An Ecology of the Net
Message Morphology and Evolution in NetNews

by
Michael Lloyd Best
Bachelor of Science
University of California, Los Angeles
1989

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning, in partial fulfillment of the
requirements for the degree of
Master of Science

at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
June, 1996

© Massachusetts Institute of Technology, 1996.
All Rights Reserved.

Author .............................................

Program in Media Arts and Sciences
May 10, 1996

Certified by ..................................

Kenneth B. Haase, Jr.
Assistant Professor of Media Arts and Sciences
Program in Media Arts and Sciences
Thesis Supervisor

Accepted by ........

Stephen A. Benton
Chair
Departmental Committee on Graduate Students
Program in Media Arts and Sciences
An Ecology of the Net
Message Morphology and Evolution in NetNews

by

Michael Lloyd Best

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning, on May 10, 1996,
in partial fulfillment of the requirements for the degree of
Master of Science

Abstract

We introduce the concept of an ecology based on a corpus of text and apply it to the analysis of posts to USENET News. In our corporal ecology posts are organisms, the newsgroups of NetNews define an environment, and human posters situated in their wider context make up a scarce resource and describe a dynamic fitness landscape. We propose eigen-text, a technique which through statistical analysis distills from a corpus those self-reproducing units of text. Eigen-text arrives at suitable replicators because it discovers word co-occurrences which segregate and recombine with appreciable frequency. We argue that natural selection is necessarily operating on eigen-text because all of the requirements for its occurrence obtain: self-reproduction, mutagenicity, and metabolism. We describe a set of experiments performed on a static corpus of over 10,000 posts. In our experiments two fundamental elements of population ecology are investigated: fitness and diversity. Our study of fitness arrives at the unhappy discovery that a flame-war and an overly prolific poster is the king of the jungle. In investigating diversity we make the surprising conclusion that as the number of posts on a particular subject increases, the diversity between those posts decreases.

We have built three visualization tools. Two of them allow the user to examine the spatial patterns and activity levels of clusters of posts across the landscape defined by NetNews. The third systems builds and displays a taxonomy of posts to NetNews describing a tree-like set of relationships between post clusters.

We go on to suggest a number of practical systems based on our work which we would like to build. These include content based news readers and tools to filter and personalize information on the Net.

Thesis Supervisor: Kenneth B. Haase, Jr.
Title: Assistant Professor of Media Arts and Sciences

This research was sponsored by the News in the Future consortium.
An Ecology of the Net
Message Morphology and Evolution in NetNews

The following people served as readers for this thesis:

Reader .............................................
W. Daniel Hillis
Adjunct Professor
Program in Media Arts and Sciences

Reader .............................................
Mitchel Resnick
Assistant Professor of Media Arts and Sciences
Fukutake Career Development Professor of Research in Education
Program in Media Arts and Sciences
### Table of Contents

Acknowledgements ........................................................................................................ 9

1 Introduction .................................................................................................................. 11
   1.1 Evolution of an Idea ......................................................................................... 11
   1.2 Overview of the System .................................................................................... 12
   1.3 On Sociobiology ............................................................................................... 13
   1.4 Related Work .................................................................................................... 14
   1.5 The Contributions Made Explicit ....................................................................... 16
   1.6 Why Make the Connection? ............................................................................... 17
   1.7 Organization of Chapters .................................................................................. 18

2 An Overview of NetNews ............................................................................................ 19

3 The Text Mechanisms ............................................................................................... 23
   3.1 System Overview .............................................................................................. 23
   3.2 Stop Words and Stemming .............................................................................. 24
   3.3 Vector Space Representation ............................................................................ 26
   3.4 Eigen-text ........................................................................................................... 29
   3.5 Cosine Similarity Measure ............................................................................... 34
   3.6 Nearest Neighbor Clustering .......................................................................... 38
   3.7 Direction Cosines .............................................................................................. 39

