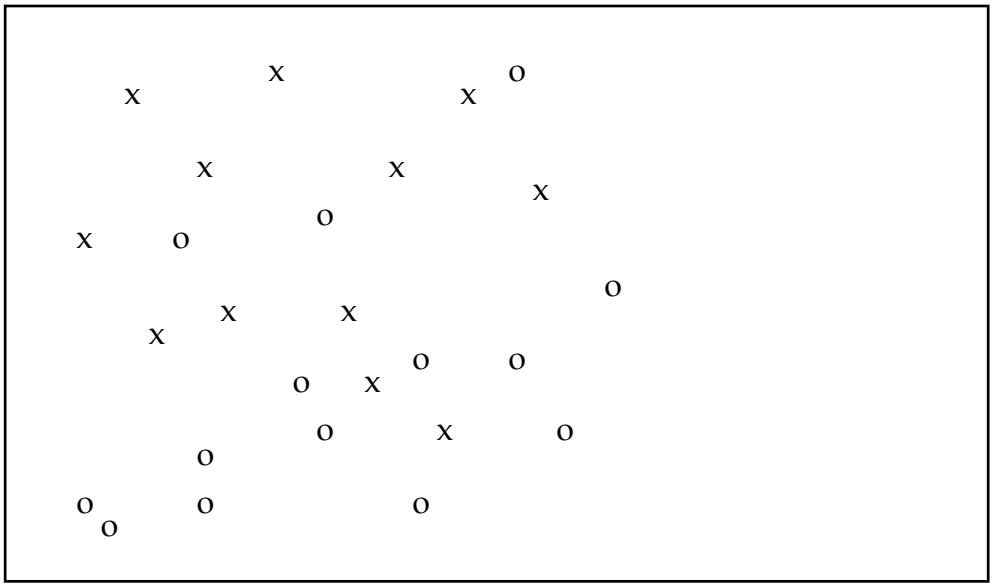


Figure 2 – The Xs and Os are not linearly separable

Types of Artificial Neural Networks

Artificial neural networks consist of processing units or nodes. Each node of an ANN is connected by links to one or more other nodes creating a network in which any one of the n nodes may have up to n connections (i.e. a connection to every other node in the network as well as a connection to itself). The connections allow a node to communicate to the other nodes to which it is connected in an excitatory or inhibitory way. The links have a strength or weight property that causes the effect of the signal from a node to be modified depending on the strength (and sign) associated with the link. The adjustment of these strengths is the basis for learning or self-organization.

ANNs can be classified in a variety of ways. There is the division between networks that learn under supervision and those that do not require supervised learning. The number of layers can vary. There are feedforward and feedback networks. Finally, there are many learning rules or algorithms that are employed (Simpson, 1990).

Supervised and unsupervised learning

Networks that learn under supervision are trained by a process of presentation of patterns and their associated correct or target responses. A pattern is presented to the network which, after some processing, produces an output which is compared to the target or target output. The strengths of the links between the nodes is then adjusted according to one of many learning rules with the objective being a reduction in the error between the generated and desired outputs. Another pattern is presented, processed and the resulting error used to modify the links again. This is generally repeated many times until the network produces the desired outputs for every input pattern.

Networks that learn without supervision are allowed to discover the patterns inherent in the features presented to them. Repeated presentation of feature patterns causes the links between nodes corresponding to features that occur together to become strengthened and the links between nodes which do not co-occur to become weakened, a process known as self-organization. Thus, after some number of presentations the network can be probed to determine, for instance, which features are related, and how closely. This probing involves activating a node and allowing the activation to spread without self-organizing or learning taking place. The level of activation of the other nodes are indicative of their degree of association to the probe. Alternatively, calculating eigenvectors of the network matrix and using these to plot the nodes in some representation of n-dimensional space can be useful. An example of this kind of network is the self-organizing map (SOM). This type of network can also be used, after self-organization, to reconstruct output from partial or noisy input (Abdi, 1999, p. 22), a task to which they are widely applied in the field of signal processing.

Network layers

The nodes of a network are organized into layers, commonly an input, hidden and output layer. The nodes of one layer communicate with (or are connected to) nodes of the adjacent layer. Nodes that communicate with the outside world are input and/or output nodes. This type of network is a multilayer network most often instantiated as a multilayer perceptron (MLP). It has been shown that it is not necessary to have a large number of such hidden layers to solve most problems (Garson, 1998, p. 28). An exception to this (multilayer) model is any of the variations on the Hopfield net or Associative Memory models which typically use only one layer as both the input and output layer. In this case each node is connected to every other node, including itself (Abdi, 1999; Jordan, 1997). The type of ANN used in this study contains only one type of node which serves to accept and process input, produce output.

Feedforward and Feedback networks

Multiple layer neural networks can be classified by the direction in which the signals propagate in the network.[1] In the more common feedforward network, the input signals are presented to the input layer which, after some processing, sends signals to the next layer, and so on. In one of the most common forms of feedforward network the error (i.e. the difference between the output of the final layer and the target output) is fed to preceding layers in a process called backpropagation of error, hence the name *backpropagation network*. In the backpropagation network it is the error that is fed backwards through the layers as part of the weight adjustment process. The signals are fed forward through the network, thus the backpropagation network is a feed forward network. Examples of the feedback network include recurrent and Hopfield networks

[1] Single layer networks are generally fully connected, in other words, each node is connected to every other node.

(Garson, 1998). In these models signals flow in both directions resulting in systems that are less likely to stabilize, are less sensitive to new input and may develop resonance. Feedforward networks are the more commonly used model (Garson, 1998).

Learning rules

Many learning rules have been derived from the original proposed by Donald Hebb in 1949 (Gurney, 1994). The basic principle behind them all is that the connections between output nodes that should be activated and the hidden nodes activating them should be strengthened while those between output nodes that should not be activated and those hidden nodes that are activating them should be weakened. The variations in rules involve the inclusion of momentum factors, learning rates, and decay (forgetting) rates, among others. There are a variety of learning rules used to modify the strengths of the links between neurons during training or learning. One of the simplest and oldest is known as the Hebbian rule, named after its developer, Donald Hebb (Klerfors, 1998). The Hebbian rule simply causes the strength of any activated connection to be increased if both neurons are active. Modified Hebbian learning adds a learning rate which modulates the amount by which the links are strengthened or weakened. A popular rule for supervised learning is the Delta rule which seeks to minimize the mean squared error in the network and is therefore sometimes known as the least squares rule. In order to implement a Delta rule the error has to be known. This implies training on a desired outcome and is therefore limited to supervised learning models. Teuvo Kohonen developed a method whereby the network trains itself through a competitive strengthening scheme (Garson, 1998, p. 5). This depends on the determination of a "winner" node which is that node which has the highest activation. Only weights of the connections to the

winner node (and in some variations its neighbors) are modified. This type of network is the basis of the Self Organizing Map (SOM) and the Interactive Activation and Competition (IAC) model (McClelland, 1981).

Interactive activation and competition networks

This study uses a type of network known as an Interactive Activation and Competition (IAC) network in which connections between the nodes in the network take on values that can then be used to perform associative recall (McClelland, 1981; Rumelhart, 1986). IAC networks can be used to find associations in a text. A familiar example is the commercial program Catpac™ which analyzes text and produces graphical output in the form of a dendrogram that the researcher can then examine for meaningful patterns, clusters, and the like (Woelfel & Styanoff, 1993). Catpac can produce output based on the eigenvalues of the network matrix which can be used by other programs, such as ThoughtView, to visualize the relationships in a variety of ways (Woelfel & Styanoff, 1993). The IAC used in this work is fully recurrent which is to say every node is connected to every other node (Jordan, 1997). After self-organization takes place the connection values can be thought of as similarities, strengths, distances, or other measures of relatedness. They may be excitatory or inhibitory. The connection values are refined by a process of repetitive coactivation of nodes by the co-occurences of the features that these nodes represent in the input under analysis. Nodes are coactivated when their associated features appear together within a region called the scanning window which is defined by the parameters of the network. Figure 3 illustrates this process in a very simplified way. The input (in this case text) is processed by examining all of the features (words) in the scanning window (in this case a sentence). The nodes whose features appear in the window (in this case: “I”, “like”, and “Ike”) are externally activated. Their

external activations cause them to fire which is to say send a unit signal to all connected nodes (in this case all nodes). These signals are increased (or decreased) according to the values of the links that they traverse. An activation function is applied to the sum of the signals received by a node and the result compared to a threshold value. If the result is greater than the threshold then the receiving node itself fires causing the activation to spread. This is illustrated in the step labeled "Spread Activation" (Figure 3) in which none of the non-externally activated nodes fire. When the spreading activation is complete the learning rule is applied causing the connection strengths to be modified. In Figure 3 -"Apply Learning Rule" two links have been strengthened as shown by the heavy lines. Next, all external activations are typically set to zero, the scanning window moved ahead in the input stream and the process repeated. The number of features that the window is advanced is determined by an adjustable parameter. This study used the length of the current sentence for the width of the window. After a number of cycles of moving the window, coactivating the nodes, spreading the activations and, applying the learning rule, the connections begin to assume stable values which can then be used to recall associations expressed in the text. When the end of the input (text) is reached the network used in this research stores the values of the links, the activation level of each node, the terms associated with the nodes and other parameters for future use.

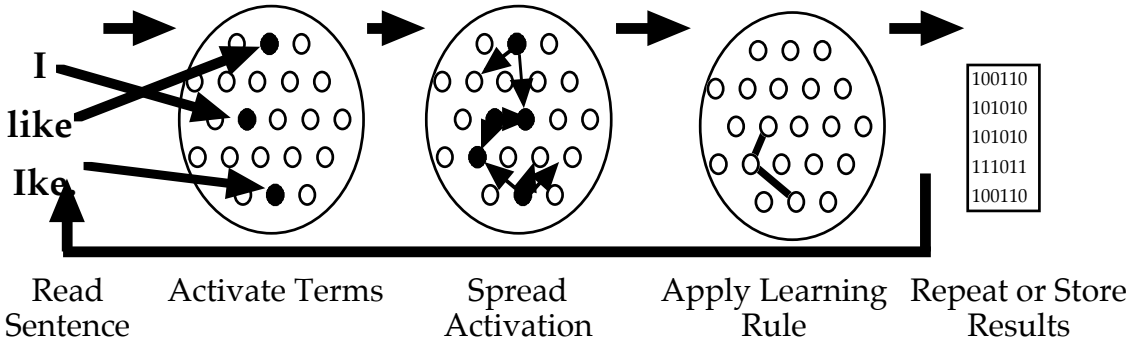


Figure 3 – Basic operation of a single layer ANN during self-organization

After many patterns of input have been processed associations can be recalled. Figure 4 shows the recall process, again in a very simplified way. Recall involves probing the network by exciting one or more nodes (the probes) and allowing the activation to spread as described above. This will cause all nodes to become activated to the degree to which they have become associated with the probe(s). In this case one probe term's node is activated causing it to fire. Its signal traverses every link, in some cases being augmented and in other cases attenuated depending on the link values. To reduce clutter the example in Figure 4 only shows two strong links (solid arrows) and two weaker links (dashed arrows). It should be remembered that each node signals all of the other nodes in the fully recurrent model used in this research. The end result, in this case, is that two additional nodes have become activated due to their strong (solid line) links to the probe node. During recall the learning step is not performed. Thorough reviews of the development of artificial neural networks can be found in Garson (1998) and Simpson (1990).

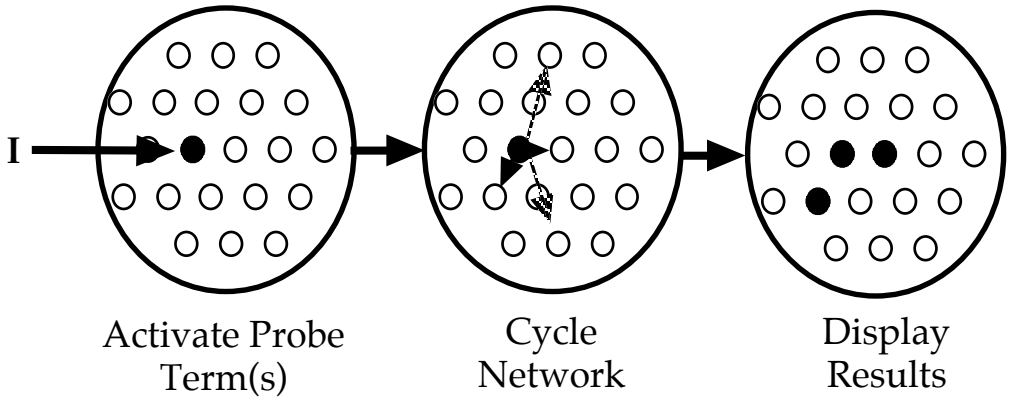


Figure 4 – Probing a single layer ANN

Internet communication

This research focuses on technology mediated communication (TMC) which takes many forms. This section briefly describes them in the approximate chronological order of their implementation. Some, like the telephone and email, predate the Internet being originally implemented on point-to-point analog networks or within single time-sharing (so called "mainframe") systems and later on circuit switched and local area networks (LANs).

