Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 41 (2005) 1081–1099
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Incorporating context in text analysis by interactive activation with competition artificial neural networks

Peter Jörgensen *

School of Informatics, 534 Baldy Hall, University at Buffalo, Buffalo, NY 14260, United States

Received 12 December 2003; accepted 12 October 2004
Available online 24 November 2004

Abstract

Many segments of modern society, including marketing, politics, government, activism and public safety, desire the ability to find relationships, thus meaning, in public discourse. This can be accomplished by analyzing communication documents according to their content. The increasing use of the Internet for public dialog has made Internet communication a potentially rich source of data in this regard. This study explores the use of an interactive activation with competition (IAC) artificial neural network (ANN) to find relationships in email texts. A modified fully recurrent IAC network was used to process 69 email messages from two threads in the Open Library/Information Science Education Forum using two variations of the self-organizing phase of network formation. These variations were: (1) with and (2) without a linear decay function applied between sentences to the external activation of nodes. The use of the linear decay function, which could be considered a method for including context, produced three positive effects: the entire network was more differentiated from keywords; the keywords were more highly associated with each other, and; roughly half the number of noise strings were highly associated with keywords.

Keywords: Text classification; Artificial neural networks

1. 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

* Present address: College of Information, 101 Shores Building, MS2100, Florida State University, Tallahassee, FL 32306. United States. Tel.: +1 850 644 4139; fax: +1 850 544 6253.
E-mail address: peter.jorgensen@fsu.edu

0306-4573/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2004.10.003

mind of "algorithmic compression". (Barrow, 1991, pp. 10–11) This can be applied to the social sciences where the goal is to make sense of the diversity of human activities. Of particular interest to scholars of communication and informatics are the patterns and associations that occur in texts, such as newspaper articles, works of literature, conversations or email messages. 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 & Meunier, 2001), mass media (Danielson & Lasorsa, 1997), political science (Franzosi, 1997), philosophy (Pasquale & Meunier, 2001) and others.

In the applied realm, individuals and organizations are interested in accurate and robust means to track opinions and attitudes for marketing, political, and social reasons (Lasswell, 1931). Personal attitudes and opinions are expressed and shared through a variety of communication modalities such as email. Email messages in public forums, such as listserv lists, can be a steady and plentiful source of personal expression in machine readable 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 & Varian, 2000). This is clearly too much for existing manual methods to analyze. Analysis of these texts, therefore, requires automated techniques that can deal with natural language. Large corporations, for instance, employ automated text analysis methods to help them discover information in focus group transcripts (Woelfel & Styanoff, 1993).

This paper reports on modifications of an interactive activation with competition (IAC) artificial neural network (ANN) algorithm to incorporate the notion of context, defined here as the words in the sentences preceding a sentence being processed. The results from applying two approaches to external activation of words during self-organizing were compared. In the first (sentence) approach the ANN set the external activations of all words to zero after each learning cycle took place, i.e. after each sentence was processed. In the second (message) approach the ANN reduced (without going below zero) the external activations of the words by a reduction factor after the learning cycle had been run—the external activations were set to zero only when the beginning of a new message was encountered.

The rationale for this second approach comes from the observation that the specific meaning of each word in a text depends on the context within which the text is set. The context is set, in part, by the other words in the text. Therefore, taking these words into account adds a degree of context to an analysis. There are, of course, other contextual clues that assist the reader in assigning specific meaning to specific words. These include, for instance, the source of the text (i.e. where it was published and by whom), the references to other texts made within the work, etc. This research focuses on the immediate contextual clues, those that can be derived from the surrounding text.

After a brief discussion of text classification and artificial neural networks, I will present the methodology used, the results obtained using two samples and some conclusions that are suggested. Finally, some suggestions for future research will be made.

2. Text classification

There are a variety of methods of text analysis including: content analysis (Breen, 1997; McKinnon, 1989; McMillan, 2000; Palmquist, 2002; Rubenstein, 1995; Schneider, 1997; Shi-Xu, 2000; and others); text mining (Dworman, 1996; Hearst, 1999; Witten, 2001); natural language processing (Liddy, 2001; Mani, 1999); latent semantic analysis (Dumais, Furnas, Landauer, Deerwester, & Harshman, 1988); probabilistic latent semantic indexing (Hofmann, 1999); and the topic of this study, artificial neural networks (Belew, 1996; Woelfel & Styanoff, 1993). It is beyond the scope of this paper to review all of the methods of text analysis. The interested reader should refer, for instance, to Belew (1996). Current methods have two limitations which this study addresses. First, in order to disambiguate terms they often rely on a multistage processes to implement contextual awareness. Second, they usually employ stop word lists which, again,

require pre-processing and preclude real-time processing of text streams. Additional pre-processing steps are often needed to conflate plurals and other variants of words to a single term.

2.1. Limitations of current methods

2.1.1. Lack of contextual awareness

Current methods of text analysis generally do not take into consideration "context". By this I mean words are assigned meanings based on a syntactic, rather than a semantic, level of analysis. Therefore, there is no disambiguation of homographs. The "bank" of a river is taken as the same as the "bank" in which we save our money. Within the neural network approach some attempts have been made to address this issue. Other methods, e.g. sublanguage approaches (Liddy, 2004), have also attempted to deal with this issue. They will not be considered further here. The approaches the ANN methods generally use rely on multi-stage analyses (for example, to identify nouns, noun-phrases and attached modifiers, etc.). Grefenstette (1994) describes one such method which involves five pre-processing steps before a weighted Jaccard measure of the terms relatedness is calculated. Aizawa (2002) reports on a clustering technique using an iterative discounted probabilistic creation and selection of microclusters. Ohgaya, Simmura, and Takagi (2003) incorporate the method reported by Aizawa (2002) in their conceptual fuzzy set method which also uses term co-occurrence in neighboring web pages to improve context sensitivity. Another approach to the problem of homography is the use of multiword terms and noun phrases (Grefenstette, 1994). While these approaches have shown some success, the extra complications and steps they introduce is a drawback. This paper describes a simple yet effective way to achieve similar results using a simple interactive activation with competition artificial neural network.

