THE ORGANIZATION OF INTERNET WEB

PAGES USING WORDNET AND

SELF-ORGANIZING MAPS

The members of the Committee approve the masters thesis of Darin Brezeale

Diane J. Cook
Supervising Professor _____

Lawrence B. Holder _____

Lynn Peterson _____

THE ORGANIZATION OF INTERNET WEB PAGES USING WORDNET AND SELF-ORGANIZING MAPS

by

DARIN BREZEALE

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

THE UNIVERSITY OF TEXAS AT ARLINGTON

August 1999

ACKNOWLEDGMENTS

I wish to thank my thesis adviser, Dr. Diane Cook, for her support and guidance. Dr. Cook consistently provided positive feedback and helped me to stay on track. This thesis is better because of her input.

July 15, 1999

ABSTRACT

THE ORGANIZATION OF INTERNET WEB
PAGES USING WORDNET AND
SELF-ORGANIZING MAPS

Publication No. _____

Darin Brezeale, M.S.

The University of Texas at Arlington, 1999

Supervising Professor: Diane J. Cook

With the Internet increasing in size at a rapid rate, locating information is becoming more difficult. Many people use traditional search engines, such as Altavista, to locate information, but they find that these search engines return links to many irrelevant sites. Alternative search engines which effectively organize web sites perform this task through human intervention. As the Internet grows in size, this manual organization process will become increasingly more difficult to perform.

The approach that we have taken is to use machine learning techniques to automate the organization of web pages. Self-organizing maps have been used previously to organize web pages represented as vectors. We believe that by using WordNet, an electronic lexicon, we can take advantage of the relationships that exist between words to improve the vector representations of the web pages, thereby improving the organizational process performed by the self-organizing map.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

1.1 Motivation

As the Internet grows in size, it is becoming increasingly more difficult to locate the exact information that you seek. Traditional search engines perform keyword searches and as a result they produce huge lists of sites containing those keywords. The site that contains the information a user desires may be on this list, but so may thousands of irrelevant sites as well. A user can reduce the number of web page links returned by the search engine by requiring that all of the search terms be in each web page, but this will exclude any related web pages that may not contain all of the terms.

Some search engines, such as Yahoo, produce lists of related web sites, but this requires indexing by humans [Taulli99]. As the number of web pages grows, this manual task will become much more difficult, if not impossible. The Internet contains an estimated 800 million web pages, but the most comprehensive search engines index at most 16% of the web pages [Dunn99]. In addition, traditional search engines do not return web pages that contain synonymous terms in place of the terms used in the search query. For example, using a traditional search engine to find web pages containing the word ‘car’ will not find those web pages that always use the word ‘automobile’ instead. Also, web pages that represent the concept being sought may not include the search terms. An example is using the search term ‘astronomy’ but not having the search engine display those web pages about the solar system because they don’t contain the word ‘astronomy’. One solution to these

problems is to use machine learning techniques to automatically organize and group related web pages.

1.2 Hypothesis

Some work with favorable results has already been performed using a self-organizing map (SOM), a type of neural network, to automate the task of classifying web pages [Chen96]. Typically, when SOMs are used to classify documents, a vector representation of the document or web page is used as input data for the SOM. These vectors show the presence or absence of the words contained in the document or web page.

Our belief is that the electronic lexicon WordNet can be used to exploit the relationships that exist between words by altering the term vectors representing the web pages in order to improve the results produced by the SOM. Looking back at the example from section 1.1, if one web page contains the term ‘car’ throughout and another web page contains the term ‘automobile’ throughout, then WordNet may decide that the two terms are both types of vehicles and use the term ‘vehicle’ in the term vectors for both web pages instead of the actual terms of ‘car’ and ‘automobile’.

Traditional search engines find web pages that share some or all of the terms given in the search query regardless of context. Instead of grouping web pages together because they share a few words, it is the intent of this work to use a combination of WordNet and a SOM to group web pages that are conceptually similar, and therefore closer to the concept that the user seeks. By organizing web pages together that represent the same concept, web pages can be found that do not contain all of the search terms. Web pages containing synonymous terms (such as ‘car’ and ‘automobile’) could also be found this way, even if they do not contain any of the search terms. Also, web pages containing the search terms but representing a much different concept can be avoided.

Using a machine learning approach to organizing web pages into related groups has a number of benefits. These include:

1. Use of the SOM for classification does not require prior knowledge of the number of categories as some classification methods do.
2. New web pages can be added to existing categories in manually indexed search engines, such as Yahoo.
3. Web pages related to the web page currently being viewed by a user can be prefetched.
4. Creating a map or hierarchy of web pages allows a user to look for information without knowing exactly how the information is classified. See figure 1.

1.3 Contributions

This work has produced a number of contributions:

a) WordNet was used to alter the vector representations of web pages.
b) A combination of WordNet and a SOM was used to organize web pages.
c) Vectors with reduced size were shown to be as effective as the original vectors in some cases.
d) Testing was performed to determine the validity of this approach.

1.4 Outline

In chapter 2 we describe the data used for the experiments and how it was obtained. Next, in chapters 3 and 4 details are given about the use of WordNet and SOMs, respectively. The experimental results are described in chapter 5. Our conclusions and thoughts for future work are given in chapter 6.

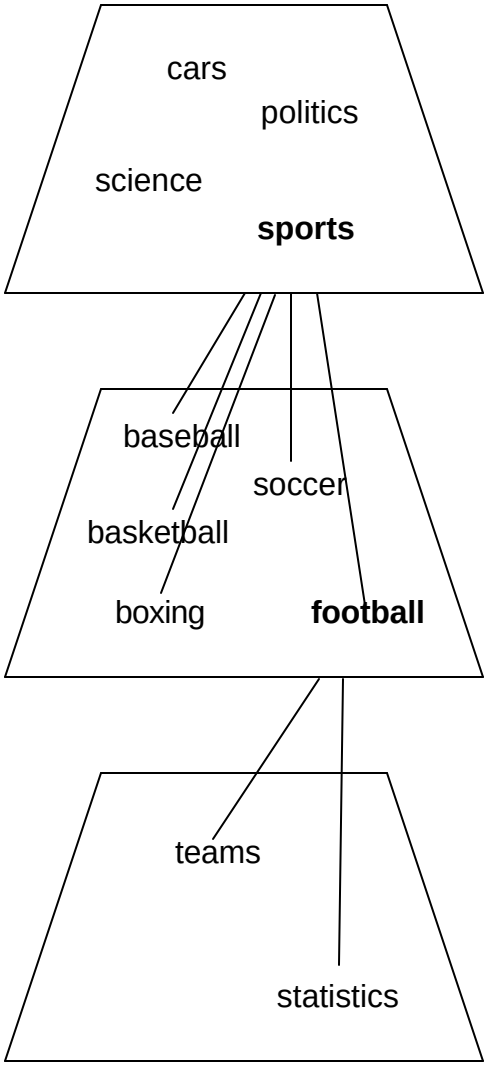


Figure 1. Example of map hierarchy.

CHAPTER II

THE DATA

In this chapter the data used for this thesis is discussed. The choice of data, why it was chosen, and how it was gathered is detailed. This is followed by a discussion of the classification and preparation of the data.

2.1 Choice of Data

When users wish to locate web pages containing information of interest, typically they access a traditional search engine and, after the user has entered several terms that represent a concept, the search engine returns a list of links to web pages containing some or all of the search terms. If the user sets up the search query so that all web pages with any of the search terms are to be found, the resulting list may contain tens of thousands of web pages. If the search query indicates that only links to those web pages containing all of the search terms are to be returned, then the list will most likely be much shorter but could still contain many links that are unrelated to what the user is seeking.

It was decided to use a traditional search engine to find web pages. These web pages were then classified by the author before being classified by the SOM. Terms representing ten concepts were used. Those terms, as well as the concept they were intended to represent, are shown in table 1.

Table 1. Search terms for web pages

1.	apache helicopter military – the Apache helicopter used by the U.S. military
2.	census bureau statistics – census statistics
3.	cortisol memory loss – the hormone cortisol and its relationship to memory loss
4.	fantasy football rules – the rules for playing fantasy football
5.	fermats last theorem – information about Fermat's last theorem
6.	growing fruit trees – the raising of fruit trees
7.	nearest neighbor classifier – a type of classifier used in machine learning
8.	performing clustering analysis – how to analyze clustering methods
9.	solar system planets – a collection of planets that orbit a star
10.	wolfgang amadeus mozart biography – a biography of the composer Mozart

2.2 Gathering the Web pages

The process of gathering the web pages was manual for this project in order to allow the author to gather suitable web pages. Because it was known in advance that the number of web pages being used for the experimentation would be small, it was desirous that web pages be chosen that could easily be classified into a small number of groups. The web pages used for the experimentation were located by using the well-known search engine Altavista (http://www.altavista.com).

Only web pages containing all of the search terms were used in order to improve the results of the search engine. For each of the ten concepts being sought, the first ten web pages returned by the search engine that were suitable for this experimentation were used, for a total of one hundred web pages. By suitable, what is meant is that the web page had to contain some text, the web page was not a duplicated link of a previously chosen web page, and the web page was not from the same site as a previously chosen web page. In all cases, the ten web pages chosen came from the first thirty links displayed by the Altavista search engine.

The online help available for the Altavista search engine states that based on some heuristic, those links returned by the search engine are ordered, with the most relevant links first. Therefore, those web pages that best represent the search terms are displayed at the top of the list [Alta99].

2.3 Classifying the Data

Once the web pages had been gathered, it was necessary to classify them. Ten concepts were chosen and then ten web pages representing each concept were retrieved using a search engine. If the search engine retrieved web pages that exactly matched the desired concepts, then the results would be one hundred web pages that could be classified into ten categories, with ten web pages per category.

The author viewed each of the one hundred web pages and classified them. Many of the web pages could be classified by the concept that was used to retrieve them, but others could not. For example, when using the query ‘Apache helicopter military’ to represent the concept ‘the Apache helicopter used by the U.S. military,’ several web pages that did not represent the desired concept were retrieved. One page was of military artwork that the author classified as ‘art’, another web page was of models constructed out of Lego blocks, which the author classified as ‘toy’. Instead of retrieving only web pages that correctly matched the desired concept, it was decided to retrieve the best web pages that the search engine returned and allow the SOM to determine if they should not really be classified together just as the author had done. Ultimately, the one hundred web pages fit into thirty-seven categories; many of the categories contained only a single web page. These categories are shown in appendix B.

2.4 Preparing the Data

Before the web pages could be analyzed, they had to be preprocessed. On a large scale, this process would need to be automated. Preprocessing was necessary to reduce the amount of work done by the WordNet program and the SOM and also to prevent confusing the WordNet program with punctuation, non-alphabetic characters, and so forth.