Pre-computer technologies

Early forms of TMC enabled the written word (sent via postal service, pneumatic tube, and, telegraph, or, in books, magazines, and, newspapers) and the spoken word (via telephone, radio, television, recordings, and motion pictures) to be used to communicate synchronously or asynchronously in one-way or interactive modes. The telephone is perhaps the most widely used interactive, synchronous communication technology to date, at least for direct person-to-person (as opposed to automated) communication.

Email

By 1998 email had already become the most heavily used TMC technology

among Internet users surpassing even telephone use (Georgia Tech Research Corp., 1998). Similar results were found in 2001 by Cole (2001). Email has the advantages that the communication can be private, secure and anonymous. Although complete privacy and security are not automatically guaranteed, current encryption technologies allow a high degree of protection from snooping. Likewise, there are a number of technologies that can be employed to verify the source and authenticity of email messages. Anonymity, although not impossible to achieve, is more difficult in email than other forms of internet communication. There are a number of remailers that will strip identifying information from an email header before forwarding the message to the recipients. Email communication can be one-to-one or one-to-many. A one-to-many communication may appear to the recipients as a one-to-one message because the identities of and even the fact that there are additional recipients can be hidden from the recipients.

Very little private email is archived (Lyman, 2000). Mail distributed to listserv lists, on the other hand, is and has been archived in some cases for quite some time making it a good candidate for study. List email is also generally public communication in the sense that the users know that their messages are distributed to a group of unknown composition (Sierpe, 2000). This does not apply to lists whose membership is controlled by a moderator or owner, but many lists allow unrestricted subscriptions.

Usenet News

Usenet news ("news groups" or simply "news" as it is frequently known) was started at Duke University in 1979 (Seidenberg, 1995) and has achieved truly global proportions since then. It developed as an alternative to email to handle widespread distribution of messages without the storage and bandwidth

requirements that email distribution imposes. Usenet news messages are centrally stored (on nntp servers) and jointly accessed by many users using nntp client software (news readers). Individuals can access messages at will without having to store their own copies of the messages, as would be the case with email messages. Therefore, each user does not have his own copy of each message, as is the case with email distributed messages. Most news clients provide sophisticated management capabilities that allow messages to be sorted by date, sender, topic (thread) and so forth. Although some newsgroups are archived indefinitely, most purge messages after a period of time determined by the group owner. Newsgroups can be configured to allow unrestricted posting of messages by anonymous users, approved posting of messages through a moderator, or other variations of restricted posting. Most news groups are duplicated (mirrored) by secondary servers to spread out the network traffic and reduce the load on the primary server. The relatively long history of news groups and the fact that many of them are archived makes this mode of Internet communication another good candidate for this type of research.

Gopher

When the Internet was still very young (1991) workers at the University of Minnesota developed the first widely used hypertext protocol for the Internet – Gopher (Zakon, 2001). The gopher protocol implemented the concept of hyperlinks first suggested by Vannevar Bush (1945) and Ted Nelson (1992) as a means for connecting related information in discrete documents for ease of retrieval. This technology was soon supplanted by the more versatile World Wide Web.

World Wide Web

The World Wide Web, created in 1990 by Tim Berners-Lee to serve the scientific community (Cailliau, 1995) has been the most successful technology to date in harnessing the power of the Internet for lay users. It was overshadowed at first by the Gopher protocol, but soon replaced Gopher as the hyperlinking protocol of choice. The flexibility and adaptability of the Web and its underlying standards (http, ftp, html, cgi) has allowed applications to be developed that have far exceeded the capabilities originally conceived by its inventor.

Real-time chat

There are two types of real-time internet enabled "conversations" – one-to-one and many-to-many. One-to-one chat protocols include Internet Relay Chat (IRC), America Online Instant Messaging (AIM), and others. They have their roots in console message sending facilities of early time-share operating systems. IRC was developed to allow real-time communication between internet users (Caraballo, 2000). Users can communicate with each other (individually or in groups) over what is known as channels. A channel may be open to all, or require subscription and approval of one or more of its owners to join. This concept has been adapted in many ways to include features such as nickname registration, automatic notification when an individual of interest (buddy) is online, and avatars (cartoon-like characters which represent their owner and are often able to display facial expressions and/or body motions), sound, and video. They are generally theme- or topic-based, although it is a common occurrence for people to say things in a chat room that are off topic. There is also a tendency by participants to resort to name calling and "flaming" rather than rational debate, and the same ideas tend to be expressed repeatedly. Most participants of chat rooms use "screen names" rather than their real names, resulting in a high

degree of anonymity. As Peter Steiner so aptly put it in his now famous cartoon, "On the Internet, nobody knows you're a dog." (Steiner, 1993) IRC communication is hosted on central servers making it possible to archive the text of the discussions. These archives are potential sources of large amounts of data for research.

Streaming media

The Internet also supports streaming media such as audio and video content that is meant to be experienced in real time like radio and television. Streaming audio can be live or prerecorded. Although it is widely anticipated that the Internet will become the primary distribution mechanism of all of our media in the not-too-distant future, the limited bandwidth of most connections to home users today prevents high quality audio and video from being widely streamed (Kennard, 1999). A 56Kbps modem connection will support FM quality sound and 320 x 240 pixel, 1 to 5 frames per second color video. Already it is commonplace to listen to a radio station's Internet stream far outside its normal coverage area. Streaming video is also starting to become more common as an increasing percentage of the population secures the bandwidth required to receive it. Much of the video demand and development is driven by soft and hard pornography, much as the adoption of home video technology and image scanners was a few years ago (Liddell, 2001). A common type of consumer level video stream is produced by a simple digital video camera connected to a web server – a so-called "web cam." Some chat rooms now support web cams.

Reliable high quality streaming video will not become widespread until two problems are overcome. One is the Internet's lack of quality of service (QoS) protocols which is the telecommunications industry's term for guaranteed bandwidth. All users on any particular subnetwork of the Internet share the

available bandwidth equally on a collision detection and recovery scheme. The more users there are on a subnet, the less bandwidth each has. Traffic that requires continuous, uninterrupted flow, like video, is not prioritized over traffic that can be transmitted in interrupted bursts, like email and file transfers, because such a prioritization scheme was not part of the original Internet protocol specifications. The second problem which prevents widespread adoption of streaming video is the lack of high bandwidth connections to the end-users – the so-called "last mile" problem. As more end users get high bandwidth connections, such as ADSL/DSL, cable modem, satellite and even fiber to the home, this problem will disappear, provided the local internet service providers themselves increase the bandwidth of their connections to keep pace with demand. Because streaming media does not rely on text as I have defined it, I will not consider it further in this work. Due to the well established nature of email lists, their generally well defined and narrow topical focus, the existence of archives, and the serial nature of communication flow, they will provide the data for this study.

The remainder of this dissertation deals with text streams carried on the Internet, and their automatic processing by artificial neural network software. To that end, this dissertation will investigate the modification of an interactive activation and competition neural network algorithm to include the ability to add terms to the network as they are presented by new texts for the purpose of continuous text analysis. Two models of term external activation will also be investigated to explore one alternative to using a stop word list to further enhance the automaticity of the system. In one treatment the external activations of all nodes will be set to zero after each learning cycle. In the other treatment the external activation of nodes that have appeared in the current message will be reduced by a constant factor after each learning cycle.

Chapter 3
Methodology

This chapter describes in detail the methodology used in this study. In summary, artificial neural network software based on the interactive activation and competition model was adapted to process data continuously from email messages as they arrived. The size of the context unit was varied and the results obtained thereby compared.

Overview

This research applies neural network analysis techniques to bodies of text generated over a period of time by Internet discussions for the purpose of making accessible the relationships between terms in the text. That is, once the text stream analysis has proceeded for a period of time the various terms and their relationships will have been encoded in (or learned by) the network. The terms are effectively mapped into a multidimensional space. As additional text is processed, additional terms appear which are added to the network. This, of course, causes the relationships between terms to continuously undergo slight changes causing changes in the network. These changes can be monitored. The network thus provides an up-to-date representation of the terms and their relationships. These relationships are revealed in the activation level generated in the term nodes when one or more terms is externally activated (used as a probe term). Any term can be used as a probe term to see its relationship to the other terms. Multiple terms can be simultaneously presented to the network as probes. The key question that this system answers at any given time is, how far apart are terms of interest? The distance metric provided is non-parametic, that is, it is simply an abstract measure of closeness. It does not imply similarity or direction

along any particular cognitive dimension. It reflects the degree or amount of association between the terms as they appear in the text analyzed so far.

Description of the network and definition of terms

This section describes the network used in this study and provides definitions for the terms as used in this work. The definitions appear alphabetically in Appendix A.

This study explores a method of analyzing *text* which is defined as *sentences* taken from email messages sent to the Open Lib/Info Sci Education Forum (jesse@listserv.utk.edu) which were selected by subject line. A *sentence* is a sequence of *terms* or individual words of three or more characters (in the ASCII character set). Punctuation marks and spaces mark the boundaries of words. They are the recording units of this research. No attempt is made to disambiguate terms nor to deal with hyphenations or contractions (e.g. "I'll"). A sentence starts at the beginning of an email message, or the end of the previous sentence and is terminated by: A) a punctuation stop (a period, exclamation point, question mark, etc.) or; B) the end of the message. *Email headers* (added to the message by email software to facilitate routing and delivery of the message) and *quoted text* (that authors copy from previous messages and include in their own messages to clarify context) are not included.

The *network* is a collection or matrix of *nodes* each of which represents a term that has been encountered in the text being processed. At any given time each node has two *activation levels* which are real numbers. *External activations* are assigned to terms that occur in the sentence currently being processed. They are the input to the network. *Computed activations* are calculated by the software as described

below. In this network model each node is connected to all of the others nodes, including itself. These connections, called *links,* provide the means for determining the magnitude of the effect one node has on another. When a node's level of activation is sufficiently great, that node will *fire* sending a signal through its links to all other nodes. The strength of the signal received by a node through a link is determined by the product of the value of the signal applied and the value stored in the link. Each link, therefore, has a value or *weight* which modifies the signal sent through it. The links allow the nodes to activate each other according to the parameters and algorithms in effect. Positive link values are excitatory and negative values are inhibitory.

The network has two types of operation, self-organizing and probing. The first is used when the network is "learning" about the texts. The second is used to discover what the network has "learned." During self-organizing, when the activation cycle described above is complete (at the end of each sentence) the learning cycle takes place during which a modified *Hebbian learning rule* is implemented to adjust the weights of the connections. This rule reinforces connections between nodes which are externally excited, thereby strengthening their associations. The modification used in this study is the inclusion of a learning rate – *h*. The modified Hebbian learning rule is:

$$Win_{i,j} = Win_{i_{t-1},j_{t-1}} + h((Av_{i_{t-1}} - \bar{x})(Av_{j_{t-1}} - \bar{x}))$$

where

$Win_{i,j}$	=	the connection weight between nodes i and j at the end of a cycle
t-1	=	the end of the previous cycle
Av	=	the activation level of a node
$\bar{x}$	=	the average activation
h	=	the learning rate.

After learning takes place, the external excitations of the nodes are reduced either to zero or by a decay function (which in this case was a linear reduction by *rf*) depending on the context treatment. This study compares the results obtained using two different context treatments. The first context treatment is the sentence and the second is the message. When the context treatment is the sentence, the external activations are reduced to zero after each learning cycle. When the context treatment is the message the external activations are reduced by subtracting rf (with a lower limit of zero). Finally, the network connection weights are normalized. The normalization of the weight matrix is an important step in this model. Centering the matrix on zero and limiting the values to the range of -1 to +1 is accomplished by subtracting the average value of all link weights from each link weight and then dividing each link weight by the absolute value of the largest. This results in a matrix of link weights whose average value is zero, whose minimum value is -1, and whose maximum value is +1. In future cycles, those links with negative weights will be inhibitory while those with positive weights will be excitatory. Finally, the new network state is saved in the two files mentioned above (`.lbl` and `.net`) for the processing of a future message.