2.1.2. The use of stop word lists

Current text analysis methods often rely on the use of "stop word" lists to filter out (generally) small words that are assumed to occur with a high frequency and to add little or no meaning to the text. This commonly used technique "in *some* cases *may* hope to improve retrieval" (emphasis added) (Wilbur & Sirotkin, 1992, p. 45). The increasing use of short acronyms (such as *IT*, *WWW*, *IS*, *IR*, *KM*) (Kennedy, 2001), further calls into question the value of eliminating stop words as commonly implemented, i.e. words containing fewer than some arbitrary small number of letters. Additionally stop word lists suffer from the same problem that all controlled vocabularies do, namely the slowness with which they are able to respond to changes in the vocabulary of a domain. The method presented here does not rely on the use of stop word lists, but, nonetheless, reduces the significance of these words.

2.2. 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 act together to perceive, learn, and recall patterns. The fundamental concept is that the nodes (also called units or neurons) are highly interconnected by links whose strengths (or weights) are dynamically adjusted by a learning algorithm when patterns of stimuli are presented. Stimuli may be words, pixels or anything else that can be differentiated in input to the system. Each stimulus is represented by a node in the network. [1] Thus when a pattern (of words, in this case) is presented to the system the nodes that represent these words are externally activated and, through the learning algorithm, their connections to each other are strengthened. Depending on the exact model and algorithms used,

[1] In most implementations the nodes are populated by an initial pass of the text(s) in which a node is created for each unique word. In this implementation the nodes are added to the network as new words are encountered.

connections to unactivated nodes may also be weakened during the learning step. In this way ANNs become sensitized to patterns of stimuli which can be recalled at a later time when a fragment or an inexact copy of the original pattern is presented. ANNs used in this way are a type of associative memory. Thus, ANN software can be used to discover relationships between words in bodies of natural language text (Merkl & Rauber, 2000; Woelfel & Styanoff, 1993) by developing associations between words based on exposure to those words as they occur together in the text. However, ANN systems typically do *not* differentiate between homographs because they operate on the syntactic rather than the semantic level. This makes them less than ideal for real time text analysis (Elman, 1990). ANN tools would, therefore, be better able to deal with homography if they could be made to take context into account.

Generally speaking, ANNs have two modes of operation, self-organizing and probing. The first, or self-organizing phase, is when the network is "learning" about the texts. The second, or recall phase is used to discover what the network has "learned". The connection values (which are set to zero when a node is created) are refined by a learning algorithm applied to the links as the software encounters the features that these nodes represent in the self-organizing mode. Nodes representing words that are encountered within a scanning "window", in this case defined as a sentence but in many implementations an arbitrary or user-define number of adjacent words, are co-activated. A learning step is performed next in which the activations spread through the network via the links. Each node is activated to a degree determined by the weights of the links from nodes that have fired. Connections between highly activated nodes are strengthened and those between less activated nodes are weakened by the application of a Hebbian learning rule (Klerfors & Huston, 1998). The final step in the learning process is a normalization of the weight matrix resulting in an average weight of zero and a range of -1 to $+1$.

Repeatedly presenting similar patterns causes the link weights to gradually assume values that allow patterns to be recalled given partial or inexact pattern presentation. Recall is performed by exciting one or more (probe) nodes and allowing the activation to spread. This causes each node to become activated to the degree to which it has become associated with the probe(s) during the learning phase. During recall the learning step is not performed, i.e. connection weights are not modified. The interested reader can find thorough reviews of the development of artificial neural networks in Garson (1998) and Simpson (1990).

3. Methodology

3.1. Definitions

For the purpose of this study, *text* is defined as email messages exclusive of smtp headers, signature "lines" and quoted text. Email was chosen because it is a readily available machine readable example of natural language expression of ideas and opinions typical of those of interest to text analysts. A *sentence* is a series of words delimited by final punctuation marks ("."", "!" and "?"), or the end of the text block. A *word* is any series of characters delimited by a space or punctuation mark and not containing spaces or punctuation marks. (All characters in words are converted to uppercase.) The *network* is an $N \times N$ matrix of link weights where N = the number of *nodes* each of which represents a unique word that has been encountered in the text being processed. No attempt is made to conflate plurals and other variants to single terms nor to filter out "stop words". Nodes are added dynamically as new words are encountered. At any given time each node has two *activation levels* which are real numbers. The *external activation* of all nodes is initially zero. When a word is encountered in the sentence currently being processed the external activation of its node is set to 1. *Computed activations* are calculated during the spreading activation phase. In this network model each node is connected to all of the others nodes, including itself through *links*, whose weights form the $N \times N$ matrix of the network. When a node's level of activation reaches or exceeds a programmatically determined threshold it will *fire* sending a unit signal 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 (in this case always +1) and the value, or weight, stored in the link. An *activation function*, in this case a sigmoid function (Atiya, 1990), is applied to the sum of the signals reaching a node through all of its links. The result of the function is the computed activation of the node which determines whether or not the node will fire, and what its final activation level will be.