The first step was the removal of such things as the HTML tags, graphics, menus, and so forth. This process was performed manually. The next step, which was automated by using a Java program, was to gather the first twenty lines of each web page. This step greatly reduced the number of terms to be analyzed in each web page. Finally, several Perl scripts were used to filter out terms that would not be helpful, such as single characters, prepositions, articles, forms of the verb ‘be’, and so forth. The punctuation was also removed.

CHAPTER III

WORDNET DETAILS

Using WordNet was a major component of the work done for this project. This chapter provides details of the use of WordNet in this project.

3.1 Background

The WordNet project was started in 1985 by George Miller [Miller93]. Initially the intent of the WordNet project was to design a system that allowed dictionaries to be searched conceptually. WordNet currently divides words into the categories of nouns, verbs, adjectives, and adverbs. WordNet attempts to organize information according to the meanings of the words instead of the forms of the words.

WordNet contains the standard information found in dictionaries and thesauri. An additional feature of WordNet is its information about the relationships between words; the most important of these for this thesis is that of the hypernym. A hypernym of a term is a more general term that fits the statement "___ is a kind of ___." For example, a dog is a kind of canine, so canine is a hypernym of dog. A canine is a kind of mammal, so mammal is a hypernym of canine. This continues until some top, very general, term is reached; in this case, the top hypernym for dog and canine is entity. These terms, related through the concept of hypernymy, form a hypernym tree, which will be used later in this thesis. The inverse relationship is called hyponymy. The hypernym trees for the words 'orange', 'apple', and 'potato' have been combined and are shown in figure 2.

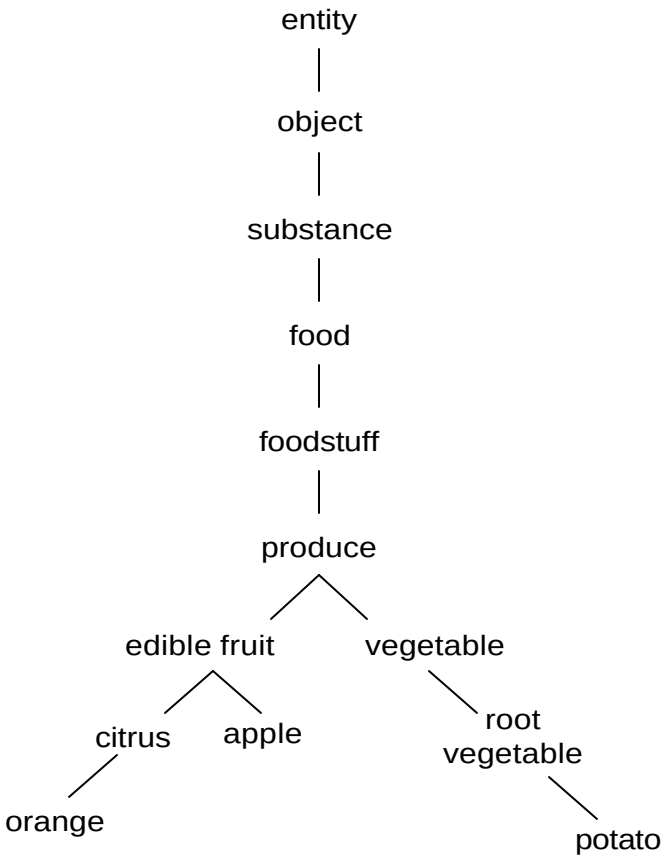


Figure 2. Example hypernym tree.

3.2 Identification of Semantically Related Terms

The identification of semantically related terms, those that are related in meaning, can be accomplished in a variety of ways. This variety is largely dependent on how tight or loose the relationship must be. For example, in the sentence “His house is big” the word ‘big‘ can be replaced with the word ‘large’ with no loss in meaning because these two words are synonymous. This can be considered a tight relationship. However, in the sentence “I ate an orange” the word ‘orange’ cannot be replaced with the word ‘apple’ without altering the

meaning of the sentence. However, in some situations, documents (or web pages) about apples or oranges may be grouped together under the classification ‘fruit’. In this case, this might be considered a loose relationship.

WordNet provides several ways to identify related terms, including synsets, hypernyms, and coordinate terms. Synsets are sets of synonyms. A word can be part of several different synsets, each representing a different context. Hypernyms, as previously defined, are words that are more general forms of a word that fit the statement “___ is a kind of ___.” A hypernym for ‘oak’ is the word ‘tree’. Coordinate terms are those terms that are on the same level, one level below the same hypernym. They can be considered as sibling nodes since they share the same parent. The coordinate terms for the first sense of the word ‘dog’ are: bitch, dog, wolf, jackal, wild dog, hyena, and fox since they are all types of canines.

Each of the previously described relationships has its strong and weak points. The use of synsets has the advantage that terms are being replaced with another term that is considered interchangeable within the correct context. However, since context is not known, it will be necessary when using this approach to make some assumptions, such as the first sense of the word will be used. The disadvantage of using synsets is that it may be considered too limiting in many cases. In the example using the terms ‘orange’ and ‘apple’, many people might be comfortable grouping together web pages about these terms under the classification ‘fruit’ even though the two terms cannot be considered synonymous. If synonyms of synonyms are allowed in order to expand the number of interchangeable terms, then words that are not related may be considered interchangeable. For example, WordNet shows that a synonym of "banana" is "fruit" and a synonym of "fruit" is "yield.” "Banana" and "yield" are not closely related. Another potential problem with this approach is that by

trying to find the synonyms of synonyms of synonyms and so forth, you create an exponential search and ultimately all words will be considered related if enough synonyms are checked.

Coordinate terms share some of the strengths and weaknesses as synsets. It is easy to see the relationship that exists between coordinate terms. From the previous example, ‘dog’ and ‘wolf’ are both recognized as canines and for the sake of clustering it would be easy to group web pages about either of these canines together, at least in a general sense. However, because WordNet is so detailed in its construction of hypernym trees and has so many levels, some terms that humans might believe to be coordinate terms are not. For example, a partial list (there are seventy-five coordinate terms) of the coordinate terms for the first sense of ‘apple’ is: apple, citrus, berry, apricot, peach, fig, plum, and grape. The term ‘orange’, however, does not appear on the list because it is a hyponym of citrus and therefore not a coordinate term for ‘apple’. ‘Citrus’ and ‘apple’ are the coordinate terms here even though humans might feel that ‘orange’ and ‘apple’ should be at the same level.

The hypernym relationship has its own strengths and weaknesses. Its strengths, though, were considered greater than those of the two previously described relationships and therefore it was the relationship used in this thesis. It is described in detail in the next section.

3.3 Hypernyms

When using hypernyms to determine if two words are semantically similar, the hypernym tree of each word must be constructed. Once the trees have been constructed, it is necessary to navigate up each tree to see if they share a common ancestor. By counting the distance in words from each of the initial words to the common hypernym ancestor and summing these two counts, a distance measure can be calculated [Hadd98]. This distance

can be viewed as a measure of the semantic similarity of the two words. A short distance indicates that the two words are more closely related than two words separated by a longer distance.

WordNet has twenty-five categories of nouns [Miller98]; these are shown in table 2. This means that many words that are similar only in the loosest sense are in the same category and therefore share a common hypernym ancestor. For example, the words ‘lock’ and ‘elephant’ are both in the same category; ‘entity’ is the shared hypernym for both.

Table 2. Noun categories in WordNet

{act, activity}	{animal, fauna}	{artifact}
{attribute}	{body}	{cognition, knowledge}
{communication}	{event, happening}	{feeling, emotion}
{food}	{group, grouping}	{location}
{motivation, motive}	{natural object}	{natural phenomenon}
{person, human being}	{plant, flora}	{possession}
{process}	{quantity, amount}	{relation}
{shape}	{state}	{substance}
{time}		

In order to tighten the semantic similarity of two words, a distance threshold is used. Words that share a common hypernym ancestor but have a distance value greater than the threshold are not considered to be semantically similar. In the example shown in figure 3,

the distance from ‘orange’ to ‘edible_fruit’ is 2 and the distance from ‘apple’ to ‘edible_fruit’ is 1. The total semantic distance from ‘orange’ to ‘apple’ is therefore 3. If this semantic distance of 3 is less than or equal to the threshold for determining semantic similarity, then the words ‘apple’ and ‘orange’ can be replaced by the common hypernym ancestor ‘edible_fruit’.

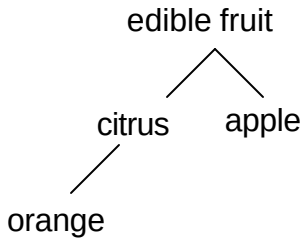


Figure 3. Partial hypernym tree for orange and apple.

3.4 The Algorithm

One assumption that was made about how users locate information using traditional search engines is that nouns are typically used to describe the concept being sought. The English language contains many more nouns than verbs. As described in [Fell90], the *Collins English Dictionary* contains 43,636 different nouns and 14,190 different verbs. The nouns have fewer senses as well. Because verbs have more senses than nouns on average, the meanings of verbs are more flexible than the meanings of nouns, that is, the meaning of a verb is much more dependent on the kinds of nouns in the sentence with it than nouns are on the kinds of verbs present [Fell90]. Because of this, we believe that using nouns only will be sufficient to represent a web page.

Because only nouns will be used to represent a web page, ideally, a program such as the Brill tagger [Scott98] could be used to identify the nouns in a web page so that only the

nouns would be fed to WordNet. Since incorporating the Brill tagger into this work was beyond the scope of this thesis, WordNet was used to determine if a word was a noun and, if so, it returned the hypernym tree of the word. This method will not be entirely accurate if a term that is a noun is not found in WordNet or if the term can be both a noun and a verb, such as “spring.” In the case that a word can be both a noun and a verb, we assume that the word is a noun. WordNet (version 1.6) contains a substantial number of nouns: 94,474 unique strings in 66,025 synsets, with a total of 116,317 senses.

Each term in each web page is read and, if WordNet determines that it is a noun, the term is added to a master list of terms (assuming it has not been added already). The master list of terms contains all of the distinct noun terms in all of the web pages. In addition, the hypernym tree for the term is stored as well as the level of the input term in the hypernym tree and the root hypernym.