Self-organizing mode

The network has two modes of operation, self-organizing and probing. In the *self-organizing* mode the software loads an existing network into memory from two files, one of which contains the n terms already in the network (the `.lbl` file) and the other which contains an n x n matrix of the network connection strengths (the `.net` file).[2] The software next reads a `.dat` file containing sentences from a new email message. The `.dat` file is read line by line. Lines are equivalent to records, i.e. they are delimited by the operating system line ending

[2] For the first message this step is, of course, skipped as there is no pre-existing network.

token, which in the case of Darwin (the operating system used in this study) is a carriage return/linefeed pair (ASCII 13 followed by ASCII 10). To simplify programming, lines were also limited to 1024 characters. This limitation was never reached as the email server through which each message was received automatically inserts an end-of-record character at white spaces at most every 80 characters. Next, each line is parsed (by the neural network program) into terms based on white space and punctuation. Each term thus encountered is checked against existing terms in the network. If the term is not already in the network it is added to the array of terms and a column and a row are added to the matrix of links. The term just encountered is given an external activation (ext_i) of 1 (Appendix B) . This process is repeated until punctuation marking the end of a sentence or the end of the message itself is found. When the end of a sentence (or the end of the message itself) is reached an activation cycle is initiated. During the activation cycle the activation level of each node ($act_{(i,2)}$) is calculated using the node's current activation, the external activation levels of the other nodes, the connection strengths and an activation function. A standard sigmoid (-1 to +1) activation function is used. Its mathematical form is given in Equation 1. In this equation act_t is the new activation of the node after the cycle; act_{t-1} is the activation level calculated by the previous cycle; *rf* is the decay rate which is a factor that is subtracted from all link strengths at the end of each learning cycle and *anet* is the sum of the current activations multiplied by their associated link strengths (*win*), given in Equation 2.

$$act_t = \frac{1 - e^{-anet}}{\left(1 + e^{-anet}\right) + (rf \bullet act_{t-1})}$$

Equation 1. The sigmoid activation function

$$anet = \sum_{i=1}^{n} \sum_{j=1}^{n} act_{t-1}^{i} \bullet win_{i,j}$$

Equation 2. The anet value

Probe Mode

The probe mode of operation is the same as the self-organizing mode with the following exceptions. After the network is initialized as described above the program sets the external activation(s) of the *probe term*(s) (selected by the researcher and passed as command-line parameters) to +1. There is no `.dat` file to read in probe mode. The network then cycles spreading the activation via the connections and the sigmoid activation function. The nodes are then sorted by activation level and this sorted list is sent to `stdout`. A parameter controls whether the program outputs a list of all terms in the network, or only those terms falling between user-specified or default lower threshold and upper cutoff values. In either case, the nodes are output with their activation levels. The program defaults (threshold=.0001, cutoff=1.001) can be overridden by parameters passed in the command-line or via a text file. For the purposes of this experiment this output was largely ignored except to monitor the processing. A tab-delimited file (See Appendix C) of all nodes and their activations is also saved and was the data that was analyzed in this experiment. The learning step is skipped in the probe mode so there is no change in the network connection strengths, though for simplicity the program saved the (unchanged) network and label files after every execution.

Software adaptation

The ANN used in this research is of the Interactive Activation with Competition (IAC) type (McClelland, 1981). In order to perform the experiment presented in this research the core engine of a commercially available IAC was adapted. The

software, originally written in FORTRAN77 as an interactive command-line driven program to run under MS-DOS, was modified to run under the Darwin operating system (a Unix derivative) as a cgi.[3] This required modifications to the input routines which originally were designed to display prompts on the screen and accept interactive keyboard input. These routines were combined into an initialization routine that reads the parameters from a text file. The parameters file can be reused, as was the case in the self-organizing runs or generated on the fly by a cgi script, as was done during the probes. In addition to the original parameters of threshold, learning rate, decay rate, activation function, input data and network file names and output file names the following parameters were added – *cutoff* (a real number), and *context* treatment (sentence or message).

An important addition to the original program was the ability to add terms to an existing network. Each time a message was processed, previously unencountered terms were added. At the end of the process the system (optionally) wrote the current network matrix and labels to files. These files could then be used as the basis of the network to analyze the next message. A limited amount of filtering was done during message processing on a line-by-line basis. This was an attempt to reduce the processing of unwanted lines, such as email headers and quoted text. This proved unnecessary, however, as the separate program that created the message text files automatically removed headers. The lack of standardization in formatting quoted text, necessitated manual editing of the messages before extraction and processing to eliminate quoted text.

Data collection

Archives of Internet Mailing Lists

Messages posted to the topical internet mailing list, Open Lib/Info Sci Education

[3] Compiling was performed by Absoft Pro FORTRAN 7.0 for Mac OS X. (Absoft Corp., 2001)

Forum (jesse@listserv.utk.edu) henceforth known as the Jesse list, were used in this study. The Jesse list is the official electronic channel of communication for the Association for Library and Information Science Education (ALISE) and "promotes discussion of library and information science education issues in a worldwide context." (Whitney, 2002) This list was chosen because of it familiarity to the researcher and the diversity of topics covered by it. The Jesse list had 955 subscribers at the time of the study. (L-Soft, 2002) The messages were obtained by subscribing to the list under an email account set up for this purpose. Each message was edited by hand, as described above, and saved as a text file without header lines. All messages in a thread were also appended to a single text file for the purpose of generating descriptive statistics. Two streams of messages were used. The first stream consisted of messages concerning the use of web citations in evaluation of scholarship (henceforth, the *web-citation thread*.) The second stream consisted of messages concerning the writing abilities of students (henceforth the *writing thread*.) Members of these two threads were selected by inspection of the subject lines of incoming messages. Messages whose subject line indicated membership in one of the two threads were processed. The web-citation thread consisted of 23 messages. Overall this amounted to 3214 terms of which 1121 were unique. The longest message had 387 terms and the shortest 21. The writing thread consisted of 26 messages. There were 3055 terms of which 1262 were unique. The longest message contained 329 terms and the shortest 17. For reproducibility the system simultaneously processed each message under both treatments. For the sake of processing efficiency the messages were first manually edited to remove quoted text, headers, signature lines, and decorative non-text elements, such as emoticons. Initial tests revealed that lines beginning with the asterisk, underscore and hyphen characters are almost universally used as decorative elements so they were also eliminated. A

typical message (with identifying material obscured) from one of the threads is shown in Appendix D in both its original and edited forms. A script (Appendix E) written in AppleScript™ (Apple Computer, 2001) controlling the commercial email client, PowerMail 3.4.2 (Christe, 2001) saved the body of each selected message to a text file named `newmsg.dat` (overwriting any existing `newmsg.dat` file) and appended the body to another text file named `allmsg.dat`. After extracting a message the script executed the neural network program passing it the appropriate parameters via text files that had been manually created. The script executed the neural network program as a terminal process twice for each file – once each for two parameter files creating the two test treatments. Finally it changed the status of the message in the email client from "unread" to "read" to enable the researcher to know which messages had been processed. All software ran under the Mac OS X environment. The neural network program ran as a terminal program under Darwin allowing it to be used as a CGI program during the analysis phase as well.

Network manipulation

The parameter that was manipulated was the context unit size. Two variations of the context unit size were used on two sets of messages as described above. For one condition the context unit size was a sentence, as defined above. This condition will be called the *sentence treatment*. For the second condition (the *message treatment*) the context unit size was the entire message. In both treatments the external activations of all nodes were set to zero at the begining of each message. Under the sentence treatment, after the terms in a sentence were subjected to a learning cycle, the external activations of all terms was set to 0. Under the message treatment the external activations of all terms were reduced incrementally (rather than to zero) by subtracting rf after each sentence in the

message was processed. Because the external activation of every term is set to 0 at the beginning of each message, only terms that have appeared in the current message have an activation level greater than zero. The reduction factor (rf) was held constant (at .05, the default for the software that was modified for this experiment) for all messages. So, for example, in a message with n sentences, the non-recurring terms in the first sentence will have been processed with an external activation level of 1 (when the first sentence itself was being processed), a level of 1-rf during the processing of the second sentence, and so forth. until the nth sentence was processed at which time the external activation level of these terms would have been $1 - n \bullet rf$. The activation level was not permitted to go below zero. These two treatments are compared graphically in Figure 5.

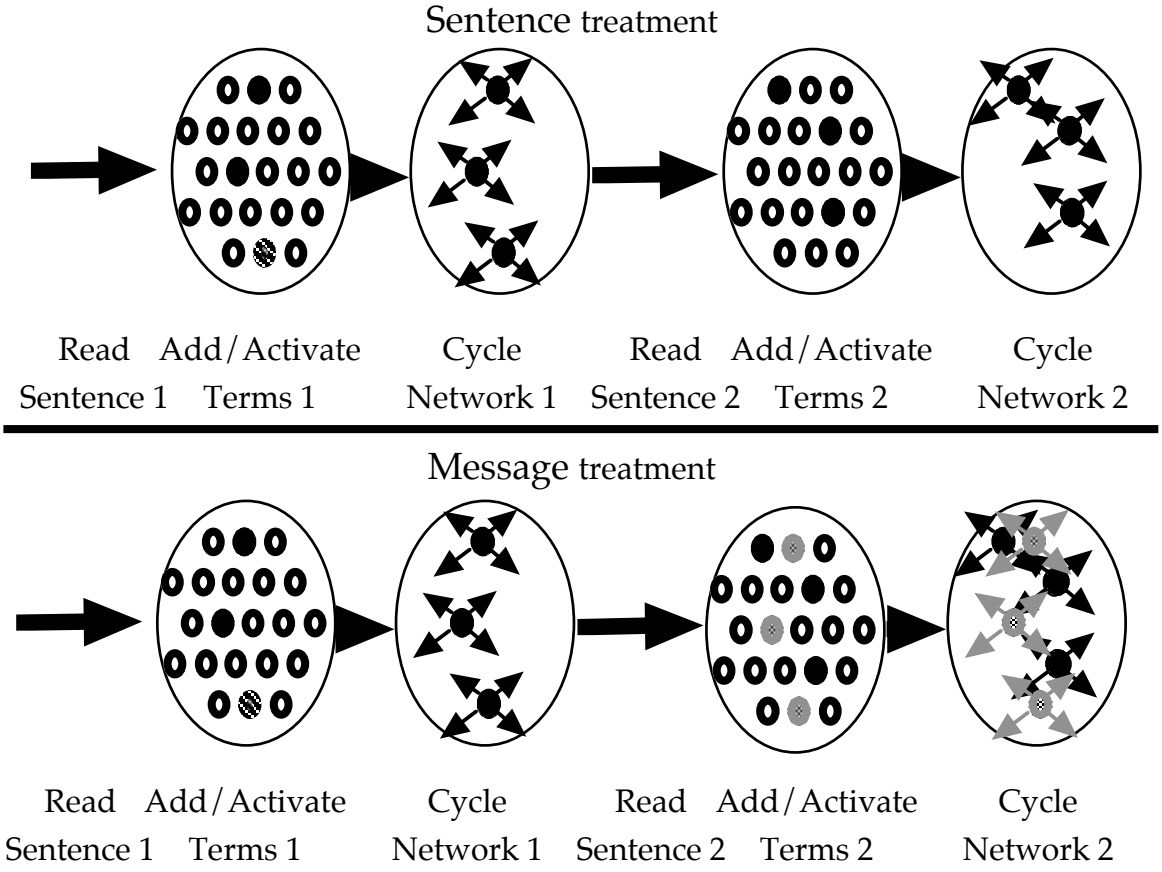


Figure 5 – The two treatments of self-organizing:

Referring to Figure 5, the sentence treatment is depicted at the top of the figure and the message treatment at the bottom. The large arrows indicate direction of flow, that is, the first action is at the left and the last action is at the right. Each diagram shows two cycles of the network stretched out in linear form for readability. Both networks are reading the same data. The small elliptical shapes are the nodes. The large ellipses are the network. At the beginning of a message (far left) the operation of the network in both treatments is identical. A sentence is read and the nodes corresponding to the terms contained therein are activated, as indicated by filled shapes. If a term has not been encountered then a node is added for it and that node is activated as well. When the end of the sentence (containing in this case three terms) is reached the network is cycled. In order to show signals emanating from the activated nodes the other nodes are not shown in the "Cycle Network" steps. In both treatments there are three nodes (corresponding to the three terms in the first sentence) firing in the first cycle. After the learning cycle takes place the networks behave differently. Under the sentence treatment all external activations are set to zero. Under the message treatment the external activation level of the activated nodes is decreased by the decay function, which subtracts the decay factor from the activation level. Therefore, after reading sentence two and activating the nodes for the three terms in sentence two, the network operating under sentence treatment has three activated nodes while the network operating under message treatment has six nodes activated. The nodes whose activations have been decreased by the decay function are shown in gray to indicate that they are still activated, though to a lesser degree than nodes representing terms encountered in the current sentence. In anthropomorphic terms, under the sentence treatment the network only pays attention to the words in the current sentence. Under message treatment the network remembers the words in previous sentences of the current message. As

the message is processed, sentence by sentence, the network pays less attention to words appearing earlier, unless, of course, they reoccur in later sentences.

For each thread a file containing all of the messages was created for the purpose of generating descriptive statistics such as the number of unique words and word frequencies. K*ey terms* were assigned to each thread as described below. Each of the two *threads* (web citation and writing) was processed using the two context *treatments* (sentence and message) producing four networks. When a term is used to probe a network, it is known as a *probe term* and the resulting output is a *case* . The file of primary interest for a case is the `.tab` file which contains a list of all of the nodes, in network entry order, with the activation levels produced by the probe (Appendix C). Although the system allows multiple simultaneous probe terms, this experiment used just one probe term for each probe cycle. All of the key terms were used individually as probe terms with their respective networks producing two case files per key term (one for each treatment) for a total of 22 case files, 12 for the web citation thread (2 treatments x 6 key terms) and 10 for the writing thread (2 treatments x 5 key terms). The IAC system also produced two other files for each case: a `.net` file which contains a matrix of connection strengths representing the network itself, and a `.lbl` file containing the terms in the order they were added to the network.