4 The Biology of Eigen-Text ......................................................................................... 41
   4.1 Biological Connections ..................................................................................... 41
   4.2 Natural Selection and Replication .................................................................... 44
   4.3 Fitness Landscapes ............................................................................................ 48

5 Fitness .......................................................................................................................... 53
   5.1 Dataset ................................................................................................................ 53
   5.2 Fitness .................................................................................................................. 55
   5.3 Champion Quasi-species .................................................................................. 56
   5.4 Adaptive Meme Complexes ............................................................................. 58
   5.5 Summary .............................................................................................................. 61

6 Diversity ...................................................................................................................... 63
   6.1 Levels of Diversity ............................................................................................ 63
   6.2 Ecosystem Diversity .......................................................................................... 65
   6.3 Population Diversity .......................................................................................... 68

7 Visualization Methods ............................................................................................... 71
   7.1 Newsgroup Tree ............................................................................................... 71
   7.2 Phylogenetic Display ......................................................................................... 74

8 Conclusions and Future Work ................................................................................... 79
   8.1 Conclusions ......................................................................................................... 79
   8.2 Future Work ......................................................................................................... 80

References ....................................................................................................................... 85
I gratefully acknowledge the guidance and support of my advisor, Ken Haase. Ken has given me the freedom to explore; who would have thought that my initial interests in parallel memory based reasoning would lead to a thesis on ecologies of discourse! Thanks also go to my readers Danny Hillis and Mitchel Resnick who both have provided very helpful critiques of this work.

I have greatly benefited from and enjoyed the support, intellectual partnership, and friendship of the other members of the Machine Understanding Group. Thanks to Janet Cahn, Sara Elo, Warren Sack, Pushpinder Singh, and Zhong-hui Xu. Sincere thanks to Janet Cahn and Warren Sack for their close readings of a draft of this thesis.

Walter Bender and the News in the Future Consortium have provided nutriment and support for this work which is gratefully acknowledged. Walter’s enthusiasm for all of our work is inspiring.

Thanks also to Richard Pocklington and David Christini for their early inspiration and for their thoughtful criticism and encouragement.

Finally, thanks to my parents who set me on this path and to my sister who has been a support as we struggled in parallel through the turmoils of graduate school.
Chapter 1

Introduction

*We choose the theory which best holds its own in competition with other theories; the one which, by natural selection, proves itself the fittest to survive.*

Sir Karl Popper, 1959

1.1 Evolution of an Idea

*Do ideas evolve?* The quote above, by the noted philosopher of science Karl Popper, gives one early affirmative response to this question. Biologists have also, quite naturally, taken up this question. At one extreme has been the work of the sociobiologists, most notably E.O. Wilson, who have advanced theories which purport to unite all of social and cultural behavior at the most macroscopic level with the theories of evolutionary biology and the neo-Darwinian Synthesis (Wilson 1975). Richard Dawkins, at the other extreme, has worked more at the “microscopic” level; in other words, he starts with much smaller phenomena. Dawkins advocates the *meme*, an analogue to gene, as the unit of idea replication and the target of evolutionary selection. “Examples of memes are tunes, ideas, catch-phrases, clothes fashions, ways of making pots or of building arches,” (Dawkins 1989, pp 192).

*How do we build a formal theory of the evolution of ideas?* This is the question we will attempt to answer. To really build such a theory we need first to distill exactly the evolving units which undergo self-reproduction. To accept Dawkins’ formulation, we are in search of the meme. Furthermore, we need to build a model for the environment in which these units undergo reproduction with selection. A set of evolving interacting agents within some environment defines an ecology -- so, in the end it is an evolutionary-theoretical framework for this “idea ecology” which we seek.
We are building a set of models for self-reproduction and evolution of ideas within an environment. However, we are not studying the evolution of ideas within the natural world; at least not directly. We instead are studying the structure and evolution of ideas within a textual discourse. That is, we propose an ecology based on the interactions and evolution of connected text within a growing single corpus or set of corpora -- a corporal ecology. Unlike the traditional locale for biological investigations, the environment of the corporal ecology is not the natural world around us but the digital environment defined by the Net. The organisms of our model are not carbon-based “survival machines” (Dawkins 1989), but text encoded as bits, existing in situ a digital environment, and travelling at light-speed over (more and more often) optical fibers. While the Net affords a variety of corporal ecologies (email, the Web, etc.), we concentrate our analysis on a particularly well-developed one: messages posted to USENET News or NetNews.