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 (Klerfors & Huston, 1998). It reinforces the connections between nodes which are externally excited, thereby strengthening their associations. The modified IAC ANN used in this study is that used in the commercial program Catpac™ (Woelfel & Styanoff, 1993) which includes a learning rate h in the Hebbian learning rule. The default learning rate (0.005) used in Catpac™, having been arrived at as a good starting point during the development and extensive use of that program, was used here as well. The modified Hebbian learning rule can be expressed mathematically as

$$Win_{i_t,j_t} = Win_{i_{t-1},j_{t-1}} + h((Av_{i_{t-1}} - \bar{x})(Av_{j_{t-1}} - \bar{x})) \quad (1)$$

where Win_{i_t,j_t} is the connection weight between nodes i and j at the end of cycle t, $Win_{i_{t-1},j_{t-1}}$ the connection weight between nodes i and j at the end of cycle $t-1$, Av the activation level of a node (i or j) at the end of cycle $t-1$, $\bar{x}$ the average activation levels of all nodes and h is the learning rate.

After a learning cycle takes place, the external excitations of the nodes are reduced before the next sentence is processed. Finally, in both algorithms, the network connection weight matrix is normalized, i.e. centered on zero and limited to the range of -1 to $+1$ (Jordan, 1997).

During recall one or more nodes (corresponding to one or more *probe words*) are externally activated. The network is then cycled without a learning step. The activation levels that result are known as a *case* which can be output as a list of all of the nodes, in network entry order, with the activation levels produced by the probe.

3.2. Software adaptation

This research is investigating a method of introducing context into the learning phase of an interactive activation with competition (IAC) network (McClelland & Rumelhart, 1981). To this end the core engine of Catpac™ (Woelfel & Styanoff, 1993), a commercially available IAC-based application, was adapted (Jorgensen, 2003). The software, originally written in FORTRAN77 as an interactive command-line driven program to run under MS-DOS, was ported to FORTRAN95 and modified to run under Darwin (a BSD derivative) as a cgi. [2] A *context* (sentence or message) selector was added to the original parameters of threshold, learning rate, decay rate, activation function, input file name, network file name and output file name. The context selector was used to determine which of two external excitation reduction algorithms would be used as described above and thus forms the basis of this experiment. The two excitation reduction algorithms apply only in learning mode. In the first (sentence) algorithm, the external excitations of all nodes are reduced to zero after a sentence is processed. This is the standard behavior for this type of network. In the second (message) algorithm the external excitations are reduced by a decay function after a sentence is processed and reduced to zero at the end of the message. The decay function in this case is a linear reduction (with a lower limit of zero) by a factor *rf*. For this reason, reduction by factor *rf* is called the "message" algorithm. The rationale behind using the decay function in the second algorithm is based on the observation that when humans process the text of an email they do so in a context that includes, among other things, the prior sentences of the email. The words in any sentence are most strongly related to each other, but they are also related to words in sentences preceding and following them, i.e. their context.

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

3.3. Data collection

Two sets of email messages were collected to serve as test data. Messages were manually selected by inspection of the subject lines of incoming messages posted to the topical internet mailing list, Open Lib/Info Sci Education Forum (L-Soft, 2002) and saved as text files. Twenty-three messages concerning the use of web citations in evaluation of scholarship were selected and will be called the *web citation set*. Twenty-six messages concerning the writing ability of students were collected and will be called the *writing set*. Email headers, signature lines and quoted text were removed from the messages when they were saved as these elements were not relevant to the research question.

3.4. Network manipulation

The parameter that was manipulated was the context selector, as described above. Descriptive statistics such as the number of unique words and word frequencies were generated. Processing the messages from each thread under the two algorithms (sentence and message) produced two networks. Although the system allows multiple simultaneous probe words, this experiment used just one probe word at a time for each probe cycle. All of the keywords (defined below) were used individually as probe words with networks generated by both algorithms producing two cases per keyword (one for each treatment), for a total of 12 cases for the web citation set of data and 10 cases for the writing set.

For the purpose of choosing keywords, the messages were analyzed by a professor who teaches indexing in an accredited library science program and by the commercial software program Catpac™ (Woelfel & Styanoff, 1993). Appendices A and B show the frequencies and dendrograms created by Catpac™ for the two sets of messages. The indexer extracted 17(10) web citation (writing) keywords that she felt represented the overall content of the collections while Catpac™ placed 16 (9) web citation (writing) words in the main cluster. 6 (5) web citation (writing) words identified by the indexer also appeared (along with their plurals) in the main cluster created by Catpac™. An identical procedure was followed to select keywords for the writing set. These words are shown in italics in Tables 1 and 2 and were chosen as keywords to be used as probes and indicators of performance. They are: *citation, counts, engine, links, search*, and *web* for the web citation set and: *English*, *papers*, *student*, *students*, and *write* for the writing set.

Table 1
Indexer and Catpac™ keywords for web-citation set

Indexer words	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 2
Indexer and Catpac™ keywords for writing set

Indexer words	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	

4. Results

4.1. Summary statistics

Summary statistics are shown in Table 3.