For the purpose of choosing key terms, both sets of messages were analyzed by a professor who teaches indexing in an accredited library science program and by the commercial software program Catpac (Woelfel, 1993). The indexer extracted 17 key terms that she felt represented the overall content of the web-citation thread while Catpac extracted 16 (Table 1). Appendix F contains the frequencies and dendrograms created by Catpac for the two threads. Of these terms six of

those identified by the indexer also appeared in the main cluster created by Catpac (along with their plurals.) These terms are shown in italics in Table 1. In the writing thread the indexer identified ten terms and Catpac clustered nine. Between these two groups there was agreement on five terms which are shown in italics in Table 2. The following "key terms" were chosen and used as probes and indicators of performance: for the web-citation thread – *citation, counts, engine, links, search,* and *web*; for the writing thread – *English, papers, student, students,* and *write*. Thus, the web-citation experiment consists of six pairs of cases and the writing experiment consists of five.

Indexer Terms	**Catpac Cluster**
activity	*citation*
citation	*citations*
counts	*counts*
databases	*engines*
engine	etc
evaluation	*link*
intellectual	*links*
links	mentions
print	pages
promotion	question
publication	research
refereed	results
search	*search*
syllabi	*web*
teaching	will
tenure	work
web	

Table 1 - Indexer and Catpac terms for web-citation thread.

Indexer Terms	**Catpac Cluster**
center/s	*English*
content	level
course/s	*paper*
English	*papers*
format	*student*
papers	*students*
scholarly	two
skills	work
standards	*write*
student/s	years
write	

Table 2 - Indexer and Catpac Terms for writing thread

Chapter 4
Results

Presentation

The two sets of messages will be known as the *web-citation thread* and the *writing thread*.

Summary statistics

The web-citation thread consisted of 23 messages. Overall this amounted to 3214 terms of which 1121 were unique. The longest message had 387 terms and the shortest 21. The writing thread consisted of 26 messages. There were 3055 terms of which 1262 were unique. The longest message contained 329 terms and the shortest 17. In the 23 web-citation messages the most frequently used term was "the" occurring more than twice as often (198 times) as the next most frequent term ("and", 98 occurrences). The next most frequently occurring words, in order of frequency, were "web", "that", "for", "research" and, "this". In the writing thread messages "the" was also the most frequently occurring term followed by "and", "that", "writing", "students", "for" and, "not". From the case files the key terms and their activation levels under both treatments and for each probe were compiled and sorted alphabetically by key term. The similarity of these vectors was measured by calculating a correlation coefficient for each pair (one for each treatment) of activation levels generated for each probe term. A correlation coefficient was also calculated for the entire network for each probe term. The correlations for the entire network were low yet those for the key terms were very high especially for the web citation thread networks (tables 3 and 4).

Sentence Treatment						
	Probe Terms					
Key Terms	citation	counts	engine	links	search	web
citation	1.00000	0.04108	0.00746	0.02428	0.03833	0.14031
counts	0.04108	1.00000	0.00629	0.02147	0.02480	0.05358
engine	0.00748	0.00629	1.00000	0.00624	0.02363	0.01274
links	0.02428	0.02147	0.00624	1.00000	0.01835	0.04189
search	0.03833	0.02480	0.02352	0.01835	1.00000	0.07306
web	0.14031	0.05358	0.01272	0.04189	0.07306	1.00000
Message Treatment						
	Probe Terms					
Key Terms	citation	counts	engine	links	search	web
citation	1.00000	0.31323	0.28720	0.30207	0.31313	0.30780
counts	0.31323	1.00000	0.33083	0.33823	0.36071	0.33987
engine	0.28720	0.33083	1.00000	0.30844	0.32788	0.30911
links	0.30207	0.33823	0.30844	1.00000	0.33915	0.32293
search	0.31313	0.36071	0.32788	0.33915	1.00000	0.34091
web	0.30780	0.33987	0.30911	0.32293	0.34091	1.00000
Correlations						
All Terms	0.18198	0.13900	0.11763	0.14172	0.16286	0.21149
Key Terms	0.99394	0.99721	0.99866	0.99831	0.99687	0.99160

Table 3 – Key term activation levels and correlations: Web citation network

Activation spreads for all terms (total spread) and for key terms (term spread) were calculated for all cases (tables 5 & 6[4]). These values represent the dispersion of the activation levels and the difference between the highest and lowest activation level for the terms in question. Tables 5 & 6 also show the rank position (by activation level) of the least highly activated key term (LAKT) for each case.

[4] Rounding off of numbers for display purposes may result in some apparent irregularities.

Sentence Mode

Probe Terms

Key Terms	english	papers	student	students	write
english	1.00000	0.01125	0.01532	0.06455	0.01815
papers	0.01125	1.00000	0.02097	0.03765	0.01170
student	0.01532	0.02097	1.00000	0.04255	0.01249
students	0.06455	0.03765	0.04255	1.00000	0.06093
write	0.01815	0.01170	0.01249	0.06093	1.00000

Message treatment

Probe Terms

Key Terms	english	papers	student	students	write
english	1.00000	-0.26056	0.24663	0.26835	0.25939
papers	-0.26056	1.00000	-0.25676	-0.26584	-0.25598
student	0.24663	-0.25676	1.00000	0.23409	0.27516
students	0.26835	-0.26584	0.23409	1.00000	0.24911
write	0.25939	-0.25598	0.27516	0.24911	1.00000

Correlations

All Terms	0.16919	0.14867	0.20466	0.32644	0.16684
Key Terms	0.87653	0.99955	0.87117	0.87824	0.87364

Table 4 – Key term activation levels and correlations: Writing network

Probe Term	citation		counts		engine		links		search		web	
Key Terms	sentence	message	sentence	message	sentence	message	sentence	message	sentence	message	sentence	message
citation	1.00000	1.00000	0.04108	0.31323	0.00746	0.28720	0.02428	0.30207	0.03833	0.31313	0.14031	0.30780
counts	0.04108	0.31323	1.00000	1.00000	0.00629	0.33083	0.02147	0.33823	0.02480	0.36071	0.05358	0.33987
engine	0.00748	0.28720	0.00629	0.33083	1.00000	1.00000	0.00624	0.30844	0.02363	0.32788	0.01274	0.30911
links	0.02428	0.30207	0.02147	0.33823	0.00624	0.30844	1.00000	1.00000	0.01835	0.33915	0.04189	0.32293
search	0.03833	0.31313	0.02480	0.36071	0.02352	0.32788	0.01835	0.33915	1.00000	1.00000	0.07306	0.34091
web	0.14031	0.30780	0.05358	0.33987	0.01272	0.30911	0.04189	0.32293	0.07306	0.34091	1.00000	1.00000
total spread	0.17798	0.74018	0.08019	0.82555	0.02898	0.75958	0.06781	0.78175	1.01356	1.45825	0.30996	0.79850
term spread	0.13283	0.02603	0.04729	0.04748	0.01728	0.04363	0.03565	0.03708	0.05470	0.04757	0.12758	0.03311
Ratio	1.34	28.44	1.70	17.39	1.68	17.41	1.90	21.08	18.53	30.65	2.43	24.12
Rank/LAKT	207	231	107	326	25	388	101	263	44	305	208	243

Table 5 – Key term activation levels with summary statistics: web citation thread (LAKT = Least Activated Key Term)

Probe Term	english		papers		student		students		write	
Key Terms	sentence	message	sentence	message	sentence	message	sentence	message	sentence	message
english	1.00000	1.00000	0.01125	-0.26056	0.01532	0.24663	0.06455	0.26835	0.01815	0.25939
papers	0.01125	-0.26056	1.00000	1.00000	0.02097	-0.25676	0.03765	-0.26584	0.01170	-0.25598
student	0.01532	0.24663	0.02097	-0.25676	1.00000	1.00000	0.04255	0.23409	0.01249	0.27516
students	0.06455	0.26835	0.03765	-0.26584	0.04255	0.23409	1.00000	1.00000	0.06093	0.24911
write	0.01815	0.25939	0.01170	-0.25598	0.01249	0.27516	0.06093	0.24911	1.00000	1.00000
total spread	0.07897	0.59258	0.05599	0.47936	0.08286	0.60347	0.19245	0.58105	0.06164	0.62064
term spread	0.05330	0.52891	0.02640	0.00986	0.03006	0.53192	0.02690	0.53419	0.04923	0.53114
term %	32.51%	10.75%	52.85%	97.94%	63.73%	11.86%	86.02%	8.06%	20.13%	14.42%
Rank of LAKT	69	1063	50	1052	48	1040	38	1071	25	1042
Percentile	93.84%	5.09%	95.54%	6.07%	95.71%	7.14%	96.61%	4.38%	97.77%	6.96%

Table 6 – Key term activation levels with summary statistics: writing thread (LAKT = Least Activated Key Term)

Chapter 5
Discussion

The results of this study can be summarized in the following comparisons: the message treatment produced networks with a larger overall spread, a smaller key term spread, and fewer highly activated noise strings.

Sentence Treatment	**Message Treatment**
Smaller network spread	Greater network spread
Greater key term spread	Smaller key term spread
More highly act. noise strings	Fewer highly act. noise strings

When processing the same data, the two treatments of self-organization produced very different networks in which the strength of association between the key terms, however, was remarkably similar. For the web citation thread, correlations of the vectors of activations for each probe term are low for all terms yet high for the key terms (Table 3). This shows that while the networks produced by the two treatments are different and the activation values of key terms are different between treatments the relationships among the key terms are similar. These results are repeated, though not as dramatically, for the writing thread (Table 4). This is evidence that both treatments are capable of clustering the key terms in similar fashion which, in turn, supports the choice of key terms.

The question of whether one self-organizing treatment produces better or more useful results than the other can be answered by looking at three things: the spread or differentiation between terms in the networks; the strength of the associations between the key terms in the networks; and, the amount of noise strongly associated with the key terms in each treatment.

Key term association strengths

In the following discussion terms are in quotes and individual cases are specified in the form: *"probe term"/treatment*. Using the "citation"/sentence case of the web-citation thread as an example, the spread of key term activations was .14031 - .00748 = 0.13283 (table 5).[5] The total spread of activations for all terms was 0.17798. The key terms are therefore slightly more (1.34 times) highly associated with "citation" than the least associated term. In the message treatment the key term activation spread was 0.02603 in a total activation spread of 0.74018 making the key terms almost thirty times (28.44) more highly associated with the probe "citation" than the term with the weakest association (table 5). The most highly activated key term ("web") was third in activation in the "citation"/sentence case, while the least activated key term ("engine") was 207th overall. In the "citation"/message case, the most highly activated key term ("counts") was ninth while the least activated key term ("engine") was 231st (table 5).[6] The cluster of key terms is shifted and compressed in the "citation"/ message case.

The writing thread produced slightly different results (Table 6). The most striking difference has to do with the key term "papers" which consistently responded to probing by the other key terms with a negative activation in the message self organizing treatment. Likewise, when "papers" is the probe term the other key terms assume negative activations in the message self organizing treatment. This is may be evidence that "papers" does not actually belong in the key term group or that this network model is reacting incorrectly to associate

[5] In all of these calculations the probe term, whose activation is, by definition, equal to one, is not included.

[6] It is interesting to note that the plural form ("engines") is the most highly activated term (after the probe term itself) in the "citation"/message case.

certain key terms.

The exclusion of noise is an indication of the utility of a text analysis system. Noise includes stop words which the *Online Dictionary of Library and Information Science* defines as:

> A frequently used word – usually an article, conjunction, or preposition – ignored when a computer executes a keywords search because it adds no value to the search statement, and its omission does not affect retrieval. Each database has its own stop list. Examples: a, an, as, at, by, for, from, of, on, the, to. Also spelled stopword. (Reitz, 2002)

Also included in concept of noise are numbers and character strings that are not words (e.g. "****"). Another way to define noise is *that which is not interesting or provides little or no insight into the meaning of the text being analyzed.* All of this together defines "noise strings" (as opposed to "noise words" because some of them are not words.) The number of noise strings in the 20 most highly activated nodes for each probe/self-organizing treatment case are shown in tables 7 and 8. The mean number of noise strings in the web citation thread was 14.6667 in the sentence treatment and 6.5 in the message treatment. For the writing thread the mean number of noise strings was 15.9 and 7.4 for the sentence and message treatments respectively.