1.2 Overview of the System
We have built a software system to distill elements of replication, ideas which reproduce themselves, within a corpus of messages posted to NetNews. Given these replicators, we have developed a set of theories which describe their evolution and interaction and have tested some of these theories through the analysis of an example corpus. In our corporal ecology, posts to NetNews are organisms, the newsgroups define an environment, and human posters situated in their wider context make up a scarce resource and describe a dynamic fitness landscape. (For the uninitiated, the next chapter will detail the NetNews system.)

To discover replicators we begin with a collection of traditional text analysis methods which convert each post into a vector representation of the words which compose it. These vectors, one for each post, are all collected into a matrix representing the entire corpus. A principal component analysis finds statistically significant word co-occurrences across the
corpus. It is the representations of these word co-occurrences which we treat as replicators within the corpus. We call this method of distilling statistically significant corporal replicators \textit{eigen-text}. Next, we re-represent each document in terms of these replicators and proceed to cluster the semantically related documents through a traditional classification algorithm. We claim that these clusters represent a sort of quasi-species within our ecology, that is, they group posts which share underlying structures.

Thus, we have created a system for distilling and representing replicators, organisms, and quasi-species all within an environment defined by NetNews. It is this system which will form the basis for our ecological modeling.

\textbf{1.3 On Sociobiology}

Our investigations into the evolution of ideas within a corpus may appear related to the sociobiology research program. The efforts of the sociobiologist has come under considerable attack (Caplan 1978; Gould 1991) and we feel it is worthwhile to make clear the critical differences between our work and theirs.

The term “Sociobiology” was coined by E.O Wilson (Wilson 1975). Wilson is an ant-man by training. Anyone who has observed a colony of ants cannot help but be struck by the emergent complexity of their society. Given a close study of their neurological structures, it is evident that the social insects are relatively simple organisms with small capacity for learning. Thus, one can conclude that much if not all of their social behaviors are hereditable, that is, programmed in their genes. Given this reading, an unfettered application of Darwinian theories to their social behaviors makes good sense (and succeeds to a large degree).

The strongest programs of sociobiology attempt to extend the rubric of genetic heritability to describe all of culture. But this has gotten them into considerable trouble, in par-
ticular when the theories have been applied to humans and have taken on social behavior such as "xenophobia, religion, ethics, social dominance, hierarchy formation, slavemaking, etc." (Sociobiology Study Group 1978 pp 285).

Our work does not fit into the program of sociobiology because we fundamentally do not claim that memes can be inherited in our genes. We claim that genes live or die in competition with other genes; similarly memes live and die due to their competition with other memes. We do not tangle our targets of selection, claiming, as the sociobiologists do, that humans or human societies live and die due to the success of their memes which originate in their genes. In fact, our work holds a radically human-orthogonal view: in other words, we take a meme-eyed view. Although humans do not appear explicitly anywhere in our model we will later show their implicit role.

1.4 Related Work
Our work, admittedly, spans a number of fields; to situate this research into its context requires a discussion of a number of disciplines.

Philosophers have provided some of the earliest and best links between the life and death of ideas and natural selection. We see this in the work most notably of Karl Popper, but there are others including Campbell, Ackermann, Dennett, and surely more (Popper 1972; Campbell 1974; Ackermann 1970; Dennett 1995).