4.1.1. Web-citation set

The 23 messages in the web-citation set consist of 3214 words of which 1121 are unique and therefore formed the 1121 nodes of the network. The longest message had 387 words and the shortest 21. As would

Table 3
Twenty most frequently occurring words in data sets

Data set	Web citation set		Writing set	
Total words	3214		3055	
Unique words	1121		1262	
Rank	Word	Frequency	Word	Frequency
1	the	198	the	145
2	and	98	and	90
3	web	74	that	55
4	that	63	writing	52
5	for	52	students	47
6	research	37	for	39
7	this	30	not	36
8	not	28	are	33
9	are	27	was	28
10	citation	26	with	26
11	about	25	this	22
12	with	24	have	21
13	citations	22	done	21
14	from	21	But	21
15	use	20	our	19
16	more	20	from	19
17	have	20	all	19
18	can	19	work	17
19	one	18	their	17
20	link	17	English	17

Table 4
Summary of findings

Learning treatment	Sentence	Message
Term differentiation (total activation spread)	Smaller	Greater
Keyword clustering (activation spread)	Greater	Smaller
Number of highly activated noise strings	More	Fewer

be expected without the use of a stopword list, the most frequently used word was "the" occurring more than twice as often (198 times) as the next most frequent word ("and", 98 occurrences). The next most frequently occurring words, in order of frequency, were "web", "that", "for", "research" and, "this".

4.1.2. Writing set

The 26 messages of the writing set consist of 3055 words of which 1262 are unique creating a 1262 node network. The longest message contained 329 words and the shortest 17. The most frequently used word again was, "the" followed by "and", "that", "writing", "students", "for" and, "not".

4.2. Data analysis

The question of whether one text analysis algorithm produces better or more useful results than another can be explored by looking at three indicators: the differentiation between words in the networks; the clustering of keywords; and, the amount of noise associated with the keywords. This study's findings (Table 4) can be summarized by the following generalizations: the message algorithm produced networks with a greater term differentiation, as measured by total activation spread; tighter keyword clustering, as measured by (smaller) keyword activation spread and; better rejection of noise words as measured by fewer highly-activated noise strings. It should be noted that all of these values were obtained with no attempt to adjust parameters to improve performance. Each of these findings will be discussed next.

4.2.1. Term differentiation

Term differentiation is measured by the activation spread which is the difference between the highest and lowest activation levels for nodes in the network after probing. A network (or portion thereof) with a small spread suggests that the nodes are more equally related to the probe word(s) than does a larger spread. Activation spreads for all words (total activation spread or TAS) and for keywords (keyword activation spread or KAS) were calculated for all cases (Tables 5 and 6). As would be expected, the activation spread for keywords is generally less then that for the entire network because the keywords are each more closely related to the probe word than are all the words in general. This is true for both the sentence and message algorithms.

4.2.2. Total activation spread

Comparisons of the mean TAS values for the entire network across the two self-organizing algorithms reveals one of the differences that the message algorithm makes. For the web-citation data the mean TAS was 0.27975 [3] for the sentence algorithm and 0.89397 for the message algorithm. For the writing data the figures are 0.09438 for the sentence algorithm and 0.57542 for the message algorithm. In both cases the message algorithm created a more highly differentiated network (Table 7).

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

Table 5
Keyword activation levels[a] with summary statistics—web citation set

Probe word algorithm	citation		counts		engine		links		search		web	
	sentence	message	sentence	message	sentence	message	sentence	message	sentence	message	sentence	message
keywords												
citation	1.00000	1.00000	0.04108	0.31323	0.00748	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
TAS	**0.17798**	**0.74018**	**0.08019**	**0.82555**	**0.02898**	**0.75958**	**0.06781**	**0.78175**	**1.01356**	**1.45825**	**0.30996**	**0.79850**
KAS	**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

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

Table 6
Keyword activation levels with summary statistics—writing set

Probe word algorithm	english		papers		student		students		write	
	sentence	message	sentence	message	sentence	message	sentence	message	sentence	message
keywords										
English	1.00000	1.00000	0.01125	−0.26056	0.01531	0.24662	0.06455	0.26834	0.01815	0.25938
papers	0.01125	−0.26056	1.00000	1.00000	0.02097	−0.25675	0.03764	−0.26584	0.01170	−0.25597
student	0.01531	0.24662	0.02097	−0.25675	1.00000	1.00000	0.04254	0.23409	0.01249	0.27516
students	0.06455	0.26834	0.03764	−0.26584	0.04254	0.23409	1.00000	1.00000	0.06093	0.24910
write	0.01815	0.25938	0.01170	−0.25597	0.01249	0.27516	0.06093	0.24910	1.00000	1.00000
TAS	**0.07897**	**0.59258**	**0.05598**	**0.47936**	**0.08285**	**0.60346**	**0.19245**	**0.58105**	**0.06163**	**0.62064**
KAS	**0.05329**	**0.52890**	**0.02639**	**0.00986**	**0.03005**	**0.53192**	**0.02690**	**0.53418**	**0.04923**	**0.53114**
Ratio	1.48	1.12	2.12	48.62	2.72	1.13	7.15	1.09	1.25	1.17

Table 7
Total activation spreads compared

Data set	Algorithm	TAS	implication
Web-citation	Sentence	0.27975	Less differentiated
	Message	0.89397	More differentiated
	Ratio	0.3	
Writing	Sentence	0.09438	Less differentiated
	Message	0.57542	More differentiated
	Ratio	0.2	

Table 8
Total activation spread and keyword activation spread compared

Data set	Algorithm	TAS	KAS
Web-citation	Sentence	0.27975	0.06922
	Message	0.89397	0.03915
Writing	Sentence	0.09438	0.03717
	Message	0.57542	0.42720

4.2.3. Keyword clustering

Comparing the means of the TAS and KAS shows that the keywords have a smaller activation spread than the whole network and therefore the software is, indeed, clustering the keywords. For the web-citation data the mean TAS (across all probes) for the sentence algorithm is 0.27975 while the mean KAS is 0.06922 for this treatment. [4] Similarly, for the writing set sentence algorithm the mean TAS is 0.09438 while the mean KAS is 0.03717. For the web-citation set message algorithm the mean TAS is 0.89397 while the mean KAS is 0.03915 and for the writing set/message algorithm the values are 0.57542 and 0.42720 respectively. Table 8 summarizes these results.