Once all of the web pages have been read, it is time to determine what the replacement terms are for the words in the master list. The replacement terms are the ancestor hypernym terms of a word that will be used in place of the actual word. By default, the replacement for each term is itself; the replacement only changes if a shared hypernym is found. Each term in the master list is then compared to the other terms to determine if they share a common hypernym term. If the two master terms being compared share a hypernym term and the distance between them is within the threshold, then the replacement term for each of the master terms will be the common hypernym ancestor of each word. For example, if the words ‘orange’ and ‘apple’ are on the master list of terms (those that actually appear in a web page), the partial hypernym tree for these terms will be constructed as in figure 4. For this example, assume the threshold for semantic similarity is four. The semantic distance from ‘orange’ to ‘apple’ is three, which is within the threshold. The term ‘edible_fruit’ will

be used as a replacement for each. If, however, the words being compared are ‘orange’ and ‘potato’, we find that the semantic distance is six. This is beyond the threshold and the two are not considered semantically similar. The replacement term for each remains the same. Had the terms being compared been ‘apple’ and ‘produce’, ‘produce’ would have become the replacement term for itself and ‘apple’ since ‘produce’ is a hypernym of ‘apple’.

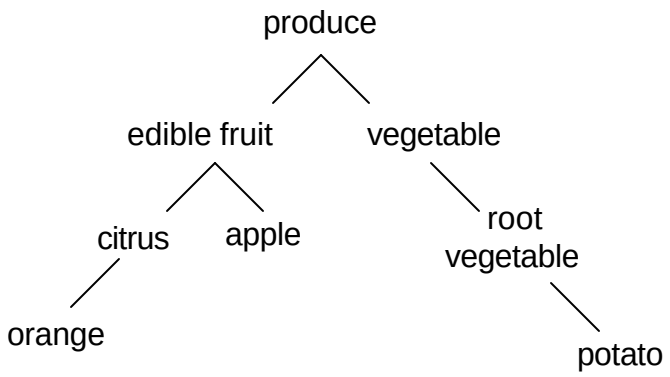


Figure 4. Partial hypernym tree for orange, apple, and potato.

In order to reduce the computational complexity of the search, two heuristics are used. Because WordNet has twenty-five categories of nouns, only terms that have the same root hypernym term are compared since terms that are in different noun categories cannot possibly share a common hypernym term. The second heuristic used is that the levels of the input terms are checked and, if the difference between them is greater than the threshold, then there is no reason to check the hypernym trees for each term. For example, the word ‘food’ is at level three and the word ‘orange’ is at level eight. Without even constructing the hypernym trees for each word, we can determine that even if the two words happen to share a common hypernym ancestor the closest semantic distance between the two words will be

five, which is beyond the threshold. It so happens that ‘food’ is a hypernym of ‘orange’ and the semantic distance between them is in fact five. This can be seen back in figure 2.

After the replacement terms have been determined, a list of the replacement terms is created and a binary flag is set to zero for each term as the default. This list is used to create the feature vectors representing each web page. The size of the vectors will be equal to the number of replacement terms, which will be less than or equal to the number of terms on the master list. The feature vectors use the bag-of-words representation [Scott98] – each element of the feature vector represents one of the replacement terms. In this case, the values are binary; for each term in the original web page, the corresponding element in the feature vector will have a value of one. The elements in the feature vector that represent terms not found in the web page will have a value of zero. As an example (using the actual words instead of a binary representation), suppose we have two web pages composed of the following words: page1 = {apple, tree, box} and page2 = {orange, tree, man}. The master list of terms is then {apple, box, man, orange, tree} and the size of the feature vectors will be five. The vectors representing the two web pages would be page1 = [apple, box, *, *, tree] and page2 = [*, *, man, orange, tree], where the character * indicates a missing word. Looking at the hypernym trees of the words, it is discovered that ‘apple’ and ‘orange’ are related by the word ‘edible_fruit’ and the word ‘edible_fruit’ can be used as a replacement term. Using replacement terms, the list of terms is {box, edible_fruit, man, tree} and the vectors representing the two web pages would be {box, edible_fruit, *, tree} and {*, edible_fruit, man, tree}. The feature vectors with the replacement terms have a size of four.

Each web page is read a second time and for each term in the web page, its replacement term is found in the list and its binary value is set. When all of the terms in a web page have been read and checked, the vector representation of the web page is written

to a file as well as a label identifying that particular web page. This process is repeated for all of the web pages.

The algorithm can then be summed up as follows:

1. Read each web page, creating a master list of the distinct nouns present.
2. Use WordNet to determine the replacement terms.
3. Read the web pages again and generate the vector representation of each page.

The overall process, including the clean-up of the web pages and the use of the SOM, is shown in figure 5.

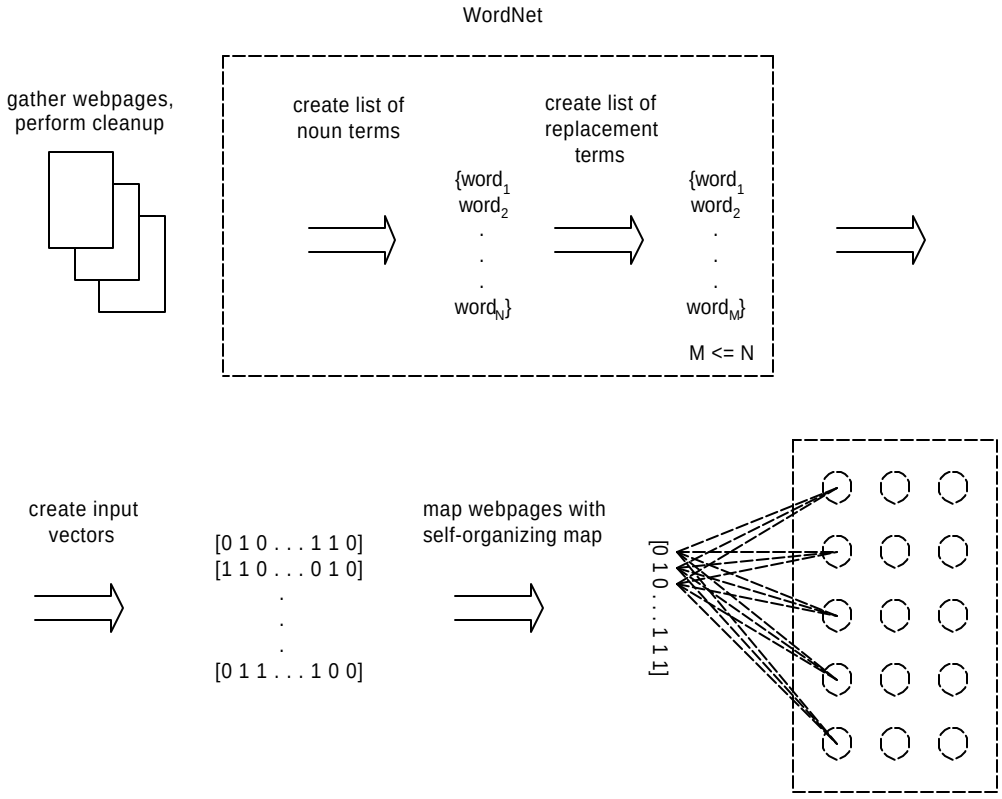


Figure 5. Overall process.

3.5 Related Work

Haddock used WordNet to calculate the semantic distance between terms in order to determine how similar the terms were [Hadd98]. This is the same approach used in this thesis. Using an alternative methodology known as the *hypernym density function*, Scott and Matwin used WordNet hypernyms to classify text [Scott98]. By using a combination of synonymy and hypernymy, the synsets that represent a document are determined in an automatic process. For each synset, the number of occurrences of the synset within the document is divided by the number of words in the document. This ratio is the hypernym density for this synset in this document. Synsets with higher density values are thought to be more representative of the document. Rodríguez et al. used WordNet to complement training information in text categorization [Rodrig97]. WordNet synsets are used to expand the set of terms that represent a category. Li et al. used WordNet to determine the sense of words in text for use in natural language processing [Li95]. In their work, the semantic similarity of two words is inversely proportional to the semantic distance between the words in the hypernym tree. Many additional WordNet-related papers can be found in [Rosen98].

3.6 Limitations of This Approach

This approach has several known limitations. First, as was stated earlier, the context of the words is not known and therefore the hypernym tree for the wrong context of a word may be used. In order to reduce this potential problem, the first context of a word is used, which represents the context with the greatest frequency of occurrence.

Next, this approach is not currently capable of identifying compound words, such as fountain pen. WordNet does support compound words, but they must be identified and queried as fountain_pen.

Finally, the layout of web pages is often quite different from traditional documents. Some web pages make heavy use of graphics in place of text, and these pages may not provide enough terms to accurately classify the page. Until the technology exists to classify images, this problem will remain. Also, web pages that represent a document are often divided into a number of different files that are linked together through the use of hyper links. Any particular page may not accurately represent the concept that the entire document represents. It should be possible to identify the files that are part of the same document through their use of embedded links to each other, and this entire collection used as if it were a single web page.

CHAPTER IV

SELF-ORGANIZING MAP DETAILS

In this chapter the self-organizing map is described. Details are given about its background, the general algorithm, and why it was used for this project. Also, the self-organizing map software is discussed.

4.1 Background

Research began on the self-organizing map (SOM) in 1981 by Teuvo Kohonen. It is a type of neural network that is particularly well-suited to clustering and visualization. [Koho95]. The SOM allows multidimensional inputs to be mapped to a two- or three-dimensional map. Because the algorithm maps similar vectors to the same node or to neighboring nodes, the SOM map can be used to visually identify clusters within the data.

One thing that separates the SOM from some other methods of clustering is that the SOM is an unsupervised algorithm [Lin91]. The data can be input into the SOM and the SOM will attempt to determine the clustering of the data. Other methods require the number of clusters be known in advance. This makes the SOM particularly well suited for this project since the number of categories describing web pages on the Internet is unknown but obviously large.

Another reason the SOM is well-suited to the task of organizing web pages is that the resulting map preserves the distance relationships between the input vectors. This means that input vectors that are similar, and therefore representative of similar web pages, will appear closer on the map than will input vectors that are not as similar [Lin91].

4.2 The Algorithm

WordNet generates vector representations of the web pages as input to the SOM. Each vector consists of ones and zeros and the size of the vectors is equal to the number of distinct replacement terms. The SOM used for this thesis produces a two-dimensional grid of nodes as output. Each node of this map has an associated reference vector of the same size as the input vectors. The value of each reference vector is randomly assigned during the initialization process.