The word "meme" was coined by Richard Dawkins and he has provided some in-principle arguments about their existence and dynamics (Dawkins 1982, 1989). Some additional work has been done to develop these theories (Koch 1986). Other researchers have tried to develop theories of memetic dynamics and have built computer simulations of their theories (Gabora 1994; Bura 1993; Belew 1990). We are surprised, though, by the general lack of this sort of work and the mixed results so far. To our knowledge, ours is the
first work to actually study memetic dynamics without simulation or synthesis. That is, we are analyzing the memetic dynamics of a pre-existing textual ecology rather than beginning tabula rasa and building whole-cloth a new model and test data.

The value of the meme-eye view is argued by Dennett (Dennett 1995). This radically human-orthogonal view has been taken by others and is analogous to the gene-eyed view supported elsewhere (Dawkins 1989). Hillis develops a meme-eyed argument for the evolution of human language by envisioning songs as parasites of our ape-ancestors (Hillis 1988). We note that this idea is also related to the anti-humanist arguments of the structuralists and post-structuralists (Eagleton 1983).

We believe that this work is situated within the field of Artificial Life (Langton 1989; Langton, et. al. 1992). The traditional Alife research program has been described by Charles Taylor as something “concerned with human-made systems that exhibit behaviors which are more or less characteristic of natural living systems” (Taylor 1992). We believe a weakness of this research program has been the almost complete reliance on simulation and synthesis. We consider our work to be an example of a new form of Alife research -- the analysis of pre-existing Alife communities in their natural habitat.

We are developing theories of the memetic dynamics within a corporal ecology. Some use of biological theories for problems of text analysis can be found in the literature. One type of such work has been the application of the genetic algorithm (Holland 1975) to problems in text retrieval and classification (Yang and Korfhage 1993; Moukas 1996). In related research, Robert O'Hara and colleagues have made very novel use of phylogenetic techniques (the computation of “trees of life” for related organisms) to the analysis of manuscript traditions. These manuscript traditions, such as old Norse tales, have undergone significant evolution over centuries of copying and cladistic analysis attempts to
piece together this historical lineage (Robinson and O’Hara forthcoming; O’Hara and Robinson 1993).

The text analysis methods we employ borrow heavily from the rich literature on the subject. In particular we use techniques developed over many years primarily by Gerard Salton and colleagues at Cornell University (Salton and McGill 1983; Salton and Buckley 1988). The application of a principle factor analysis to text vectors was first proposed by Dumais and co-authors with a method called latent semantic indexing (Furnas, et.al 1988; Deerwester, et.al 1990; Dumais 1992; Dumais 1993). The particular principle factor analysis we have employed, the singular value decomposition, was studied and implemented for text applications by Berry and colleagues (Berry 1992; Berry, et.al. 1993: Berry and Fierro 1995). We have borrowed the techniques of Dumais and Berry, with some departures; however our application area is novel.

1.5 The Contributions Made Explicit

We believe these to be the primary contributions of this work:

1. We introduce the notion of a *corporate ecology*. This ecology is made up of interacting evolving organisms (posts to NetNews) within an environment (newsgroups). We use a number of mechanisms, many culled from the text-retrieval literature, in novel ways to study the morphology and evolution of our text-based organisms.

2. We are building theories of population memetics. This has included the study of a method to derive replicating units -- those elements which undergo selection. We call this method *eigen-text*.

3. Our work forms a new type of Alife research -- one in which we analyze a pre-existing community of artificial organisms rather than synthesize a community based on our own parameters.
4. We have developed a technique to measure the average population fitness of clusters of posts. We have applied this technique, along with methods of measuring diversity and computing post taxonomies, to a large corpus of posts. We have found that for this corpus the most fit cluster of posts is a flame-war and that as the number of posts on a topic increases the diversity decreases.

5. We have developed a toolbox of techniques which can be applied to information analysis and retrieval on the Net. These text processing and classification techniques will be critical if we wish to make information on the Net more accessible, more personalized, and more interesting.