Comparing the KAS values for the two algorithms suggests the conclusion that the message algorithm creates a tighter cluster of keywords, just the opposite of the effect it has on the TAS, i.e. creating a more diffuse overall network. A tighter cluster of keywords is desirable because it means that the relatedness of the keywords is being discovered to a greater degree (Tables 5 and 6).

[4] Smaller spread means tighter clustering.

4.2.4. Stop words

The exclusion of noise is a desirable feature of a text analysis system and is usually accomplished by the use of stop word lists. The words in these lists can be considered noise in the information processing paradigm and will be referred to as "noise strings". The number of noise strings based on the stop list found in

Table 9
Twenty most highly activated words for each probe word in the two self-organizing algorithms showing probe in italic, noise in bold and total number of noise strings web-citation set

Sentence	Message	Sentence	Message	Sentence	Message
citation	*citation*	*counts*	*counts*	*engine*	*engine*
the	engines	**the**	thelwall	**the**	alltheweb
web	more	**and**	alltheweb	search	thelwall
and	thelwall	web	**following**	**that**	**particular**
that	alltheweb	**that**	search	**and**	**following**
for	**also**	citation	**give**	**use**	informationr
this	following	**for**	engines	web	**prof**
not	informationr	**not**	different	**this**	study
about	counts	**with**	particular	**for**	net
more	search	**this**	examples	**with**	examples
with	prof	search	questionable	**also**	questionable
can	particular	**are**	**wanted**	**which**	command
their	promotion	**can**	reliability	**same**	reliability
use	net	**should**	informationr	altavista	wanted
citations	**as**	**use**	**also**	provides	hence
from	**did**	more	**prof**	**not**	aids
have	searching	**about**	**ditto**	**can**	numbers
are	**give**	**from**	**said**	citation	–
pages	**all**	impact	mining	results	**finding**
impact	command	links	**couple**	**using**	**ditto**
some	examples	link	command	**was**	**said**
15	5	13	8	14	7
links	*links*	*search*	*search*	*web*	*web*
the	alltheweb	**the**	alltheweb	**the**	engines
and	**thelwall**	**and**	engines	**and**	thelwall
web	following	**that**	**thelwall**	**that**	alltheweb
that	particular	web	**did**	**for**	**also**
for	informationr	engines	**following**	citation	**did**
this	prof	**for**	**give**	**this**	**the**
can	search	**use**	informationr	**not**	search
not	**did**	**with**	prof	citations	following
citation	counts	**this**	counts	**about**	informationr
from	command	**not**	**also**	more	more
with	net	results	particular	**are**	counts
are	engines	citation	net	pages	prof
use	**give**	**are**	examples	**use**	particular
counts	**hence**	**from**	questionable	**with**	promotion
about	examples	**can**	**wanted**	**from**	**all**
more	questionable	**did**	reliability	**their**	net
their	wanted	**about**	different	**can**	**and**
search	reliability	more	command	**than**	study
pages	numbers	**one**	**produced**	**have**	different
research	–	**than**	**hence**	search	**as**
13	5	15	8	15	6

Fox (1989) in the 20 most highly activated nodes for every case was tabulated (Tables 9 and 10). In every case, the message algorithm included fewer noise strings in the 20 most highly activated nodes. The mean number of noise strings was 14.7 (15.8) in the sentence algorithm and 6.5 (7.4) in the message algorithm.

Table 10
Twenty most highly activated terms for each probe term and self-organizing treatment showing probe in italic, noise in bold and total number of noise strings writing set

Sentence	Message	Sentence	Message	Sentence	Message
english	*english*	*papers*	*papers*	*student*	*student*
the	**what**	**and**	**done**	**the**	instructor
writing	instructor	**the**	**maybe**	**and**	mary
and	mary	**that**	colleagues	**that**	**hold**
students	**should**	writing	median	**not**	**what**
that	different	**for**	age	**for**	different
but	**hold**	**not**	bosses	writing	**yes**
not	**their**	students	northern	students	**tried**
was	**through**	**one**	iowa	**was**	layers
are	mls	**was**	mexico	**but**	educate
for	**kinds**	**are**	submitted	**are**	basics
this	items	**when**	poorly	work	awareness
who	internalization	**but**	grouchy	**with**	internalization
their	awareness	**who**	**obvious**	paper	7-jan
to	7-jan	**with**	joke	**who**	**kinds**
with	**tried**	student	anything	**their**	items
people	layers	work	decreased	**two**	fortunately
many	educate	**their**	**ten**	**this**	**felt**
when	basics	**have**	**********	**one**	degree
have	terms	**this**	wave	**like**	teacher
there	**saw**	**what**	crested	papers	**saw**
16	8	16	5	15	7
students	*students*	*write*	*write*		
the	mary	students	mary		
and	**what**	**the**	instructor		
writing	**should**	**and**	**hold**		
that	instructor	**for**	different		
not	**hold**	**but**	**what**		
are	**their**	**are**	**yes**		
for	mls	**that**	tried		
their	**the**	writing	layers		
but	different	**not**	educate		
with	**and**	**to**	basics		
was	**good**	**who**	awareness		
from	**you**	english	internalization		
who	**through**	**was**	7-jan		
english	**tried**	learn	**kinds**		
work	layers	**had**	**items**		
have	educate	**like**	fortunately		
write	basics	**this**	**felt**		
all	awareness	**all**	degree		
were	internalization	**our**	teacher		
to	7-jan	**with**	**saw**		
16	10	16	7		