The writing thread results provide another interesting insight into the data which can also be seen in the web citation thread with closer inspection. Examination of the activation levels for the term "papers" indicates that it does not belong in the key term group. Its activation level is much lower (in fact, negative) for the message treatment. When used as the probe term it causes many of the other key terms to be negatively activated. Looking closely at the

web citation thread we see a similar, though not as pronounced, effect for the term “citation”. For every probe term in the message treatment of self organization “citation” is the least activated of the key terms. Its activation is not negative, as is that for “papers” in the writing thread, but it is consistently the lowest. Likewise the mean key term activation for each probe term in the message treatment is lowest when “citation” is the probe term. For neither term does this effect show up as strongly in the sentence treatment. It is detectable in the writing thread in that probing with “papers” resulted in the lowest mean key term activation in sentence treatment as well as message treatment. In the original messages “papers” occurs in 7 of the 26 whereas “students” for example occurs in 19 of the messages. One possible explanation for this effect is that the message treatment is better able to classify terms into the key term group and that “papers” does not belong in the key term group. Other explanations, such as a hypersensitivity to some factor in the texts, surely exist which should be tested in future research.

Sentence	Message
CITATION	CITATION
THE	ENGINES
WEB	MORE
AND	THELWALL
THAT	ALLTHEWEB
FOR	**ALSO**
THIS	FOLLOWING
NOT	INFORMATIONR
ABOUT	COUNTS
MORE	SEARCH
WITH	PROF
CAN	PARTICULAR
THEIR	PROMOTION
USE	NET
CITATIONS	**AS**
FROM	**DID**
HAVE	SEARCHING
ARE	**GIVE**
PAGES	**ALL**
IMPACT	COMMAND
SOME	EXAMPLES
15	5

Sentence	Message
COUNTS	COUNTS
THE	THELWALL
AND	ALLTHEWEB
WEB	**FOLLOWING**
THAT	SEARCH
CITATION	**GIVE**
FOR	ENGINES
NOT	DIFFERENT
WITH	PARTICULAR
THIS	EXAMPLES
SEARCH	QUESTIONABLE
ARE	**WANTED**
CAN	RELIABILITY
SHOULD	INFORMATIONR
USE	**ALSO**
MORE	**PROF**
ABOUT	**DITTO**
FROM	**SAID**
IMPACT	MINING
LINKS	**COUPLE**
LINK	COMMAND
13	8

Sentence	Message
ENGINE	ENGINE
THE	ALLTHEWEB
SEARCH	THELWALL
THAT	**PARTICULAR**
AND	**FOLLOWING**
USE	INFORMATIONR
WEB	**PROF**
THIS	STUDY
FOR	NET
WITH	EXAMPLES
ALSO	QUESTIONABLE
WHICH	COMMAND
SAME	RELIABILITY
ALTAVISTA	WANTED
PROVIDES	HENCE
NOT	AIDS
CAN	NUMBERS
CITATION	**--**
RESULTS	**FINDING**
USING	**DITTO**
WAS	**SAID**
14	7

Sentence	Message
LINKS	LINKS
THE	ALLTHEWEB
AND	**THELWALL**
WEB	FOLLOWING
THAT	PARTICULAR
FOR	INFORMATIONR
THIS	PROF
CAN	SEARCH
NOT	**DID**
CITATION	COUNTS
FROM	COMMAND
WITH	NET
ARE	ENGINES
USE	**GIVE**
COUNTS	**HENCE**
ABOUT	EXAMPLES
MORE	QUESTIONABLE
THEIR	WANTED
SEARCH	RELIABILITY
PAGES	NUMBERS
RESEARCH	**--**
13	5

Sentence	Message
SEARCH	SEARCH
THE	ALLTHEWEB
AND	ENGINES
THAT	**THELWALL**
WEB	**DID**
ENGINES	**FOLLOWING**
FOR	**GIVE**
USE	INFORMATIONR
WITH	PROF
THIS	COUNTS
NOT	**ALSO**
RESULTS	PARTICULAR
CITATION	NET
ARE	EXAMPLES
FROM	QUESTIONABLE
CAN	**WANTED**
DID	RELIABILITY
ABOUT	DIFFERENT
MORE	COMMAND
ONE	**PRODUCED**
THAN	**HENCE**
15	8

Sentence	Message
WEB	WEB
THE	ENGINES
AND	THELWALL
THAT	ALLTHEWEB
FOR	**ALSO**
CITATION	**DID**
THIS	**THE**
NOT	SEARCH
CITATIONS	FOLLOWING
ABOUT	INFORMATIONR
MORE	MORE
ARE	COUNTS
PAGES	PROF
USE	PARTICULAR
WITH	PROMOTION
FROM	**ALL**
THEIR	NET
CAN	**AND**
THAN	STUDY
HAVE	DIFFERENT
SEARCH	**AS**
15	6

Table 7 – 20 most highly activated terms for each probe term in the two self organizing treatments (Web citation thread) showing noise in bold and total number of noise strings, probe terms shown at top of each list.

Sentence	Message
ENGLISH	ENGLISH
THE	**WHAT**
WRITING	INSTRUCTOR
AND	MARY
STUDENTS	**SHOULD**
THAT	DIFFERENT
BUT	**HOLD**
NOT	**THEIR**
WAS	**THROUGH**
ARE	MLS
FOR	**KINDS**
THIS	ITEMS
WHO	INTERNALIZATION
THEIR	AWARENESS
TO	7-Jan
WITH	**TRIED**
PEOPLE	LAYERS
MANY	EDUCATE
WHEN	BASICS
HAVE	TERMS
THERE	**SAW**
16	8

Sentence	Message
PAPERS	PAPERS
AND	**DONE**
THE	**MAYBE**
THAT	COLLEAGUES
WRITING	MEDIAN
FOR	AGE
NOT	BOSSES
STUDENTS	NORTHERN
ONE	IOWA
WAS	MEXICO
ARE	SUBMITTED
WHEN	POORLY
BUT	GROUCHY
WHO	**OBVIOUS**
WITH	JOKE
STUDENT	ANYTHING
WORK	DECREASED
THEIR	**TEN**
HAVE	**********
THIS	WAVE
WHAT	CRESTED
16	5

Sentence	Message
STUDENT	STUDENT
THE	INSTRUCTOR
AND	MARY
THAT	**HOLD**
NOT	**WHAT**
FOR	DIFFERENT
WRITING	**YES**
STUDENTS	**TRIED**
WAS	LAYERS
BUT	EDUCATE
ARE	BASICS
WORK	AWARENESS
WITH	INTERNALIZATION
PAPER	7-Jan
WHO	**KINDS**
THEIR	ITEMS
TWO	FORTUNATELY
THIS	**FELT**
ONE	DEGREE
LIKE	TEACHER
PAPERS	**SAW**
15	7

Sentence	Message
STUDENTS	STUDENTS
THE	MARY
AND	**WHAT**
WRITING	**SHOULD**
THAT	INSTRUCTOR
NOT	**HOLD**
ARE	**THEIR**
FOR	MLS
THEIR	**THE**
BUT	DIFFERENT
WITH	**AND**
WAS	**GOOD**
FROM	**YOU**
WHO	**THROUGH**
ENGLISH	**TRIED**
WORK	LAYERS
HAVE	EDUCATE
WRITE	BASICS
ALL	AWARENESS
WERE	INTERNALIZATION
TO	7-Jan
16	10

Sentence	Message
WRITE	WRITE
STUDENTS	MARY
THE	INSTRUCTOR
AND	**HOLD**
FOR	DIFFERENT
BUT	**WHAT**
ARE	**YES**
THAT	TRIED
WRITING	LAYERS
NOT	EDUCATE
TO	BASICS
WHO	AWARENESS
ENGLISH	INTERNALIZATION
WAS	7-Jan
LEARN	**KINDS**
HAD	**ITEMS**
LIKE	FORTUNATELY
THIS	**FELT**
ALL	DEGREE
OUR	TEACHER
WITH	**SAW**
16	7

Table 8 – 20 most highly activated terms for each probe term and self organizing treatment (writing thread) showing noise in bold and total number of noise strings, probe terms shown at top of each list.

Conclusion

Interactive Activation with Competition artificial neural networks are adaptable to continuous real time analysis of Internet email text streams. They can be designed to deal with new terms as they are encountered. Message-level context

can be implemented by allowing the external activations of the current message terms already processed to decay gradually rather than go to zero at the end of each sentence cycle. By employing a context oriented model several benefits are realized. For instance, by incorporating message level context in the model high frequency "stop words" are effectively ignored. Message level context also may allow better discrimination between related and unrelated terms.

Future Research

This research has explored the application of an Interactive Activation with Competition Artificial Neural Network to the analysis of text streams generated over a period of time. The development of this technology for text stream analysis has many potential applications including tracking public opinion, identifying shifts in consumer attitudes, detecting and following the adoption of new ideas and, monitoring the attitudes and thereby helping to predict the behavior of well defined groups. Before these capabilities are realized there needs to be additional research in several areas. The ANN model that was used employs several parameters (threshold, decay rate, learning rate and, others) that were not manipulated. The effect of altering these, and other parameters that could be added to this model such as the size of the message or the activation decay function, should also be explored. Terms encountered in short messages might warrant a greater effect on the network than terms in long messages.

Future research should be done to determine why the key term "papers" was negatively associated with the other key terms in the message treatment. An outlier like this presents an excellent opportunity to test a variety of hypotheses due to the apparent sensitivity of the network to the factor(s) contributing to its unusual handling.

Two areas that need further research are the handling of stop words and noise in general and appropriate treatment of sense shifters such as "not". Several techniques for allowing the network to learn to ignore noise should be explored including latching the activation level of words that are frequently activated and, desensitizing noise term nodes by altering their activation function.

Disambiguation is another hotly debated topic in the field of artificial intelligence. After sufficient text has been processed it is possible (this, itself, should be verified) that the network will be able to discriminate between "banks" of a river and savings "banks". As a practical matter, the automated application of thesauri or other disambiguation techniques should be researched.

The model used "attends" equally to all sentences, yet that does not reflect how humans communicate (Liddy, 1991). Therefore, another avenue of refinement which requires more research is the inclusion of additional discourse-level information in the model. The subject line was used to manually select the messages to be processed in this study. A network that recognizes subject lines could eliminate this manual step. One that treats the subject lines of messages differently could potentially process unrelated messages allowing intra- and inter-thread relationships to be explored. This would move such a system closer to the text mining model in terms of utility. Yet another discourse-level concept that should be researched is sequence. In other words, should terms and/or sentences that occur early, in the middle, or late in a message be given more or less weight in the self-organizing process?

Appendix A
Definitions - in alphabetical order

Activation – a value that is assigned to a node either manually or computationally. Manually applied activations are known as external activations (*ext(i)*)and result from a term being detected in the scanning window during learning or being used as a probe term during a probe. Computed activations are the products of the calculations involving external activations, activations from other nodes link strengths and the activation function of a node after a probe has been cycled (*act(j,2)*). External activations can be thought of as stimuli to the system and henceforth will be called external activations.

Activation Function – the mathematical function that is used to combine the inputs of a node to determine the node's level of activation. The activation function used in this network is the sigmoid function ranging from -1 to +1. Its mathematical form is given in Equation A1. In this equation act_t is the new activation of the node after the cycle; act_{t-1} is the activation level calculated by the previous cycle; *rf* is the decay rate and *anet* is the sum of the current activations multiplied by their associated link strengths (*win*), given in Equation A2.

$$act_t = \frac{1 - e^{-anet}}{\left(1 + e^{-anet}\right) + (rf \bullet act_{t-1})}$$

Equation A1. the sigmoid activation function

$$anet = \sum_{i=1}^{n} \sum_{j=1}^{n} act_{t-1}^{i} \bullet win_{i,j}$$

Equation A2. the anet value

Context unit – the terms which are excited during a learning cycle. This study compares the results obtained using two different context units. The first context

unit is the sentence and the second is the message.

Decay rate – the factor which is subtracted from all link strengths at the end of each learning cycle.

Email header – text that is added to the message by email software to facilitate routing and delivery of the message.

Internet text – Messages in ASCII taken from email messages sent to the Open Lib/Info Sci Education Forum (jesse@listserv.utk.edu). Email headers and quoted text (both defined in this section) are ignored.

Learning cycle - a processing of all terms in the network when a message is being analyzed for the first time. Nodes whose terms are present in the current scanning window are given an external activation of 1. The activations of all of the nodes are then calculated and the weights adjusted using a variation of the Hebbian rule. This rule reinforces connections between nodes which are externally excited, therefore strengthening their associations.

The learning rule used in this study is also based on the Hebbian rule and is mathematically:

$$gWin_{i,j} = gWin_{i_{t-1},j_{t-1}} + h((gAv_{i_{t-1}} - \bar{x})(gAv_{j_{t-1}} - \bar{x}))$$

where $gWin_{i,j}$ = the connection weight between nodes i and j at the end of a cycle
t-1 = the end of the previous cycle
gAv_i = the activation level of node i
$\bar{x}$ = the average activation
h = the learning rate.

Figure A3 - The Modified Hebbian rule

Learning rate – the factor which is used to calculate the amount of increase in the strength of links during a learning cycle. It is one of the parameters of the activation function. Figure A3 above shows this parameter (h) in the formula employed in this system.