1.6 Why Make the Connection?
Finally, why make this connection at all? No matter how strongly we argue for a connection between the models of an evolutionary-theoretic ecology and the progress of ideas within text, sooner or later we must consider the value of such connections -- independent of their veracity. Clearly, these connections are not perfect. There are many ways in which the constraints on a gene are different or relaxed for the meme. Much of the work in understanding population genetics is left to considering certain physical constraints on the gene; how it gets about. Whereas, the meme in cyberspace is under relatively relaxed physical constraints.

We believe there to be considerable value in this connection. In our discussion of future work we mention some practical applications of these models. However, we believe the real value will rest on how predictive and descriptive these models are. For instance, modeling the progress of memes within discourse may allow us to predict the duration or extent of transmission for some elements of content. Or we may be able to predict the success of certain elements of content without actually having to introduce them. But further,
we hope that these models will help us understand natural selection in general. That is, our models will help in building (at least by providing a case study) a more robust and general theory of evolution.

1.7 Organization of Chapters
In the next chapter we overview the USENET News system, describing its properties and interfaces. We devote Chapter 3 to a description, using the vocabulary of text analysis, of our software systems that form the foundation of the ecological model. In the chapter that follows we recast much of the text analysis mechanisms into a biological model of discourse. We go on to make a number of in-principle arguments as to why our methods are soundly connected to their biological counterparts. In Chapter 5 and 6 we describe the results of some static analyses of posts to NetNews; we’ve investigated the population fitness of posts within a corpus as well as their diversity. Chapter 7 is devoted to a brief description of three NetNews visualization tools. We end with our conclusions and suggestions for future work.
Chapter 2

An Overview of NetNews

USENET News, or NetNews, provides an excellent setting for our corporal ecology due mostly to a number of practical concerns. First, it is a set of documents through which runs a precise arrow of time; each document is timestamped when posted. Second, all documents are public and readily available in digital form. Third, the defined set of newsgroups provides an ideal environmental setting in which to study the interaction of documents. Fourth, the sheer volume of posts (thousands per day) means our analysis is not easily swayed by small local effects. And finally, the in-reply-to structure (described below) means that many documents are posted as essentially modified “descendents” of previous documents.

NetNews originated in 1979 as a software mechanism to distribute amongst computers connected to the early Internet, “bulletins, information, and data... items of interest such as software bug fixes, new product reviews, technical tips, and programming pointers, as well as rapid-fire discussions of matters of concern to the working computer professional,” (Kantor 1986). This distribution was for the benefit of the members of the ARPA-Internet community and within the first year fifty UNIX sites were participating. Like all of the Internet, NetNews is defined solely by its protocols (rather than by ownership or licensing or governance). The Network News Transport Protocol (NNTP) stipulates how NetNews messages are posted, distributed, and retrieved over the Internet (Kantor 1986). A further Internet memo specifies the actual format of each NetNews message (Horton 1987).
The collection of messages over NetNews are organized into subject groups, called newsgroups. These subjects, in turn, are organized in a tree-like hierarchy. At the top of the hierarchy are a collection of broad categories, most notably:

- alt: Alternate groups
- bit: Gatewayed BITNET mailing lists
- comp: Computer professionals and hobbyists
- misc: Groups not fitting anywhere else
- news: USENET News network and software
- rec: Hobbies and recreational activities
- sci: Research/applications in established sciences
- soc: Social issues and socializing

Top-level categories of local interest may also be created, for instance, “mit” for news items of interest to the MIT university community. Underneath each top-level category exists newsgroups as well as possible further hierarchical categorization. A newsgroup name is defined as the entire path from the top-level category through any subsequent refining categories down to the name of the group itself. Category and group names are delimited by the period symbol. Thus, “sci.physics” is the name of a scientific oriented newsgroup devoted to general physics subjects. However, “sci.physics.optics” is a more specific group devoted to the study of optics (see Figure 2.1). There exists a voting mechanism in which new groups are proposed and approved for addition to NetNews (though groups within the “alt” domain require no vote to be created). Today there are thousands of newsgroups dealing with every possible subject matter and each day many thousands of documents are posted to the system.