5. Conclusion

This study has explored the effect of modifying the way external activation levels of intra-message words are reduced between sentences during the learning phase of an interactive activation with competition artificial neural network (IAC ANN) processing email text. The results from this initial investigation suggest that IAC ANN analysis of email text is improved by the use of message context during learning. Message context can be implemented by allowing the external activations of the words already processed from the current message to decay gradually rather than go to zero at the end of each learning cycle. The results demonstrate that by employing a context oriented model several benefits may be realized; keywords are more closely related; better discrimination between related and unrelated words is achieved and; high frequency "stop words" are effectively ignored.

6. Future research

This research has explored a potential method of improving the performance of an interactive activation with competition artificial neural network in the analysis of email text. The development of this technology for text 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 fully realized the method needs to be tested with other sets of messages and 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. For instance, words encountered in short messages might warrant a greater effect on the network than words in long messages. Words at the beginning (introduction) and/or end (summary/conclusion) of a message may warrant greater weighting.

Two additional 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 limiting the activation level of words that are frequently encountered and desensitizing noise word nodes by altering their activation function.

Disambiguation is another hotly debated topic in the field of automated text analysis. After sufficient text has been processed it may possible that the network will be able to discriminate between homonyms such as "banks" of a river and savings "banks".

The model presented here treats all sentences equally, yet that does not reflect how humans communicate (Liddy, 2004). 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. Yet another discourse-level concept that should be researched is sequence. In other words, should words 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?

The purpose of this research was to test the essential viability of this novel method to reduce the need for preprocessing of text before analysis by ANN. Future trials will test the basic model with other sets of messages and explore the effects of altering network parameters that were not manipulated in this study.

Appendix A. Catpac™ output

Catpac™ lists—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

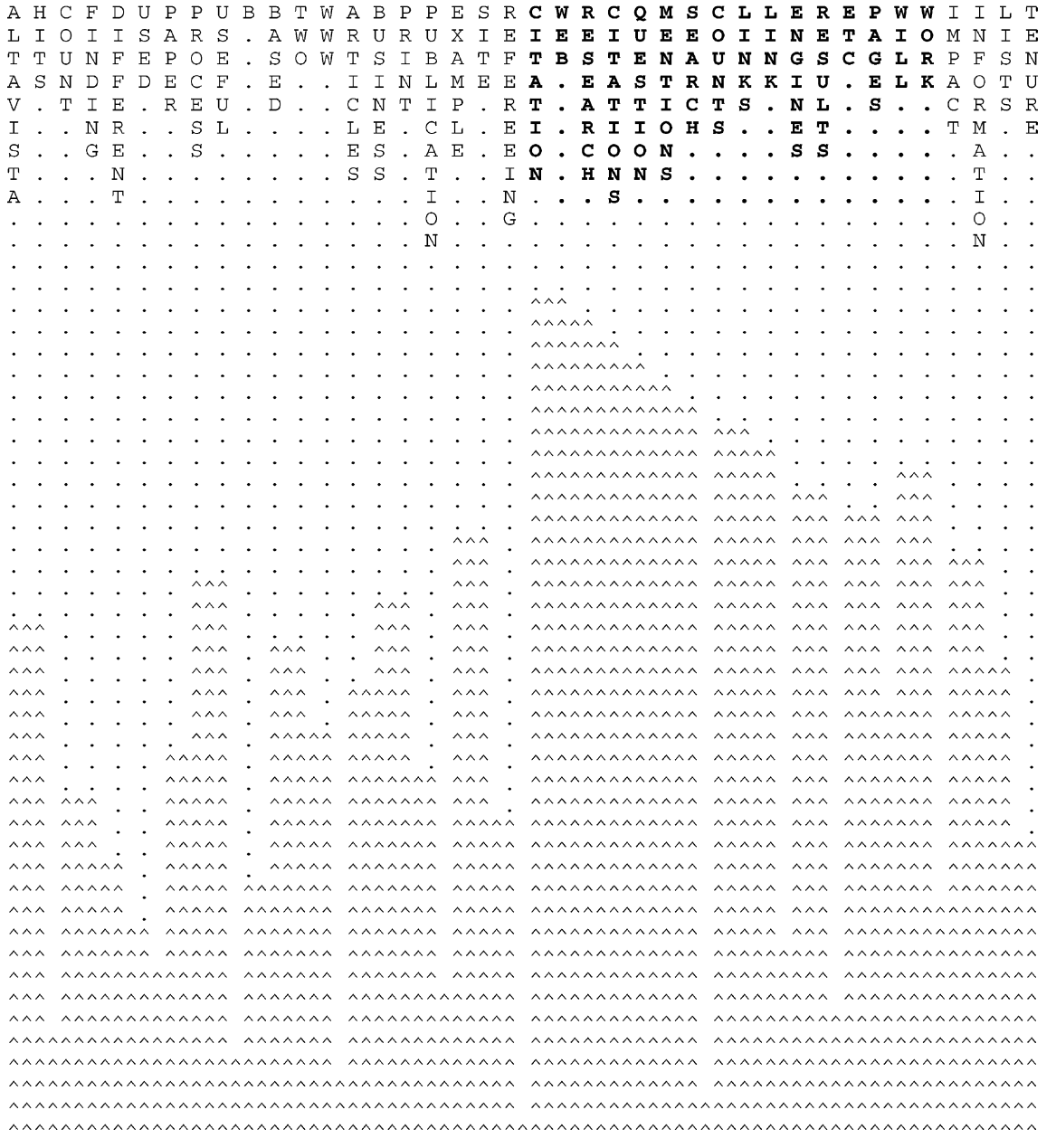
WORD	FREQ	PCNT	CASE FREQ	CASE PCNT
WEB	71	14.5	326	67.5
CITATION	27	5.5	155	32.1
SEARCH	27	5.5	124	25.7
CITATIONS	22	4.5	116	24.0
LINK	17	3.5	81	16.8
ENGINES	16	3.3	78	16.1
INFORMATION	15	3.1	87	18.0
PAGES	13	2.7	81	16.8
COUNTS	12	2.5	76	15.7
IMPACT	12	2.5	67	13.9
LINKS	12	2.5	74	15.3
PRINT	12	2.5	55	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