An input vector is chosen from the input set and compared to the reference vector for every node. The reference vector that is closest to the input vector according to some metric is the winning node. The Euclidean distance is a commonly used metric for determining the similarity of vectors and was the metric used for this thesis. The Euclidean distance, $\| x-y \|$, where $x = (a_1, a_2, \ldots, a_n)$ and $y = (b_1, b_2, \ldots, b_n)$ is given by the equation:

$$\text{Euclidean distance} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + ... + (a_n - b_n)^2}$$

Once the winning node has been found, the network must be updated. The winning node and those within some neighborhood are updated according to the equation:

$$\mathbf{w}_i(t + 1) = \mathbf{w}_i(t) + \alpha(t)(\mathbf{x} - \mathbf{w}_i(t)), \qquad i \in N_c$$

where $\mathbf{w}_i$ is the weight vector of the *i*th unit and $\mathbf{x}$ is the input vector [Free91]. N_c is the list of unit indices that make up the neighborhood, *c* is the winning node, and $\alpha(t)$ is a gain term with a value between zero and one that decreases in time to converge to zero. These adjustments are made in order to increase the likelihood that similar input vectors will choose this node again. By adjusting the neighborhood vectors as well, similar data vectors are pulled together, causing them to cluster [Lin91]. The process of choosing input vectors

and finding the most similar reference vector continues for a predetermined number of iterations.

The general algorithm [Pape98] is as follows:

1. A set of input vectors is created, with each input vector consisting of ones and zeros, for example, [1 0 1 . . . 0 0 1].
2. Every node in the output map is represented by a reference vector with the same size as the input vectors. The values of the reference vectors are randomly initialized.
3. An input vector is chosen from the input set and compared to the reference vector for every node.
4. The reference vector that produces the smallest Euclidean distance (or other metric) is considered the best match for that input vector.
5. The weights of the winning node and its neighbors are adjusted using the equation $\mathbf{w}_i(t + 1) = \mathbf{w}_i(t) + a(t)(\mathbf{x} - \mathbf{w}_i(t))$.
6. Return to step 2 and repeat the process for a predetermined number of iterations.

4.3 Related Work

The self-organizing map has been around since the early 1980s and since then it has been used extensively. Kaski et al. created a bibliography of 3,343 works that have been based on the SOM [Kaski98].

The applications of SOMs that are most relevant to this work are in the areas of document classification and web page categorization. Chen et al. [Chen96] used a SOM to automate the categorization of web pages. Vector representations of web pages were used as input to the SOM, creating a map on which related web pages are shown grouped together.

Many people have used SOMs to classify documents and Usenet newsgroups. Honkela et al. use SOMs to organize Usenet newsgroups [Honk96] in a system known as

WEBSOM. A visual map is then constructed with similar messages mapping closer to each other. More details of the WEBSOM system can be found in ([Kaski96], [Koho96b], [Koho98], [Lagus96], [Lagus98]). An extension to the SOM that improves the visual representation of the input data and makes cluster boundaries more obvious is described in ([Merkl97a], [Merkl97b]).

The use of SOMs for organizing web pages in this project is much the same as has been done in the above referenced works. This work also makes use of WordNet in a fashion similar to those works referenced in chapter 3. However, none of the previously referenced works combines WordNet and SOMs in order to improve the results of either system alone.

4.4 Limitations of This Approach

One of the limitations to this approach is that the SOM places web pages into a single category. As was discovered when the author preclassified the data, trying to categorize a web page by a single concept is not always easy. For example, when viewing the web pages that the traditional search engine produced when given the search terms {apache, helicopter, military}, one of the web pages was of a story about an Apache helicopter crash in Kosovo. This web page could reasonably be classified as ‘war in Kosovo’ or ‘helicopter crashes’. Ideally, a web page that represents multiple concepts could be found by searching for any of these concepts [Elo99].

This approach also suffers from the same problem as many other approaches—do the terms in the web page accurately describe the concept that the web page represents? This problem can be eased by web page authors embedding accurate metadata within the page, just as some technical articles have a list of keywords at the top of the cover page. Some web pages can already be found with metadata embedded in them.

CHAPTER V

EXPERIMENTAL RESULTS

In this chapter we discuss how we tested the effectiveness of the WordNet/SOM combination. We also give the results and evaluate them.

5.1 Testing Methodology

We decided to compare maps generated using the replacement terms from WordNet with maps that did not use replacement terms. The choice of a suitable testing methodology was a difficult one. While it was possible to see clusters in the maps generated by the SOM software, the author was unable to compare two maps visually in order to determine if one map was better than another. Also, each map generated by the SOM software consists of a two-dimensional grid of nodes with the nodes being at fixed distances from each other. These are the nodes that the input vectors map to as discussed in chapter 4. If, for example, three input vectors $\{V_1, V_2, V_3\}$ map to three consecutive nodes, then this is an indication from the SOM that these three input vectors are considered to be similar (see figure 6). It does not, however, mean that the distances between them are the same. The Euclidean distance from vector V_1 to V_2 is not necessarily the same as the Euclidean distance from V_2 to V_3. Because of this, a less subjective method of testing was needed.

Included with the SOM software is a program for generating a Sammon mapping of the data [Koho96a]. In the two-dimensional Sammon mapping produced, the distances between the vectors tend to approximate the Euclidean distances of the input vectors, giving the user an idea of the relative distances between the vectors. If the three input vectors from

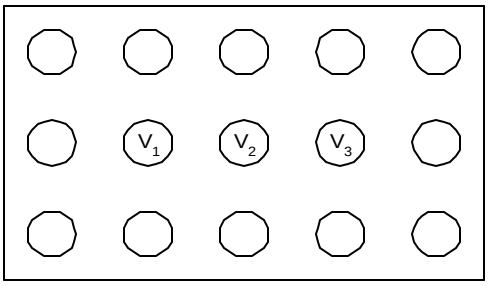


Figure 6. Sample SOM.

the previous example are used and two of them are closer to each other than they are to the third vector according to the metric used, then they will also appear physically closer to each other on the map. The SOMs and Sammon mappings used for the experiments can be found in appendix C.

In order to compare the clustering capabilities of the WordNet/SOM combination to the SOM alone, we decided to calculate the intracluster (within a cluster) and intercluster (between clusters) distances. A small average intracluster value indicates that the vectors within a category are grouped tightly together. A large average intercluster value indicates that the individual clusters are far apart. Ideally both of these will occur so that the clusters can be clearly identified.

Several metrics were used to evaluate the intracluster distances. The first looked at the individual categories. The average distance between the vectors (web pages) within a category was calculated using the X/Y coordinates generated by the Sammon map. A small average distance within a category indicates that the vectors are similar. When comparing the same category for two maps, the one with the smaller average distance is the better clustering. If all of the vectors within a category map to the same location, then the average distance will be zero. We can count the number of categories for which a particular map is

better. The map with the smaller average distance for a particular category is considered the winner for that category. The map that ‘wins’ the most is considered the better map according to this metric.

The second metric looks at the total of the average intracluster distances. When comparing entire maps, the average distances for all of the predetermined categories are summed and this value can be compared for each map. The map with the smallest total average distance is considered the better map. In our case, we will be comparing maps generated with replacement terms to maps generated without replacement terms.

The last intracluster metric is the average intracluster distance. The values from the second metric are divided by the number of categories that could be tested. This metric provides us with an idea of the size of each cluster.

A single metric was used for the intercluster distances. The center of each cluster is determined and the average distance between all of these is calculated. Because vectors with a greater size can potentially create a larger map, the values were normalized by dividing by the size of the vectors used for that map.

5.2 Testing Results

Four sets of experiments were performed. The parameters for the four were:

a) 37 author-generated categories; threshold for semantic distance of 4
b) 37 author-generated categories; threshold for semantic distance of 2
c) 10 categories represented by the original search concept; threshold for semantic distance of 4
d) 10 categories represented by the original search concept; threshold for semantic distance of 2

Several parameters had to be set when using the SOM software. A radius of ten was used for the neighborhood function and the map dimensions were 10x10. The SOM mapping is a two part process. In the first step, the gain $\alpha(t)$ was set at 0.05 and 1,000 iterations were used. In the second step, the gain $\alpha(t)$ was set at 0.02 and 10,000 iterations were used. The Sammon mapping used 500 iterations. All of these values were chosen based on their use in the examples provided with the software.

For the first two experiments, the categories that the web pages were assigned to were the thirty-seven categories that the author created after viewing the one hundred retrieved web pages. Although thirty-seven categories existed, only fifteen of the categories contained two or more web pages. This meant that only these fifteen categories could be tested using the testing methodology described in section 5.1. Different values for determining semantic similarity were considered. It was decided that a value of at most four would be used because greater values allowed the relationship between the words to be too general. Therefore, the threshold for determining semantic similarity was four for the first experiment and two for the second. With a threshold of four, words that were farther apart and therefore less related might be replaced by another term. The size of the vectors could potentially be smaller, though. When using a threshold of two, fewer words could potentially be considered related, but the size of the vectors would likely be greater than with a threshold of four.

For the experiment with thirty-seven categories and a threshold of four, the maps with and without replacement terms performed about the same according to the intracluster metrics. Of the fifteen categories that could be tested, both maps produced the smallest intracluster value for seven of the categories. One category had an average distance value of 0.0 for both methods. The total of the average distances for the map without replacement terms was 18.30 while the total of the average distances for the map with replacement terms

was 18.40, a difference of 0.55%. Both maps had approximately the same intracluster value; the map with replacement terms had a value of 1.23 and the map without replacement terms had a value of 1.22. When we look at the average intercluster values though, we see that the map generated with replacement terms was much better. The normalized average intercluster distance for the map without replacement terms was 2.10 versus 4.59 for the map with replacement terms. The size of the input vectors was decreased from 1,902 to 431 terms by using replacement terms.

The second experiment used thirty-seven categories and a threshold of two. Of the fifteen categories that could be tested, seven of the categories had a smaller average distance for the map without replacement terms. Six of the fifteen categories had a smaller average distance for the map with replacement terms. Two categories had an average distance value of 0.0 for both methods. The total of the average distances for the map without replacement terms was 18.30 while the total of the average distances for the map with replacement terms was 15.43, 15.70% less than without replacement terms. The map that used replacement terms had the smaller intracluster value; the map with replacement terms had a value of 1.03 and the map without replacement terms had a value of 1.22. As with the first experiment, the map that used replacement terms had the smaller normalized average intercluster values, although the values for each map were much closer. The normalized average intercluster distance for the map without replacement terms was 2.10 versus 2.71 for the map with replacement terms. The size of the input vectors was decreased from 1,902 to 930 terms by using replacement terms.

The next two experiments both categorized the web pages into the ten groups that were originally sought using the traditional search engine. The ten categories can be seen in table 1. As with the previous two experiments, the difference between these two experiments was the threshold value that was used. The results were similar in both cases.

For the experiment with ten categories and a threshold of four, the map without the replacement terms produced better values. Of the ten categories, eight had a smaller average distance for the map without replacement terms. The remaining two had a smaller average distance for the map with replacement terms. The total of the average distances for the map without replacement terms was 16.80 while the total of the average distances for the map with replacement terms was 20.35, 21.16% greater than without replacement terms. The map without replacement terms also had the better average intracluster value – 1.68 versus 2.04 for the map with replacement terms. The normalized average intercluster distance was still better for the map using replacement terms. It was 1.74 for the map without replacement terms and 5.43 for the map that used replacement terms. The size of the input vectors was decreased from 1,902 to 431 terms by using replacement terms.