Link – a connection between two nodes which provides a means for determining the effect one node may have on the other. If a node's level of activation is sufficiently great, that node will apply a signal to all other nodes through its links. The strength of the signal it applies to any given node will be determined by the value stored in the link. Each link, therefore, has a value which modifies the signals sent through it. In this model all signals originate with the same strength.

Network – a collection of nodes each being connected to all of the others by means of links of varying strengths or weights. The links allow the nodes to activate each other according to the parameters and algorithms in effect.

Node – a term that has been incorporated into the network.

Probe – one or more terms which are given an external activation of 1 at the beginning of a non-learning cycle for the purpose of determining which other terms are highly associated with the probe term(s).

Probe cycle – a processing of all terms in the network with the goal of discovering the associations that the network has developed during previous learning cycles. Each probe term is given an external activation of 1, the activations are allowed to spread and the resulting activation levels of the other nodes are calculated and reported.

Probe Term – a term whose external activation is set to 1 at the beginning of a probe cycle, in other words, a term of interest.

Quoted Text – Text appearing in a message that was originally in another message. This type of text is often included in messages so that the context of the reply is clear.

Recording units – those blocks of text that are specifically categorized during the analysis process. In this research the recording unit a term.

Sampling units – those blocks of text that are identified uniquely and from which the sample is drawn. Sampling units for this study will be all evaluation text blocks contained in selected (by subject line) messages posted to a specific list (jesse@listserv.utk.edu).

Sentence – a sequence of terms starting at the beginning of a message, or the end of the previous sentence and terminated by: A) a punctuation stop (a period, exclamation point, or question mark) or B) the end of the message.

Term – an individual word of three or more characters in length. No attempt is made to disambiguate terms nor to deal with hyphenations or contractions (e.g. "I'll").

Unit of analysis, – the connection strengths created in an IAC network which has processed email messages from a listserv distribution list. These units are generated by the IAC network by processing email messages which will be called the evaluation text blocks.

Appendix B
Subroutine which adds a node

```
Subroutine activate
!determines which node a term belongs to and activates the
node.
! or creates and activates a new node if no matching one
exists.

          use globals

  IF (len_trim(ANS) .EQ. 0 ) return
  DO j=1,gNcon
           if(gLabels(j) .EQ. ANS) THEN
                   ext(j)=gOn
                   return
           END IF
  END DO
  ! no match... must be a new term.
  gNcon = j !j is already incremented
  gLabels(gNcon) = ANS ! create the new node
  ext(gNcon) = gOn !activate it
  return
end
```

Appendix C
Sample Output (.tab file)

```
F= 2 T= 0.00010000 R = 0.10000000 H= 0.00500000 L= F C= 1.00010000
**************************       -0.02011338
TOM                               -0.006117421
WILSON                             0.001704015
RAISES                            -0.012098443
INTERSTING                        -0.01634914
QUESTION                           0.005586486
ONE                                0.029398214
THAT                               0.096060521
HAS                                0.01019123
BEEN                              -0.000306453
ADDRESSED                         -0.01634914
LITTLE                             0.003693493
THE                                0.15787151
LITERATURE                        -0.002844978
BEGAN                             -0.000616468
COLLECTING                        -0.007280892
MENTIONS                           0.03283193
AND                                0.1271401
CITATIONS                          0.04772453
FIRST                              0.000944492
A                                  0.0149505
CURIOUSITY                        -0.007280892
LATER                             -0.007280892
ADJUNCT                           -0.007280892
OTHER                              0.014202153
RESEARCH                           0.02306308
HOW                                0.02430037
LONG                              -0.007280892
DOES                               0.008511858
IT                                -0.002975465
TAKE                              -0.00574647
FOR                                0.084296554
WEB                                0.1403141
CITATION                           1
FALL                              -0.002566469
AWAY                              -0.007280892
EVER                              -0.002921485
ANSWER                            -0.000127478
DEPENDS                           -0.002921485
ALLOWS                            -0.002921485
DOCUMENT                           0.003765493
USES                              -0.002921485
OUR                                0.00446799
INTELLECTUAL                       0.004474485
OUTPUT                            -0.002921485
THAN                               0.03728746
ARTICLES                           0.01027621
WOULD                              0.02626864
SUGGEST                           -0.005577455
```

Note: The first line of the file is a list of the parameters used.

Appendix D.
A typical email list message
As received

··· ····· raises an intersting question, one that has been addressed a little in the literature. I began collecting "mentions" and "citations" first as a curiousity and later as an adjunct to other research (how long does it take for a web citation to fall away? - as ever the answer is "it depends")

Web citation allows us to document other uses of our intellectual output than citations in other articles. I would suggest that the inclusion of one's work as
part of the teaching of another may say something more about its "intellectual importance." It is possible to "capture" some course syllabi and reading lists from the WWW and to build citation counts. It might also be interesting to see some of Olle Persson's co-citation "venn diagrams" based on syllabi.

···· ·····

"Prof. ··· ·····" wrote:

> There have been a few mentions of Web citation searching possibly replacing
> citation indexing in time and I wondered how many people are now, as a
> matter of course, using counts of Web mentions in their cases for
> appointment, tenure or promotion.
>
> I looked at a couple of my own papers and counted the SSCI citations and
> then searched for mentions of the papers on the Web - the results left me
> wondering whether the reliance on citation indexing as a measure of
> performance is now past its sell by date.
>
> My most cited paper is "On user studies and information needs" (1981) - a
> Web search (using Google) revealed 118 pages that listed the title. The
> pages were reading lists, free electronic journals, and documents that would
> never be covered by SSCI, such as reports from various agencies. SSCI
> revealed, if I recall aright, 79 citations of the paper. The question is: is
> the Web revealing impact more effectively than SSCI? Citation in scholarly
> papers takes a variety of forms and much citation is of a token variety - x
> is cited because x is always cited. On the other hand citation on reading
> lists implies some positive recommendation of the text, and mention in
> policy documents and the like, implies (at least in some cases) that some
> benefit has been found in the cited document.
>
> It may also be that the use of Web citation would provide a more complete
> measure - I discovered, much to my surprise, that a 1971 text of mine on
> 'chain indexing' is cited on one reading list and in the bibliography of a
> document in German on classification. Greater international coverage is a
> further benefit of using Web citation.
>
> It strikes me that a move towards using Web citation as the measure of
> performance would be rather more useful than the use of citation indexes.
>

```
> No doubt others have looked at this issue - is any consensus emerging?
>
> ···· ·····
>
> ----------------------------------
> Professor ···· ·····, PhD
> Publisher/Editor in Chief
> ·········· ···········
> University of ·········
> ··············
> ··············
> Tel: +·················
> E-mail: ·········@···.·····.···
> Web site: http://··········/
> ----------------------------------
--
*******
········ ········
Associate Professor/Associate Director
Master of Library and Information Science Program
····· Library
·········· University
··············
·····, ·····
email - ·····@·········
voice: ··············

------------------ RFC822 Header Follows ------------------
Received: from [xx.xxx.xxx.xxx] ([xxx.xxx.xxx.xxx]) by jorg2.cit.buffalo.edu
(AppleShare IP Mail Server 6.3) id 36364 via TCP with SMTP; Mon, 20 May 2002 17:(
-0400
From: "·····”>
 (=?ISO-8859-1?Q?by=20way=20of=20Peter=20J=F6rgensen=20?=
 =?ISO-8859-1?Q?<titANN@jorg2.cit.buffalo.edu>?=)
X-Mailer: Mozilla 4.78 [en] (Windows NT 5.0; U)
References: <LNBBLJAEGEBGPGEBGMEMEEPDCDAA.··············>
Message-ID: <3CD67DB5.48F2F5C4@··········>
Date: Mon, 6 May 2002 08:57:25 -0400
Reply-To: Open Lib/Info Sci Education Forum <JESSE@LISTSERV.UTK.EDU>
Sender: Open Lib/Info Sci Education Forum <JESSE@LISTSERV.UTK.EDU>
Subject: Re: Web citation
Resent-Date: Mon, 20 May 2002 17:03:25 -0400
Resent-From: =?ISO-8859-1?Q?Peter=20J=F6rgensen?=
 <titANN@jorg2.cit.buffalo.edu>
MIME-Version: 1.0
Content-Type: text/plain; charset=US-ASCII
Content-Transfer-Encoding: 7bit
------------------------------------------------------------
```

As extracted and processed

> ···· ····· raises an intersting question, one that has been addressed a li
> in the literature. I began collecting mentions and citations first as a
> curiousity and later as an adjunct to other research (how long does it
> take for a web citation to fall away? - as ever the answer is it depends)

Web citation allows us to document other uses of our intellectual output t
citations in other articles. I would suggest that the inclusion of one's
work as
part of the teaching of another may say something more about its intellect
importance. It is possible to capture some course syllabi and reading list
from the WWW and to build citation counts. It might also be interesting to
some of Olle Persson's co-citation venn diagrams based on syllabi.

Appendix E
Message extraction script

```
property initialized : false
tell application "PowerMail"
	set theMessages to current messages
	repeat with msg in theMessages
		get content of msg
		my savemsg(the text of the result)
		my processMessage()
		set the status of msg to read
	end repeat
end tell
on savemsg(msgText)
	set theFile to ":phd:newmsg.dat"
	set fullFile to ":phd:allmsg.txt"
	tell application "Finder"
		if exists theFile then delete theFile
		if exists fullFile then
			set initialized to true
		else
			set initialized to false
		end if
	end tell
	try
		set outFile to open for access file theFile with write permission
		set outFile2 to open for access file fullFile with write permission
		repeat with p in (every paragraph of msgText)
			if the number of characters in p as text > 0 then
				if characters 1 through 1 of p as text is not in {">", "*",
"_"} then
					write p & (ASCII character 10) to outFile
					write p & (ASCII character 13) to outFile2 starting
at eof
				end if
			end if
		end repeat
	on error
		try
			close access outFile
		end try
		try
			close access outFile2
		end try
	end try
	try
		close access outFile
	end try
	try
		close access outFile2
	end try
end savemsg
on processMessage()
	if not initialized then
		do shell script "/Library/WebServer/CGI-Executables/listiac.exe <
/phd/win1.in"
		do shell script "/Library/WebServer/CGI-Executables/listiac.exe <
/phd/sent1.in"
	else
		do shell script "/Library/WebServer/CGI-Executables/listiac.exe <
/phd/win.in"
		do shell script "/Library/WebServer/CGI-Executables/listiac.exe <
/phd/sent.in"
	end if
end processMessage
```

Appendix F
Catpac output of the two threads
Web-citation thread

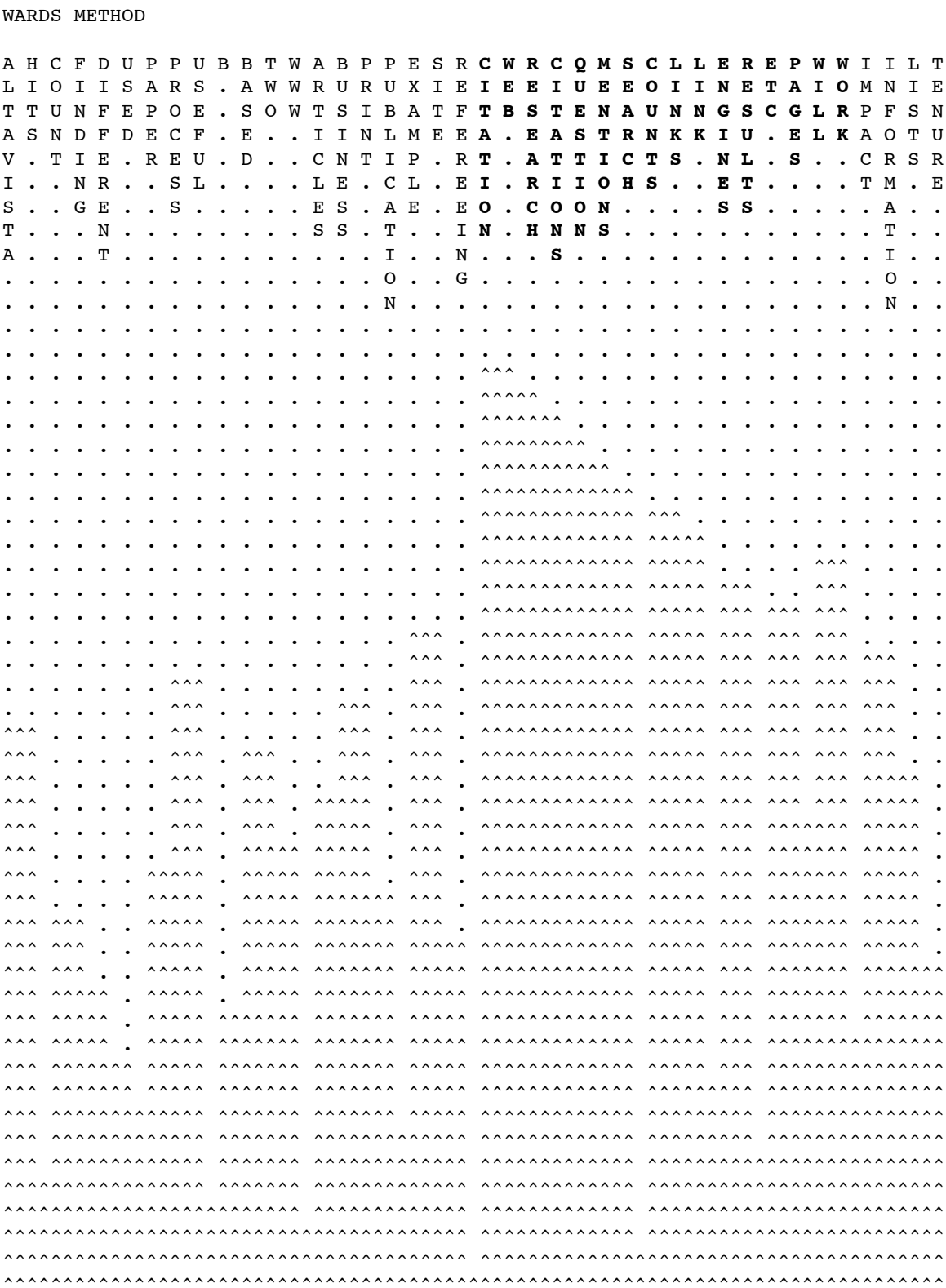
TOTAL WORDS	489	THRESHOLD	0.000
TOTAL UNIQUE WORDS	40	RESTORING FORCE	0.100
TOTAL EPISODES	483	CYCLES	1
TOTAL LINES	398	FUNCTION	Sigmoid (-1 - +1)
		CLAMPING	Yes