WARDS METHOD

Catpac™ dendrogram—web-citation set



In the dendrogram above the main cluster words have been formatted in **bold** type by the author, not by Catpac™.

Appendix B. Catpac™ output

Catpac™ lists—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

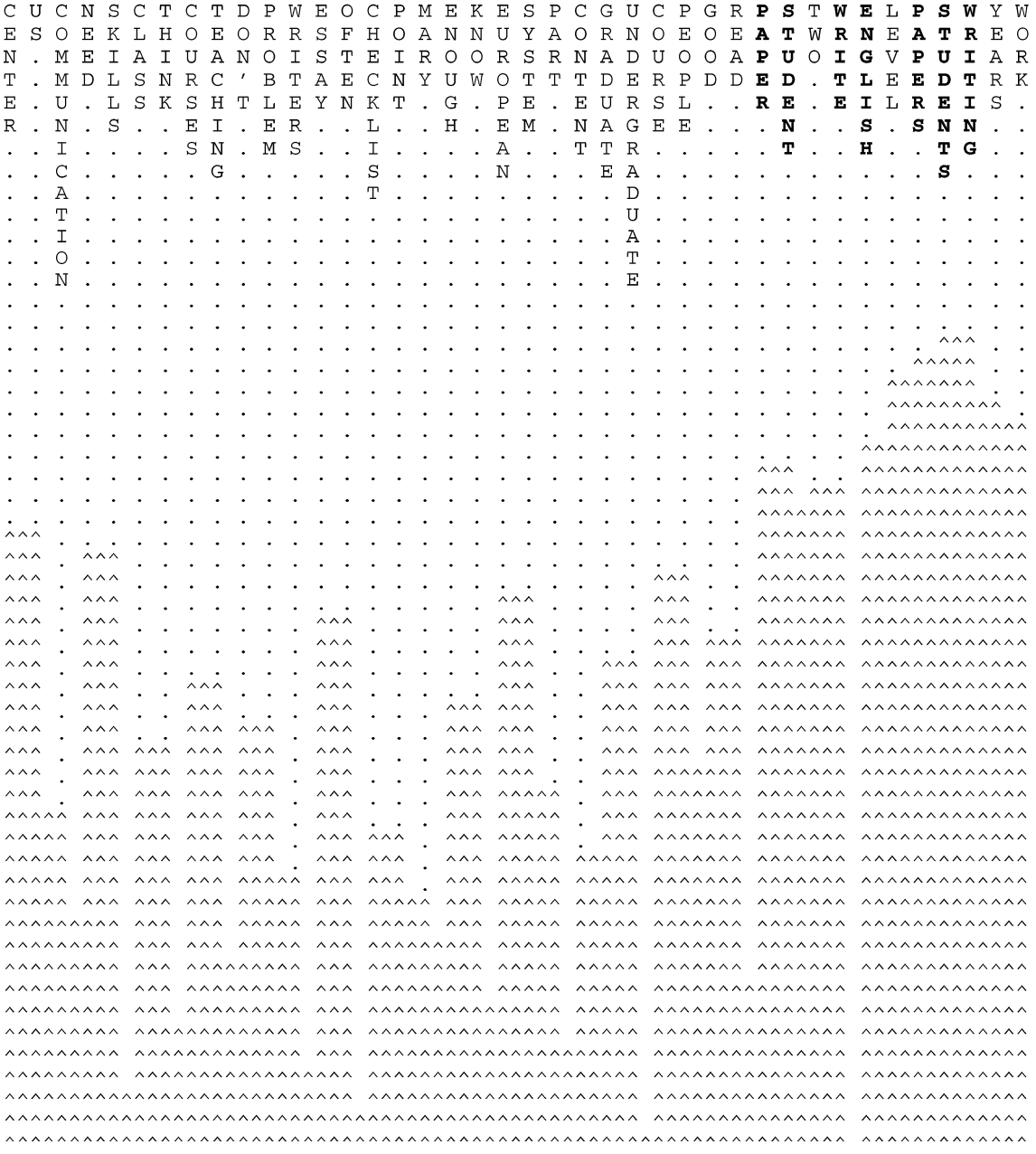
WORD	FREQ	PCNT	CASE FREQ	CASE PCNT
WRITING	55	13.3	257	63.3
STUDENTS	46	11.2	228	56.2
ENGLISH	19	4.6	92	22.7
WORK	18	4.4	108	26.6
PAPERS	16	3.9	83	20.4
STUDENT	13	3.2	81	20.0
TWO	13	3.2	79	19.5
COURSE	11	2.7	65	16.0
PAPER	11	2.7	66	16.3
PEOPLE	11	2.7	62	15.3
US	11	2.7	60	14.8
WRITE	10	2.4	64	15.8
GOOD	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

WARDS METHOD

Catpac™ dendrogram—writing set



In the dendrogram above the main cluster words have been formatted in **bold** type by the author, not by Catpac™.