The fourth experiment used ten categories and a threshold of two. As with the third experiment, the map without the replacement terms produced better values for three of the four metrics. Of the ten categories, six had a smaller average distance for the map without replacement terms. The remaining four had a smaller average distance for the map with replacement terms. The total of the average distances for the map without replacement terms was 16.80 while the total of the average distances for the map with replacement terms was 19.07, 13.52% greater than without replacement terms. The average intracluster distance was 1.68 for the map without replacement terms and 1.91 for the map with replacement terms. The normalized average intercluster value for the map without replacement terms was 1.74; it was 3.32 for the map with replacement terms. The map that used replacement terms only outperformed the map without replacement terms for the intercluster metric. The size of the input vectors was decreased from 1,902 to 930 terms by using replacement terms. The results from all four experiments can be seen in tables 3 and 4.

Table 3. Total of Average Distances

			Smaller Avg Distance per Category		Total of Avg Distances		
Experiment	Threshold	# of categories	W/O replacements	With replacements	W/O replacements	With replacements	% Diff.
1	4	37	7	7	18.30	18.40	-0.55%
2	2	37	7	6	18.30	15.43	15.70%
3	4	10	8	2	16.80	20.35	-21.16%
4	2	10	6	4	16.80	19.07	-13.52%

Table 4. Intercluster and Intracluster Distances

			Average Intercluster Distance		Average Intracluster Distance	
Experiment	Threshold	# of categories	W/O replacements	With replacements	W/O replacements	With replacements
1	4	37	2.10	4.59	1.22	1.23
2	2	37	2.10	2.71	1.22	1.03
3	4	10	1.74	5.43	1.68	2.04
4	2	10	1.74	3.32	1.68	1.91

5.3 Interpretation of Results

Of the four experiments, the experiments that produced the best results for a map using replacement terms were the two experiments that used the thirty-seven author-defined categories. Each of these two experiments produced values comparable to the maps without replacement terms for all of the intracluster metrics. When looking at the metric for evaluating the average intercluster distance, the use of replacement terms produced better maps. Because the thirty-seven author-defined categories better represented the web pages (in the author's opinion), these results were very favorable. When looking at intercluster-related metrics alone, it was observed that reducing the size of the input vectors was not detrimental in any of the four experiments.

The most likely reason for the poor performance of the WordNet/SOM combination as seen in the third and fourth experiments is that these two experiments used categories that

the author had already determined to not accurately describe the one hundred web pages. Because the ten original categories were not very accurate, getting poor results on these two experiments was not surprising.

The WordNet/SOM combination performed best when the threshold for determining semantic similarity was two instead of four. One reason for this is that as the threshold for determining semantic similarity increases, more words that are possibly unrelated in the given context may be replaced by another term. The purpose for having a threshold was to reduce the chances of this occurring, but the correct threshold to use is probably dependent on the two words being compared and is difficult to determine in advance. When the threshold was only two, this meant that replacement terms would only be used for coordinate terms (children of the same hypernym) or when one of the words was a hypernym of the other word.

Another reason is that using replacement terms may be decreasing the size of the input vectors too much in some cases. When the threshold was four, the input vectors decreased in size from 1902 terms to 431 terms, a reduction of 77%. It may be that an insufficient number of terms remained to accurately describe some web pages.

Finally, when visually inspecting the results, it was apparent that some correct clusters were forming in the maps. However, it was not possible for the author to visually determine which map was better. The metrics used may not have been satisfactory for measuring the results of this work.

CHAPTER VI

CONCLUSIONS

We provide our conclusions and thoughts for future work in this chapter.

6.1 Final Thoughts

This thesis was motivated by the ever increasing need to better organize the information found on the Internet. One of the better search engines available for searching the Internet is Yahoo (www.yahoo.com). The process for categorizing web pages within Yahoo is performed manually. This task will continue to become more difficult, if not impossible, as the Internet grows in size.

We believe that the task of organizing the information found on the Internet can be automated through the use of machine learning techniques. Some researchers have already applied self-organizing maps to the task of organizing web pages while others have applied WordNet to the task of classifying documents. We believe that a combination of self-organizing maps and WordNet can be used as well as, or more effectively than, either system alone. Even if the performance of the WordNet/SOM combination were equal to that of the SOM alone, the reduction in the size of the input vectors would be beneficial because of the reduction in computation time by the SOM.

Through visual inspection it could be seen that the WordNet/SOM combination did successfully organize some web pages. When using the author-defined categories, the intracluster performance of the WordNet/SOM combination was as good as or better than the SOM alone. The intercluster performance of the WordNet/SOM combination was better

than that of the SOM alone in all four experiments. We believe that the performance of the WordNet/SOM combination when working with well-defined categories indicates that there is merit in combining these two powerful tools, but work must continue on improving the integration of the two.

6.2 Future Work

Any future projects related to the work performed in this thesis will benefit from a better way to measure the results. Through visual inspection, it is possible to determine that the SOM is effectively organizing some of the web pages. However, it is difficult to determine how well the SOM is performing. A method is needed to quantify what a human notices when looking at the maps.

This work will greatly benefit from a way to determine the context of the words as they are used in a web page. This will allow the correct sense of the word to be used when generating hypernym trees. Also, incorporating the synsets found in WordNet will be helpful. This will be helpful with other natural language projects as well.

Future work should also be performed using a larger number of web pages to determine how well the process scales up. Using a much larger number of web pages would also allow the possibility of comparing the performance of a WordNet/SOM combination to the performance of a traditional search engine.

Finally, it would be helpful to let humans evaluate the results. The automatic organization of web pages by the WordNet/SOM combination could be compared to the manual organization of humans.

APPENDIX A

WEB PAGES

Descriptions of the web pages used for the experimentation are shown below.

Page name	Description of web page	classification of web page
ahm8	types of military helicopters	Apache helicopter
ahm2	description of Apache helicopter	Apache helicopter
ahm7	military-based artwork	art
ahm10	SRAM product to be used in helicopters	computer memory
ahm6	web site for military attack helicopter gunship enthusiasts	military enthusiasts
ahm1	story of Apache helicopter crash in Albania	news, Kosovo
ahm3	story of Apache helicopter crash in Albania	news, Kosovo
ahm5	news report from Kosovo	news, Kosovo
ahm9	news reports from Kosovo	news, Kosovo
ahm4	directions for military-based video game	video game
cbs10	article about age discrimination	age discrimination
cbs1	census bureau statistics by subject	census statistics
cbs2	census bureau statistics maps	census statistics
cbs9	economic and census statistics	census statistics, economics
cbs4	economic and census statistics	census statistics, economics
cbs3	economic and census statistics	census statistics, economics
cbs6	overview of government finance statistics	finance
cbs7	Tennessee genealogy links	genealogy
cbs5	views on racial discrimination and equality	social science
cbs8	social studies links	social studies
cml7	list of memory improvement tips	memory improvement
cml6	reviews of the book "Brain Longevity"	memory improvement, Brain Longevity
cml4	seminar for the Brain Longevity Program	memory improvement, Brain Longevity
cml3	pregnenolone - the antiaging, memory enhancing hormone	memory improvement, supplements
cml10	article about "Brain Longevity"	memory improvement, supplements
cml5	article about "Brain Longevity"	memory improvement, supplements
cml2	study of how phosphatidylserine improves memory	memory improvement, supplements
cml8	description of brain nutrient supplement Neurotone	memory improvement, supplements

cml1	mental problems caused by brain chemistry imbalances	mental problems, brain chemistry
cml9	class notes for neuroanatomy	neuroanatomy
ffr5	rules for fantasy football	fantasy football rules
ffr4	rules for fantasy football	fantasy football rules
ffr9	rules for fantasy football	fantasy football rules
ffr1	rules for fantasy football	fantasy football rules
ffr8	rules for fantasy football	fantasy football rules
ffr3	rules for fantasy football	fantasy football rules
ffr10	rules for fantasy football	fantasy football rules
ffr6	rules for fantasy football	fantasy football rules
ffr2	rules for fantasy football	fantasy football rules
ffr7	rules for fantasy soccer (British site)	fantasy soccer rules
flt4	conference about the solution of Fermat's last theorem	Fermat's last theorem
flt5	references for Fermat's Last Theorem	Fermat's last theorem
flt10	history of Fermat's last theorem	Fermat's last theorem
flt1	article about speech by Andrew Wiles about Fermat's last theorem	Fermat's last theorem
flt2	solution to Fermat's last theorem	Fermat's last theorem
flt6	solution to Fermat's last theorem	Fermat's last theorem
flt7	story about the solution Fermat's last theorem	Fermat's last theorem
flt9	history of Fermat's last theorem	Fermat's last theorem
flt8	college course: Fermat's Last Theorem in Context	Fermat's last theorem, college course
flt3	poetry about the solving of Fermat's last theorem	poetry, Fermat's last theorem
gft6	agricultural statistics CD-ROM	agricultural statistics
gft7	fertilizer for fruit trees	fertilizer
gft4	how to grow fruit trees	growing fruit trees
gft2	fruit tree links	growing fruit trees
gft3	how to grow fruit trees	growing fruit trees
gft8	list of books about growing fruit trees	growing fruit trees
gft10	how to grow fruit trees	growing fruit trees
gft5	fruit tree links	growing fruit trees
gft1	growing strawberries	growing strawberries
gft9	Thip Dhani Project in Vietnam	planned housing
nnc3	paper about Japanese character recognition	Japanese character recognition
nnc5	description of slide from machine learning presentation	machine learning

nnc4	list of machine learning research papers	machine learning
nnc10	software product that uses multipass instance learning	machine learning, multipass instance learning
nnc6	abstract for paper about Multiresolution Nearest Neighbor Classifier	machine learning, nearest neighbor classifier
nnc1	paper using machine learning methods	machine learning, nearest neighbor classifier
nnc2	proposal for work to study nearest neighbor classifiers	nearest neighbor classifiers
nnc8	abstract for paper about nearest neighbor classifier, neural network	nearest neighbor classifier, neural network
nnc7	abstract for paper about protein chemistry	protein chemistry
nnc9	abstract for paper about protein chemistry	protein chemistry
pca9	paper about time series of perceptual data	cognition
pca10	paper about collaborative information analysis	collaborative information analysis
pca3	paper on Interactive Interpretation of Hierarchical Clustering	hierarchical clustering
pca4	paper about Markush structures	molecular graphics
pca1	project summary: indexing multimedia information	multimedia
pca2	study of photon energy resolution	photon energy, clustering
pca6	bibliography for Scaling and Dimensional Analysis	scaling and dimensional analysis
pca5	paper: A Boolean Classification Analysis of Successful Military Coups	sociology, military coups
pca7	paper about Speech Analysis And Recognition	speech analysis and recognition
pca8	paper about value analysis	value analysis, finance
ssp6	description of "Sky3D" software	software
ssp9	evaluation form for "Windows to the Universe" software	software
ssp8	story of the discovery of planets around a distant star	solar system
ssp4	poster of the planets	solar system
ssp1	links to solar system web sites	solar system
ssp5	statistics of the planets	solar system
ssp3	description of the planets	solar system
ssp2	description of the planets	solar system
ssp10	links to solar system web sites	solar system
ssp7	description of the planets	solar system
wamb5	links to Mozart web sites	Mozart biography
wamb9	brief biography of Mozart	Mozart biography

wamb3	biography of Mozart	Mozart biography
wamb7	list of books about Mozart	Mozart biography
wamb1	biography of Mozart	Mozart biography
wamb6	biography of Mozart	Mozart biography
wamb10	links to biographies of Mozart	Mozart biography
wamb4	Austrian tourism site about Mozart	Mozart biography
wamb8	list of Mozart's music	music, Mozart
wamb2	list of Mozart's music	music, Mozart

APPENDIX B

WEB PAGE CLASSIFICATIONS

Below are the thirty-seven classifications that the author identified for the one hundred retrieved web pages.