DESCENDING FREQUENCY LIST

WORD	FREQ	PCNT	CASE FREQ	CASE PCNT
WEB	71	14.5	326	67.5
CITATION	27	5.5	155	32.1
SEARCH	27	5.5	124	25.7
CITATIONS	22	4.5	116	24.0
LINK	17	3.5	81	16.8
ENGINES	16	3.3	78	16.1
INFORMATION	15	3.1	87	18.0
PAGES	13	2.7	81	16.8
COUNTS	12	2.5	76	15.7
IMPACT	12	2.5	67	13.9
LINKS	12	2.5	74	15.3
PRINT	12	2.5	55	11.4
RESULTS	12	2.5	72	14.9
WILL	12	2.5	79	16.4
ETC	11	2.2	70	14.5
RESEARCH	11	2.2	69	14.3
WORK	11	2.2	74	15.3
MENTIONS	10	2.0	57	11.8
TWO	10	2.0	58	12.0
REFEREEING	9	1.8	43	8.9
TENURE	9	1.8	53	11.0
ALTAVISTA	8	1.6	51	10.6
BASED	8	1.6	56	11.6
EXAMPLE	8	1.6	49	10.1
LISTS	8	1.6	56	11.6
QUESTION	8	1.6	50	10.4
USEFUL	8	1.6	39	8.1
ARTICLES	7	1.4	45	9.3
BUSINESS	7	1.4	39	8.1
COUNT	7	1.4	37	7.7
DIFFERENT	7	1.4	39	8.1
FINDING	7	1.4	27	5.6
HITS	7	1.4	39	8.1
PAPER	7	1.4	48	9.9
PROCESS	7	1.4	44	9.1
PUBLICATION	7	1.4	41	8.5
SITE	7	1.4	40	8.3
USED	7	1.4	44	9.1
WWW	7	1.4	49	10.1
B	6	1.2	38	7.9

ALPHABETICALLY SORTED LIST

WORD	FREQ	PCNT	CASE FREQ	CASE PCNT
ALTAVISTA	8	1.6	51	10.6
ARTICLES	7	1.4	45	9.3
B	6	1.2	38	7.9
BASED	8	1.6	56	11.6
BUSINESS	7	1.4	39	8.1
CITATION	27	5.5	155	32.1
CITATIONS	22	4.5	116	24.0
COUNT	7	1.4	37	7.7
COUNTS	12	2.5	76	15.7
DIFFERENT	7	1.4	39	8.1
ENGINES	16	3.3	78	16.1
ETC	11	2.2	70	14.5
EXAMPLE	8	1.6	49	10.1
FINDING	7	1.4	27	5.6
HITS	7	1.4	39	8.1
IMPACT	12	2.5	67	13.9
INFORMATION	15	3.1	87	18.0
LINK	17	3.5	81	16.8
LINKS	12	2.5	74	15.3
LISTS	8	1.6	56	11.6
MENTIONS	10	2.0	57	11.8
PAGES	13	2.7	81	16.8
PAPER	7	1.4	48	9.9
PRINT	12	2.5	55	11.4
PROCESS	7	1.4	44	9.1
PUBLICATION	7	1.4	41	8.5
QUESTION	8	1.6	50	10.4
REFEREEING	9	1.8	43	8.9
RESEARCH	11	2.2	69	14.3
RESULTS	12	2.5	72	14.9
SEARCH	27	5.5	124	25.7
SITE	7	1.4	40	8.3
TENURE	9	1.8	53	11.0
TWO	10	2.0	58	12.0
USED	7	1.4	44	9.1
USEFUL	8	1.6	39	8.1
WEB	71	14.5	326	67.5
WILL	12	2.5	79	16.4
WORK	11	2.2	74	15.3
WWW	7	1.4	49	10.1



In the Dendrogram above the main cluster terms are in **bold**.

CatPac output – writing thread

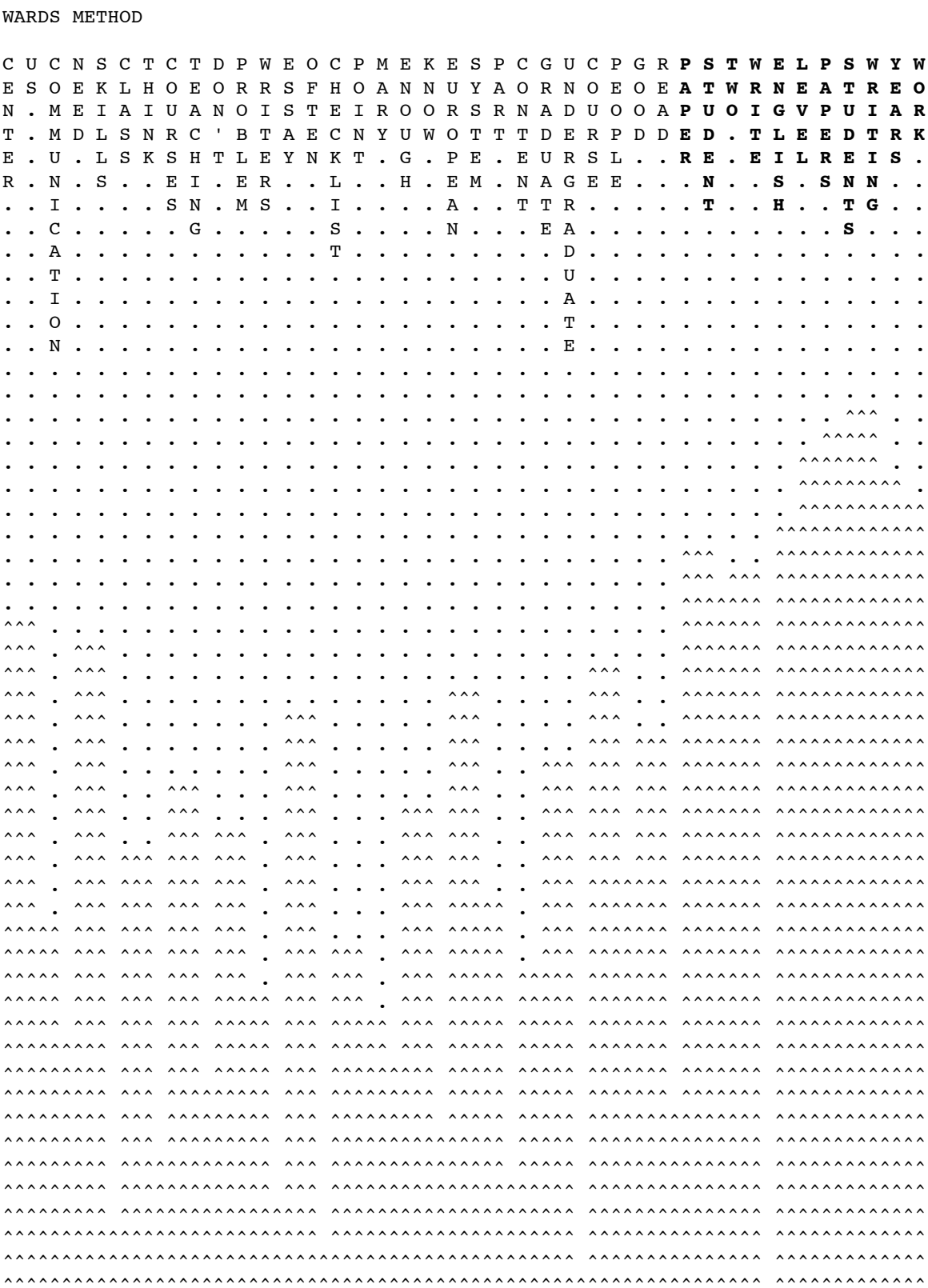
TOTAL WORDS	412	THRESHOLD	0.000
TOTAL UNIQUE WORDS	40	RESTORING FORCE	0.100
TOTAL EPISODES	406	CYCLES	1
TOTAL LINES	437	FUNCTION	Sigmoid (-1 - +1)
		CLAMPING	Yes

DESCENDING FREQUENCY LIST

WORD	FREQ	PCNT	CASE FREQ	CASE PCNT
WRITING	55	13.3	257	63.3
STUDENTS	46	11.2	228	56.2
ENGLISH	19	4.6	92	22.7
WORK	18	4.4	108	26.6
PAPERS	16	3.9	83	20.4
STUDENT	13	3.2	81	20.0
TWO	13	3.2	79	19.5
COURSE	11	2.7	65	16.0
PAPER	11	2.7	66	16.3
PEOPLE	11	2.7	62	15.3
US	11	2.7	60	14.8
WRITE	10	2.4	64	15.8
GOOD	8	1.9	54	13.3
OFTEN	8	1.9	51	12.6
POINT	8	1.9	44	10.8
CENTER	7	1.7	44	10.8
CONTENT	7	1.7	45	11.1
LEVEL	7	1.7	41	10.1
NEED	7	1.7	43	10.6
PROBLEM	7	1.7	44	10.8
READ	7	1.7	45	11.1
SKILLS	7	1.7	47	11.6
SYSTEM	7	1.7	32	7.9
WRITERS	7	1.7	49	12.1
COURSES	6	1.5	37	9.1
DON'T	6	1.5	42	10.3
ESSAY	6	1.5	36	8.9
GRADUATE	6	1.5	39	9.6
KNOW	6	1.5	42	10.3
MARY	6	1.5	42	10.3
PART	6	1.5	37	9.1
TEACHING	6	1.5	42	10.3
THINK	6	1.5	35	8.6
UNDERGRADUATE	6	1.5	33	8.1
YEARS	6	1.5	39	9.6
CHECKLIST	5	1.2	32	7.9
CLASS	5	1.2	35	8.6
COMMUNICATION	5	1.2	35	8.6
ENOUGH	5	1.2	35	8.6
EUROPEAN	5	1.2	30	7.4

ALPHABETICALLY SORTED LIST

WORD	FREQ	PCNT	CASE FREQ	CASE PCNT
CENTER	7	1.7	44	10.8
CHECKLIST	5	1.2	32	7.9
CLASS	5	1.2	35	8.6
COMMUNICATION	5	1.2	35	8.6
CONTENT	7	1.7	45	11.1
COURSE	11	2.7	65	16.0
COURSES	6	1.5	37	9.1
DON'T	6	1.5	42	10.3
ENGLISH	19	4.6	92	22.7
ENOUGH	5	1.2	35	8.6
ESSAY	6	1.5	36	8.9
EUROPEAN	5	1.2	30	7.4
GOOD	8	1.9	54	13.3
GRADUATE	6	1.5	39	9.6
KNOW	6	1.5	42	10.3
LEVEL	7	1.7	41	10.1
MARY	6	1.5	42	10.3
NEED	7	1.7	43	10.6
OFTEN	8	1.9	51	12.6
PAPER	11	2.7	66	16.3
PAPERS	16	3.9	83	20.4
PART	6	1.5	37	9.1
PEOPLE	11	2.7	62	15.3
POINT	8	1.9	44	10.8
PROBLEM	7	1.7	44	10.8
READ	7	1.7	45	11.1
SKILLS	7	1.7	47	11.6
STUDENT	13	3.2	81	20.0
STUDENTS	46	11.2	228	56.2
SYSTEM	7	1.7	32	7.9
TEACHING	6	1.5	42	10.3
THINK	6	1.5	35	8.6
TWO	13	3.2	79	19.5
UNDERGRADUATE	6	1.5	33	8.1
US	11	2.7	60	14.8
WORK	18	4.4	108	26.6
WRITE	10	2.4	64	15.8
WRITERS	7	1.7	49	12.1
WRITING	55	13.3	257	63.3
YEARS	6	1.5	39	9.6



Terms in the main cluster are in **bold**.