References

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

Aizawa, A. (2002). A method of cluster-based indexing of textual data. In Paper presented at *The 19th international conference on computational linguistics (COLING 2002)*.

Atiya, A. F. (1990). An unsupervised learning technique for artificial neural networks. *Neural Networks, 3*(6), 707–711.

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

Belew, R. K. (1996). Adaptive information retrieval: machine learning in associative networks. Unpublished PhD, University of Michigan, Ann Arbor, MI.

Breen, M. J. (1997). Agenda setting and public opinion formation: media content and opinion polls on divorce referenda in Ireland. Unpublished PhD, Syracuse University, Syracuse, NY.

Danielson, W. A., & Lasorsa, C. L. (1997). Perceptions of social change: 100 years of front-page content in The New York Times and The Los Angeles Times. 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. 103–115). Mahwah NJ: Lawrence Erlbaum.

Dumais, S. T., Furnas, G. W., Landauer, T. K., Deerwester, S., & Harshman, R. (1988). Using latent semantic analysis to improve access to textual information. In Paper presented at *The SIGCHI conference on human factors in computing systems, Washington, DC*.

Dworman, G. (1996). Homer: a pattern discovery support system. In Paper presented at *The ACM SIGCHI conference on human factors in computing systems*.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.

Fox, C. (1989). A stop list for general text. *SIGIR Forum, 24*(1–2), 19–21.

Franzosi, R. (1997). Labor unrest in the Italitan service sector: an application of semantic grammars. 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. 131–145). Mahwah NJ: Lawrence Erlbaum.

Garson, G. D. (1998). *Neural networks: An introductory guide for social scientists* (first ed.). London: Sage Publications, Ltd.

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Boston: Kluwer Academic Publishers.

Hearst, M. A. (1999). Untangling text data mining. In Paper presented at *The 47th annual meeting of the association for computational linguistics*, June 20–26 University of Maryland.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In Paper presented at *The 22nd annual international ACM SIGIR conference on research and development in information retrieval, Berkeley, CA*.

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.

Jorgensen, P. E. (2003). Continuous analysis of internet text by artificial neural network. Unpublished PhD, SUNY at Buffalo, Buffalo, NY.

Kennedy, S. D. (2001). Alphabet soup: an acronym roundup—global e-commerce has inundated us with many new abbreviations. *Information Today (United States), 18*(7), 28–30.

Klerfors, D., & Huston, T. L. (1998). Artificial neural networks. Retrieved October 1, 2001, from: http://hem.hj.se/~de96klda/NeuralNetworks.htm.

L-Soft (2002). 1 Jun 2002. JESSE@LISTSERV.UTK.EDU. Retrieved June 1, 2002, from: http://www.lsoft.com/scripts/wl.exe?SL1=JESSE&H=LISTSERV.UTK.EDU.

Lasswell, H. D. (1931). The measurement of public opinion. *The American Political Science Review, 25*(2), 311–326.

Liddy, E. D. (2001). A breadth of NLP applications. *ELSNews, 10*(4), 10–12.

Liddy, E. D. (2004). July 25–29. Context-based question–answering evaluation. In Paper presented at *The SIGIR 2004, Sheffield, UK*.

Lyman, P., & Varian, H. R. (2000). Thu, Mar 7, 2002 12:46:07 PM. How much information. Retrieved on April 18, 2002, from: http://www.sims.berkeley.edu/research/projects/how-much-info/internet.html#4.

Mani, I. (1999). *Advances in automatic text summarization*. Cambridge, MA: MIT Press.

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, S. J. (2000). The microscope and the moving target: the challenge of applying content analysis to the World Wide Web. *Journalism and Mass Communication Quarterly, 77*(1), 80–98.

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 and Co.

Ohgaya, R., Simmura, A., & Takagi, T. (2003). Meiji University web and novelty track experiments at TREC 2003. In Paper presented at *The twelfth text retrieval conference, Gaithersburg, MD*.

Palmquist, R. A. (2002.) Content analysis. Retrieved on April 7, 2002, from: http://www.gslis.utexas.edu/~palmquis/courses/content.html.

Pasquale, J.-F.d., & Meunier, J.-G. (2001). Categorisation techniques in computer assisted reading and analysis of texts (CARAT) in the humanities. In 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 June 16, 2001*.

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

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.

Shi-Xu (2000). Opinion discourse: investigating the paradoxical nature of the text and talk of opinions. *Research on Language and Social Interaction, 33*(3), 263–289.

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

Wilbur, W. J., & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of Information Science, 18*(1), 45–55.

Witten, I. H. (2001). Adaptive text mining: inferring structure from sequences. Retrieved on May 31, 2002, from: http://citeseer.nj.nec. com/rd/42058339%2C502339%2C1%2C0.25%2CDownload/http%3AqSqqSqwww.cs.waikato.ac.nzqSq%7EihwqSqpapersqSq01IHW-Ad-aptivetextmining.pdf.

Woelfel, J., & Styanoff, N. J. (1993). CATPAC™: A neural network for qualitative analysis of text. In Paper presented at *The Australian Marketing Association, Melbourne, Australia*.