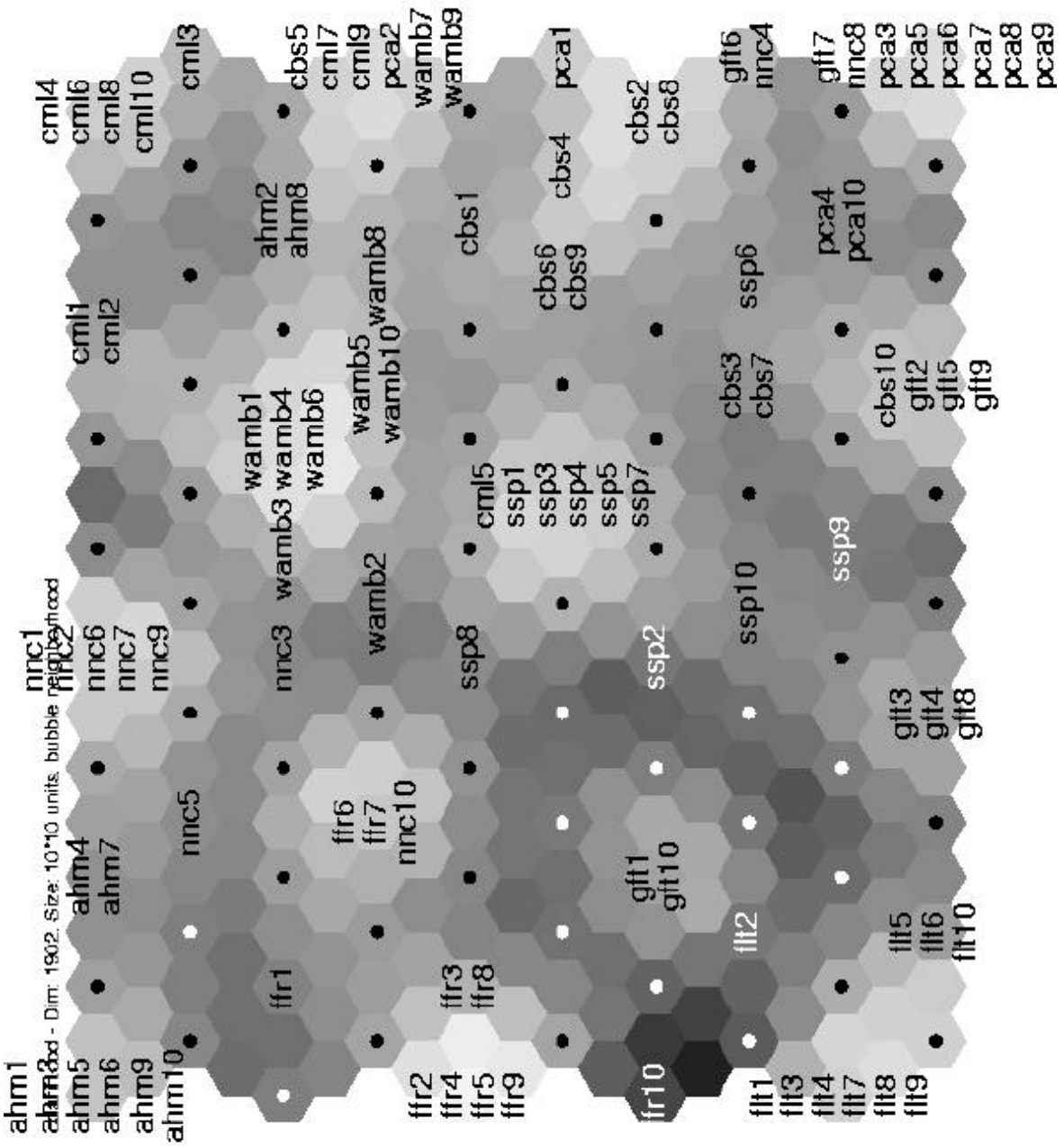
classification	web sites
Apache helicopter	ahm2, ahm8
art	ahm7
computer memory	ahm10
Kosovo, news	ahm1, ahm3, ahm5, ahm9
military enthusiasts	ahm6
video game	ahm4
age discrimination	cbs10
census statistics	cbs1, cbs2, cbs3, cbs4, cbs9
finance	cbs6
genealogy	cbs7
social science	cbs5, cbs8
memory improvement	cml2, cml3, cml4, cml5, cml6, cml7, cml8, cml10
mental problems, brain chemistry	cml1
neuroanatomy	cml9
fantasy football rules	ffr1, ffr2, ffr3, ffr4, ffr5, ffr6, ffr8, ffr9, ffr10
fantasy soccer rules	ffr7
Fermat's last theorem	flt1, flt2, flt3, flt4, flt5, flt6, flt7, flt8, flt9, flt10
agricultural statistics	gft6
fertilizer	gft7
growing fruit trees	gft1, gft2, gft3, gft4, gft5, gft8, gft10
planned housing	gft9
Japanese character recognition	nnc3
machine learning	nnc1, nnc2, nnc4, nnc5, nnc6, nnc8, nnc10
protein chemistry	nnc7, nnc9
classification, sociology, military coups	pca5
clustering	pca3, pca9
collaborative information analysis	pca10
Scaling and Dimensional Analysis	pca6
molecular graphics	pca4
multimedia	pca1
photon energy, clustering	pca2
speech analysis and recognition	pca7
value analysis, finance	pca8
software	ssp6, ssp9
solar system	ssp1, ssp2, ssp3, ssp4, ssp5, ssp7, ssp8, ssp10

Mozart biography	wamb1, wamb3, wamb4, wamb5, wamb6, wamb7, wamb9, wamb10
music, Mozart	wamb2, wamb8