References

Abdi, H., Valentin, D., & Edelman, B. (1999). *Neural networks*. Thousand Oaks, CA: Sage Publications, Inc.

Absoft Corp. (2001). Pro FORTRAN for OSX (Version 7.0 SP3) [IDE]. Rochester Hill, MI: Absoft Corp.

Apple Computer, I. (2001). AppleScript (Version 1.8.2) [programming language]. Cupertino, CA: Apple Computer, Inc.

Barrow, J. D. (1991). *Theories of everything : the quest for ultimate explanation John D. Barrow*. Oxford <England>: Clarendon Press ; New York : Oxford University Press.

Belcher, D. D. (1999). Authentic interaction in a virtual classroom: leveling the playing field in a graduate seminar. *Computers and Composition, 16*(2), 253-267.

Bradley, J., & Rockwell, G. (1992, June 1992). *Towards new Research Tools in Computer-Assisted Text Analysis*. Paper presented at the The Canadian Learned Societies Conference.

Breen, M. J. *Agenda Setting and Public Opinion Formation: Media Content and Opinion Polls On Divorce Referenda in Ireland. .*

Bush, V. (1945,). As we may think. *Atlantic Monthly, 176*, 101-108.

Cailliau, R. (1995, Feb 16, 2001). *A little history of the world wide web* (1.24), [Web Page]. W3C. Available: http://www.w3.org/History.html [2001, August

28].

Caraballo, D., & Lo, J. (2000, 6/1/2000). *The IRC Prelude* (1.1.5), [web page]. Available: http://www.irchelp.org/irchelp/new2irc.html [2001, August 28].

Carley, K. M. (1997). Network Text Analysis: The network position of concepts. In C. W. Roberts (Ed.), *Text analysis for the social sciences : methods for drawing statistical inferences from texts and transcripts edited by Carl W. Roberts* (pp. 79-100). Mahwah NJ.: Lawrence Erlbaum.

Christe, N. (2001). PowerMail (Version 3.1.2) [email client]: CTM Development.

Cole, J. I. (2001). *The UCLA Internet Report 2001 – 'Surveying the Digital Future'*. Los Angeles: UCLA Center for Communication Policy.

Dworman, G. (1996). *Homer: a Pattern Discovery Support System.* Paper presented at the SIGCHI.

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science, 14*, 179-211.

Fayyad, U., Haussler, D., & Stolorz, P. (1996.). Mining scientific data. *Communications of the ACM, 39*, 51--57.

Fielding, D. A. (2000). *Cluster Analysis: What is it?*, [Web page]. Dept of Biological Sciences at Manchester Metropolitan University. Available: http://149.170.199.144/multivar/ca.htm [2002, April 7].

Garrison, D. R., Anderson, T., & Archer, W. (1999). Critical Inquiry in a Text-Based Environment: Computer Conferencing in Higher Education. *The*

Internet and Higher Education, 2(2-3), 87-105.

Garson, G. D. (1998). *Neural Networks: An Introductory Guide for Social Scientists*. (1 ed.). London: Sage Publications, Ltd.

Georgia Tech Research Corp. (1998, Tue, Jun 22, 1999 10:55:39 AM). *GVU's 10th WWW User Survey*, [Wb Site]. Georgia Tech Graphic Visualization and Usability Center. Available: http://www.gvu.gatech.edu/user_surveys/survey-1998-10/graphs/technology/ q105.htm [2002, May 31 2002].

Gurney, K. (1994, Apr 29 1994). *Neural Nets*, [Web Page]. Available: http://www.shef.ac.uk/ psychology/gurney/notes/l10/subsection3_2_1.html#SECTION00021000000000000000 [2002, May 31].

Hearst, M. A. (1999, june 20-26, 1999). *Untangling Text Data Mining.* Paper presented at the 47th Annual Meeting of the Association for computational Linguistics, University of Maryland.

Hewett, B. L. (2000). Characteristics of interactive oral and computer-mediated peer group talk and its influence on revision. *Computers and Composition, 17*(3), 265-288.

Jordan, M. I. (1997). Serial Order: A Parallel Distributed Processing Approach. In J. W. Donahose & V. P. Dorsel (Eds.), *Neural-Networks Models of Cognition:A Biobehavioral Approach* . Amsterdam: Elsevier.

Klerfors, D., & Huston, T. L. (1998, November 1998). *Artificial Neural Networks*, [Web page]. St. Louis University. Available: http://hem.hj.se/~de96klda/NeuralNetworks.htm [2001, Oct 1, 2001].

Kennard, W. E. (1999). *The Unregulation of the Internet: laying a Competitive Course for the Future* (Remarks before the Federal Communications Bar, North California Chapter). San Francisco, CA: FCC.

Lasswell, H. D. (1931). The Measurement of Public Opinion. *The American Political Science Review, 25*(2), 311-326.

Li, Y. (2000). Linguistic characteristics of ESL writing in task-based e-mail activities. *System, 28*, 229-245.

Liddell, C. (2001, June 21, 2001). *Is Sex The Only Profitable Business Model?* , [Web Page]. autralia.internet.com. Available: http://australia.internet.com/r/article/jsp/sid/233 [2002, June 1, 2002].

Liddy, E., Jörgensen, C. L., Sibert, E., & Yu, E. S. (1989). *Processing Natural Language for an Expert System Using a Sublanguage Approach.* Paper presented at the Annual Meeting of the American Society for Information Science (ASIS-89), Washington, DC.

Liddy, E. (1991). The discourse-level structure of empirical abstracts: An exploratory study. *Information Proc. & Management, 27*(1), 55-81.

L-Soft. (2002, 1 Jun 2002). *JESSE@LISTSERV.UTK.EDU*, [Web Page]. University of Tennessee, Knoxville. Available: http://www.lsoft.com/scripts/wl.exe?SL1=JESSE& H=LISTSERV.UTK.EDU [2002, June 1, 2002].

Lyman, P., & Varian, H. R. (2000, Thu, Mar 7, 2002 12:46:07 PM). *How Much Information*, [Web Site]. School of Information Management and Systems, UC Berkley. Available: http://www.sims.berkeley.edu/research/

projects/how-much-info/internet.html#4 [2002, April 18].

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88*, 375-407.

McKinnon, A. (1989). Mapping the Dimensions of a Literary Corpus. *Literary & Linguistic Computing, 4*(2), 73-84.

McMillan. (2000). The microscope and the moving target: The challenge of applying content analysis to the World Wide Web. *Journalism and Mass Communication Quarterly*.

Merkl, D., & Rauber, A. (2000). Document Classification with Unsupervised Neural Networks. In F. Crestani & G. Pasi (Eds.), *Soft Computing in Information Retrieval* . Germany: Physica Verlag & Co.

Nahm, U. Y. (2001). *Text Mining with Information Extraction: Mining Prediction Rules from Unstructured Text.* Unpublished PhD proposal, University of Texas, Austin, TX.

Nelson, T. H. (1992). *Literary machines : the report on, and of, Project Xanadu concerning word processing, electronic publishing, hypertext, thinkertoys, tomorrow's intellectual revolution, and certain other topics including knowledge, education and*. (Ed. 93.1. ed.). Sausalito: CA : Mindful Press.

Olaniran, B. A. (1995). Perceived communication outcomes in computer-mediated communication: an analysis of three systems among new users. *Information Processing & Management, 31*(4), 525-541.

Palmquist, R. A. (2002). *Content Analysis*, [Web page]. Available: http://www.gslis. utexas.edu/~palmquis/courses/content.html [2002, April 7, 2002].

Pasquale, J.-F. d., & Meunier, J.-G. (2001, June 16, 2001). *Categorisation techniques in computer assisted reading and analysis of texts (CARAT) in the humanities.* Paper presented at the 2001 Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, New York.

Popping, R. (2000). *Computer-assisted text analysis Roel Popping*. Thousand Oaks: Calif. ; London : SAGE.

Reitz, J. M. (2002, May 15, 2002). *ODLIS: Online Dictionary of Library and Information Science*, [Web Page]. Western Connecticut State University. Available: http://vax.wcsu.edu/library/odlis.html [2002, May 31, 2002].

Rubenstein, S. M. (1995). *Surveying Public Opinion*. Belmont, CA: Wadsworth Publishing Co.

Rumelhart, D. E., McClelland, J. L., & Group, t. P. R. (1986). *Parallel Distributed Processing*. Cambridge, MA: The MIT Press.

Salton, G. (1968). *Automatic information organization and retrieval*. New York: McGraw Hill.

Sanderson, M. (1994, May 22, 1994). *list of stop words*. University of Glasgow, Department of Computing Science. Available: http://www.dcs.gla.ac.uk/ idom/ir_resources/ linguistic_utils/stop_words [2002, May 30, 2002].

Schneider, S. M. (1997). *Expanding the Public Sphere through Computer-Mediated Communication: Political Discussion about Abortion in a Usenet Newsgroup.* Unpublished PhD, Massachusetts Institute of Technology, Cambridge, MA.

Seidenberg, T. G. (1995, Oct 6, 1995). *alt.culture.usenet FAQ (Frequently Asked Questions)* (1/6), [Web page]. Available: http://www.faqs.org/faqs/usenet/culture-faq/ [2001, August 28].

Sierpe, E. (2000). Gender and Technological Practice in Electronic Discussion Lists: An Examination of JESSE, the Library/Information Science Education Forum. *Library & Information Science Research, 22*(3), 273-289.

Simpson, P. K. (1990). *Artificial Neural Systems*. (First ed.). New York: Pergamon Press.

Shapiro, G., & Markoff, J. (1997). A Matter of Definition. In C. W. Roberts (Ed.), *Text analysis for the social sciences : methods for drawing statistical inferences from texts and transcripts edited by Carl W. Roberts* (pp. 9-31). Mahwah NJ.: Lawrence Erlbaum.

Shi-xu. (2000). Opinion Discourse: Investigating the Paradoxical Nature of the Text and Talk of Opinions. *Research on Langugage and Social Interaction, 33*(3), 263-289.

Simpson, P. K. (1990). *Artificial Neural Systems*. (First ed.). New York: Pergamon Press.

Snyder, H., & Kurtze, D. (1996). Chaotic behavior in computer mediated network

communication. *Information Processing & Management, 32*(5), 555-562.

Steiner, P. (1993). On the Internet, nobody knows you're a dog . New York: The New Yorker.

Stolcke, A., & Segal, J. (1994). *Precise n-gram Probabilities from Stochastic Context-free Grammars.* Paper presented at the 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, New Mexico,.

Stone, P. J. (1997). Thematic Text Analysis: New Agendas for Analyzing Text Content. In C. W. Roberts (Ed.), *Text analysis for the social sciences : methods for drawing statistical inferences from texts and transcripts edited by Carl W. Roberts* (pp. 9-31). Mahwah NJ.: Lawrence Erlbaum.

Swanson, D. R. (1993). Intervening in the life cycles of scientific knowledge. *Library Trends, 41*(4), 606-631.

Thapalia, C. F. . *Multivariate Statistics: An Introduction*. Available: http://trochim.human.cornell.edu/tutorial/flynn/multivar.htm [2001, Oct 1, 2001].

Trushell, J., Reymond, C., & Burrell, C. (1998). Undergraduate students' use of information elicited during e-mail "tutorials". *Computers & Education, 30*(3-4), 169-182.

Whitney, G. (2002, May 8, 2002). *Jesse*, [Web Page]. University of Tennessee, Knoxville. Available: http://web.utk.edu/~gwhitney/jesse.html [2002, June 1, 2002].

Witten, I. H. (2001). *Adaptive Text Mining: Inferring Structure from Sequences*, [PDF

File]. Available: http://citeseer.nj.nec.com/rd/42058339%2C502339%2C 1%2C0.25%2C Download/http%3AqSqqSqwww.cs.waikato.ac.nzqSq %7EihwqSqpapersqSq01IHW-Adaptivetextmining.pdf [2002, May 31].

Woelfel, J., & Styanoff, N. J. (1993). *CATPAC: A Neural Network for Qualitative Analysis of Text.* Paper presented at the Australian Marketing Association, Melbourne, Australia.

Wolfe, J. L. (1999). Why do Women Feel Ignored? Gender Differences in Computer-Mediated Classroom Interactions. *Computers and Composition, 16*(1), 153-166.

Zakon, R. H. (2001, Mar 6, 1998). *Hobbes' Internet Timeline v5.4* (5.4), [Web Page]. Robert H Zakon. Available: http://www.zakon.org/robert/internet/ timeline/ [2001, Oct 8, 2001].