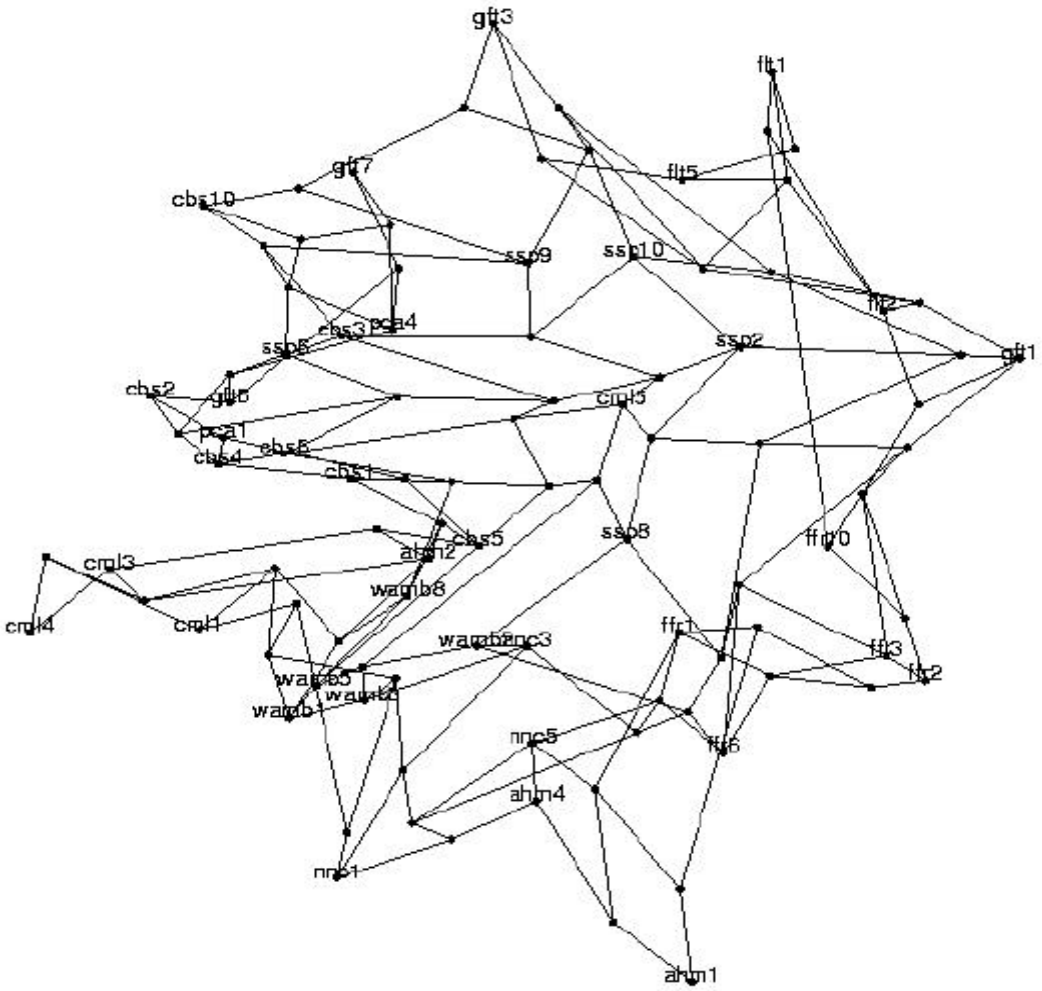
APPENDIX C

MAPS

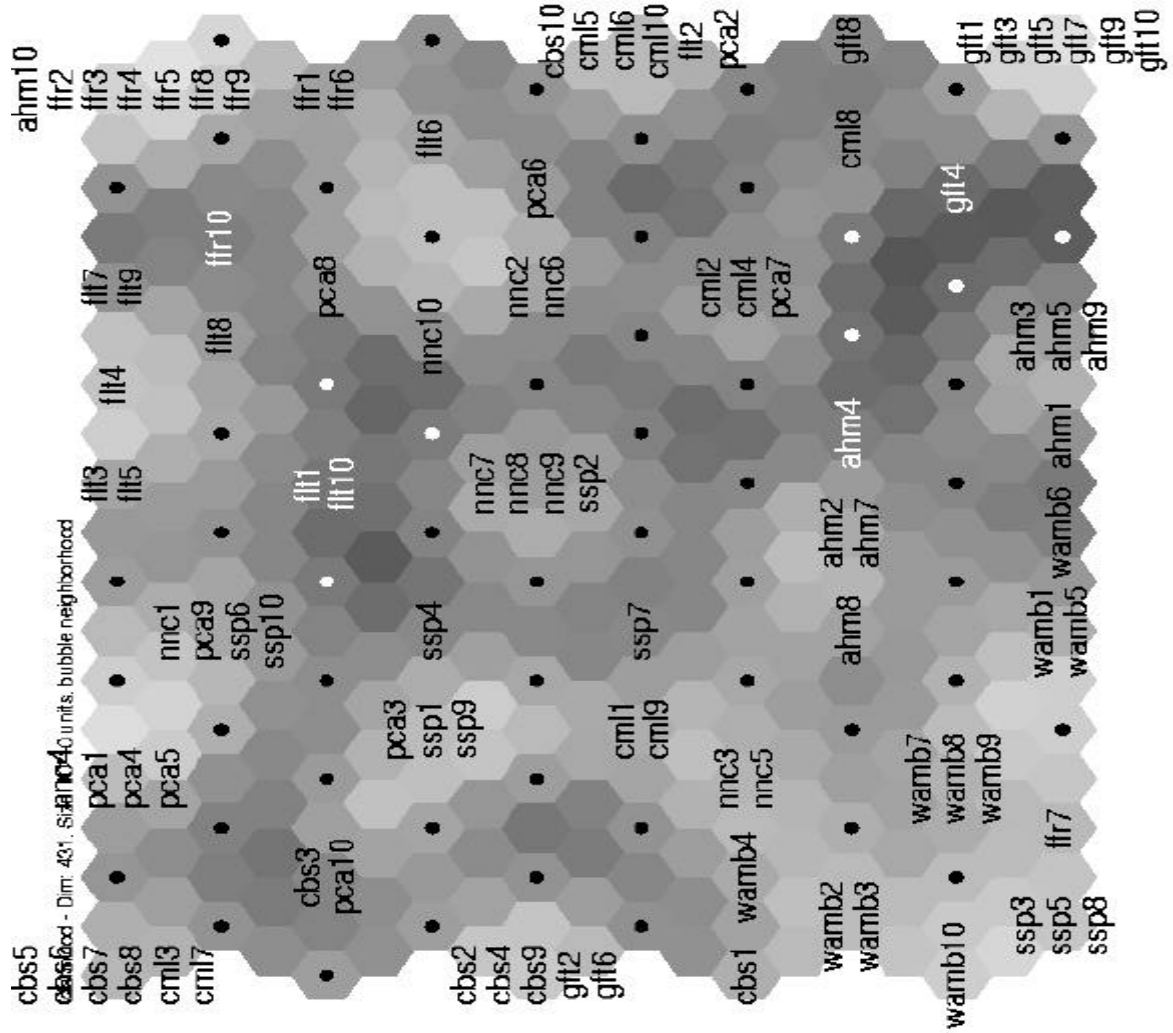
Self-organizing map generated without the use of replacement terms.



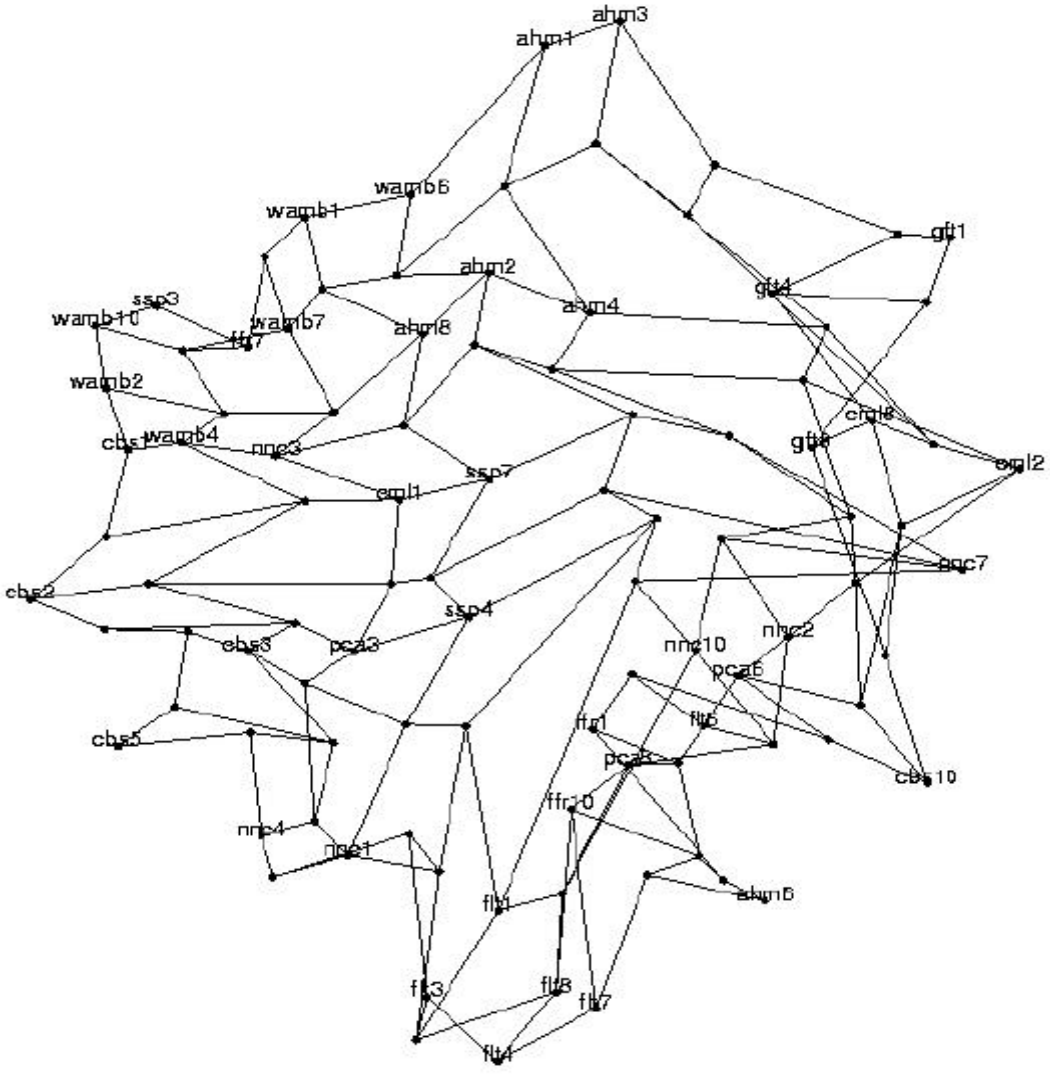
Sammon mapping without the use of replacement terms.



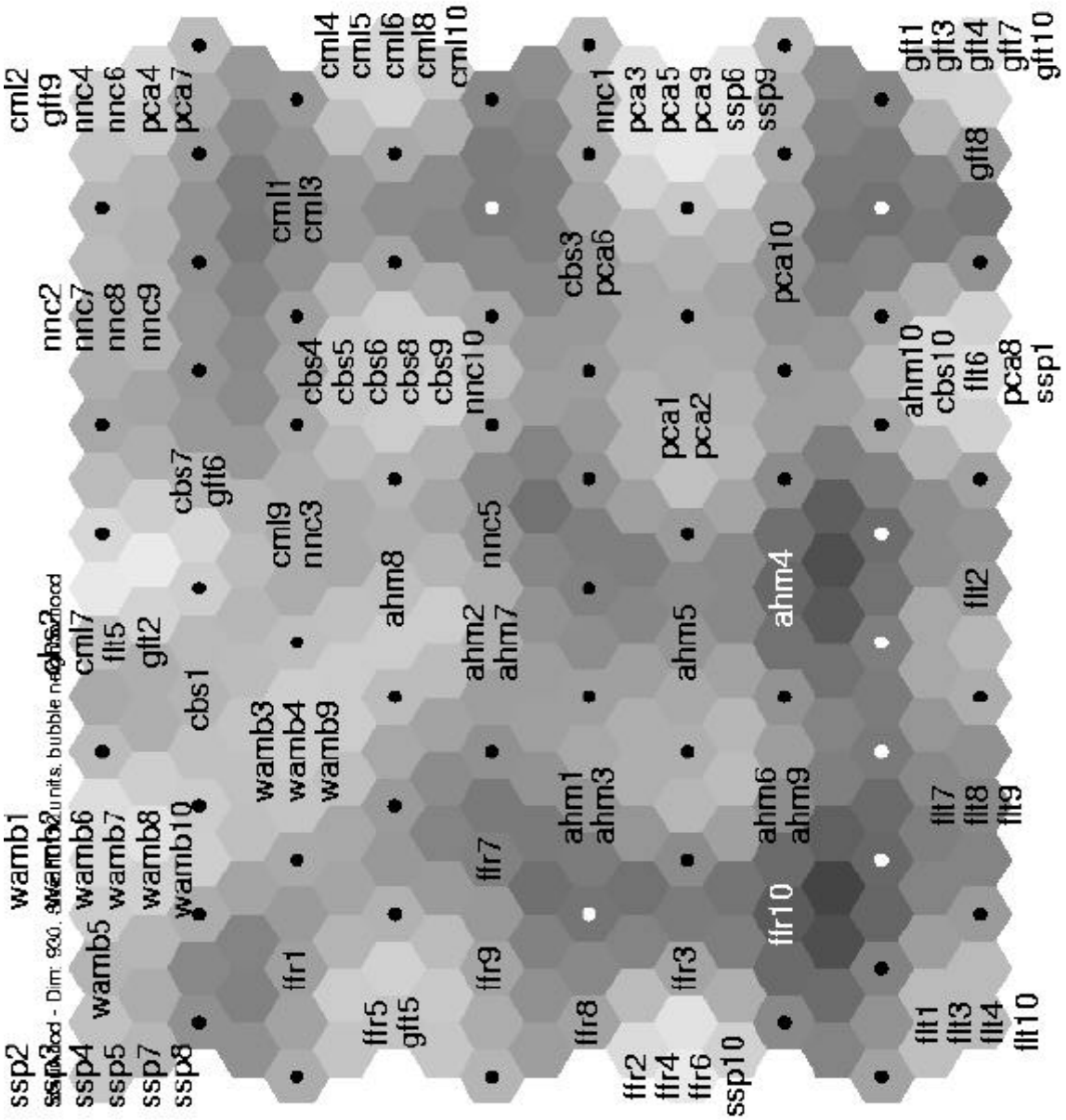
Self-organizing map generated with the use of replacement terms and a threshold of four.



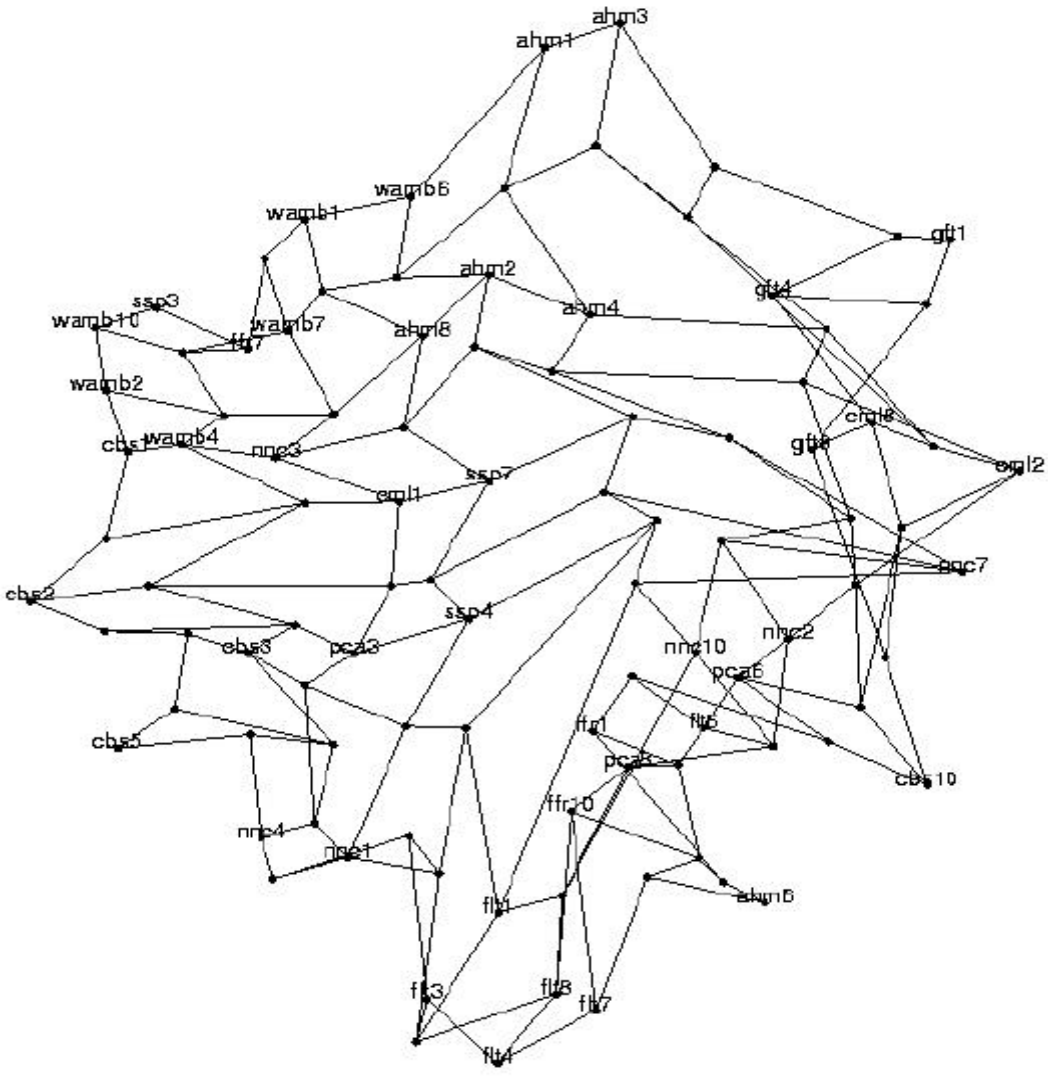
Sammon mapping generated with the use of replacement terms and a threshold of four.



Self-organizing map generated with the use of replacement terms and a threshold of two.



Sammon mapping generated with the use of replacement terms and a threshold of two.



REFERENCES

[Alta99] Altavista, "Frequently Asked Questions,"
URL: http://www.altavista.com/av/content/ques_howto.htm

[Chen96] Chen, Hsinshun, Chris Schuffels, and Rich Orwig, "Internet categorization and search: A self-organizing approach," in Journal of Visual Communication and Image Representation, 7(1):88-102, 1996.

[Dunn99] Dunn, Ashley, "Web faces searching questions," in the Dallas Morning News, pg. 1A, July 8, 1999.

[Elo99] Elo, Sara, Louis Weitzman, Christopher Fry, and Jeff Milton, "Virtual URLs for browsing and searching large information spaces," in WebNet Journal, vol. 1, no. 1, 1999.

[Fell90] Fellbaum, Christiane, "English verbs as a semantic net," in International Journal of Lexicography 3 (4):278 - 301, 1990.

[Free91] Freeman, James A., and David M. Skapura, "Neural Networks: Algorithms, applications, and programming techniques," Addison-Wesley, Reading, Massachusetts, 1991.

[Hadd98] Haddock, Gail, "Early validation of task analysis results through incremental discrepancy determination," Ph.D. thesis, The University of Texas at Arlington, 1998.

[Honk96] Honkela, Timo, Samuel Kaski, Krista Lagus, and Teuvo Kohonen, "Newsgroup exploration with WEBSOM method and browsing interface," in Technical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1996.

[Kaski96] Kaski, Samuel, Timo Honkela, Krista Lagus, and Teuvo Kohonen, "Creating an order in digital libraries with self-organizing maps," in Proceedings of WCNN'96, World congress on Neural Networks, September 15-18, 1996, San Diego, California, pp. 814-817, Lawrence Erlbaum and INNS Press, Mahwah, New Jersey.

[Kaski98] Kaski, Samuel, Jari Kangus, and Teuvo Kohonen, "3343 works that have been based on the self-organizing map (SOM) method developed by Kohonen," Neural Networks Research Centre at Helsinki University of Technology, URL: http://www.cis.hut.fi/nnrc/refs/, 1998.

[Koho95] Kohonen, Teuvo, "Self-Organizing Maps," Springer, Berlin, Heidelberg, 1995.

[Koho96a] Kohonen, Teuvo, Jussi Hynninen, Jari Kangas, and Jorma Laaksonen, "SOM_PAK: The self-organizing map program package," in Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 1996.

[Koho96b] Kohonen, Teuvo, Samuel Kaski, Krista Lagus, and Timo Honkela, "Very large two-level SOM for the browsing of newsgroups," in von der Malsburg, C., W. von Seelen, J. C. Vorbriiggen, and B. Sendhoff, editors, Proceedings of ICANN96, International Conference on Artificial Neural Networks, Bochum, Germany, July 16-19, 1996, Lecture Notes in Computer Science, vol. 1112, pp. 269-274, Springer, Berlin.

[Koho98] "Self-organization of very large document collections: State of the art," In Niklasson, L., Bodén, M., and Ziemke, T., editors, Proceedings of ICANN98, the 8th International Conference on Artificial Neural Networks, volume 1, pages 65-74, Springer, London, 1998.

[Li95] Li, Xiaobin, Stan Szpakowicz and Stan Matwin, "A WordNet-based algorithm for word sense disambiguation," in Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, 1995, pp. 1368 - 1374.

[Lin91] Lin, Xia, Dagobert Soergel, and Gary Marchionini, "A self-organizing map for information retrieval," in Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, pp. 262-269, Chicago, Illinois, October 13-16, 1991.

[Lagus96] Lagus, Krista, Samuel Kaski, Timo Honkela, and Teuvo Kohonen, "Browsing digital libraries with the aid of self-organizing maps," in Proceedings of the Fifth International World Wide Web Conference WWW5, May 6-10, Paris, France, volume Poster Proceedings, pp. 71-79, EPGL, 1996.

[Lagus98] Lagus, Krista, "Generalizability of the WEBSOM method to document collections of various types," In Proceedings of 6th European Congress on Intelligent Techniques & Soft Computing (EUFIT'98), volume 1, pp. 210-214, Verlag Mainz, Aachen, Germany, 1998.

[Merkl97a] Merkl, Dieter, and Andreas Rauber, "On the similarity of eagles, hawks, and cows: visualization of semantic similarity in self-organizing maps," in Proceedings of International Workshop on Fuzzy-Neuro-Systems, Soest, Germany, 1997.

[Merkl97b] Merkl, Dieter, "Exploration of document collections with self-organizing maps: A novel approach to similarity representation," in Proceedings of the 1st European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'97), Trondheim, Norway, June 25-27, Springer-Verlag (Lecture Notes in Artificial Intelligence), Berlin, 1997.

[Miller93] Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine J. Miller, "Introduction to WordNet: an on-line lexical database," in International Journal of Lexicography 3 (4), 1990, pp. 235 - 244. Revised 1993.

[Miller98] Miller, George A., "Nouns in WordNet," in Christiane Fellbaum (Ed.), "WordNet: An electronic lexical database," MIT Press, Cambridge, Massachusetts, 1998.

[Pape98] Pape, Daniel X., "A very brief summary of the self-organizing map," URL: http://www.canis.uiuc.edu/~dpape/dlisom/summary.html, 1998.

[Rodrig97] Rodríguez, Manuel de Buenaga, José María Gómez-Hidalgo and Belen Diaz Agudo, "Using WordNet to complement training information in text categorization," in Proceedings of the International Conference on Recent Advances in Natural Language Processing, Tzigov Chark, 1997.

[Rosen98] Rosenzweig, Joseph, "WordNet Bibliography," URL: http://www.cis.upenn.edu/~josephr/wn-biblio.html, 1998.

[Scott98] Scott, Sam and Stan Matwin, "Text classification using WordNet hypernyms," Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal, 1998.

[Taulli99] Taulli, Tom, "Google.com: Next Brainchild To Go Big?" in Techweb, URL: http://www.techweb.com/wire/finance/story/netgain/INV19990127S0005, January 27, 1999.

[Word98] WordNet version 1.6 documentation, URL: http://www.cogsci.princeton.edu/~wn/doc/, 1998.

BIOGRAPHICAL STATEMENT

Darin Brezeale received his Bachelor of Science in Electrical Engineering from The University of Texas at Arlington in 1992 and his Master of Science in Computer Science Engineering from The University of Texas at Arlington in 1999.