Users access NetNews through one of any number of news reading software systems. The systems all offer a few essential features: A user subscribes to those newsgroups that are of interest to them. The news reader will then keep track of messages sent to these groups and notify the user when new messages have arrived. The user can read posts sent
to these newsgroups and can post new messages as well. Posted messages are transmitted, via the NNTP protocols, to other USENET users across the Internet.

![Diagram of NetNews newsgroup hierarchy]

**Figure 2.1:** A small portion of the NetNews newsgroup hierarchy.

Posts are composed of a number of fields, only a few of which are relevant here. The user creating the post is responsible for the post "body," that is, the actual text of the message, as well as a subject line. The subject line is composed of a few words which describe what the post is about. NetNews software will append to posted messages a number of additional fields including a timestamp and the user name of the person who created the post.

From: bill@osl.or.gov (Bill Smith)  
Newsgroups: sci.military.naval  
Subject: wrecks to locate?  
Date: 26 Sep 1995 02:17:05 -0700

Should all of the ships sunk at Pearl Harbor and Tranto been left there because they went down with casualties aboard?

**Figure 2.2:** A fictitious example post sent to the sci.military.naval newsgroup along with some of its header information.

Posts can be either an independent message or a follow-up to a previous message. A follow-up, or "in-reply-to" message, will have special threading information in its header linking it to the previous posts to which it is a reply. This header information allows news readers to reconstruct the discussion thread. Further, in-reply-to messages will by default
share the same subject line as the original message (though the poster of a follow-up can choose to change this).

Users who post to NetNews often send their message to just a single newsgroup. However, it is possible to send a single message to multiple newsgroups. Called “cross-posting,” this is generally done when a message is relevant to multiple lists and is a way to broaden the possible readership.

NetNews today has grown considerably from its humble beginnings in the late 70’s and 80’s. Consider these example statistics generated for the NetNews system which services the Iowa State University (ISU) for the week ending March 1, 1996.

- Last week, over 83000 new articles came in from the network for 3512 newsgroups each day.

- Last week 1786 different news reading client systems including microcomputers, workstations, and mainframes connected to the ISU news server to read articles.

- The news server (on the average) services over 185000 requests for an individual news article each day.

- The [top] newsgroups on the Iowa State campus (in order of weekly readership) last week were:

<table>
<thead>
<tr>
<th>Newsgroup</th>
<th>Reads-per-week</th>
</tr>
</thead>
<tbody>
<tr>
<td>isu.market</td>
<td>5043</td>
</tr>
<tr>
<td>alt.binaries.pictures.supermodels</td>
<td>3973</td>
</tr>
<tr>
<td>rec.equestrian</td>
<td>1344</td>
</tr>
<tr>
<td>alt.binaries.pictures.erotica</td>
<td>855</td>
</tr>
<tr>
<td>isu.events</td>
<td>747</td>
</tr>
<tr>
<td>alt.binaries.pictures.nude.celebrities</td>
<td>737</td>
</tr>
<tr>
<td>alt.chinese.text</td>
<td>682</td>
</tr>
<tr>
<td>alt.support.diet</td>
<td>672</td>
</tr>
</tbody>
</table>

Iowa State University 1996

In our studies we concentrate on posts to the “sci” family of newsgroups. This is an active collection of newsgroups devoted to established scientific disciplines.
Chapter 3

The Text Mechanisms

Our overreaching goal is to analyze posts to NetNews in order to distill from each post those most salient memes, determine the activity of the memes across the corpus of posts, cluster posts into semantically related groups, and finally analyze their group behaviors. The initial analysis of posts is performed by a large software system which inputs the full-text of posts, creates a quantitative representation of each post, and clusters posts together. The output of this system is then fed into subsequent modeling and analysis software.

In this chapter we will give a complete description of the text analysis techniques without drawing connections to their biological or memetic analogues. Only after the techniques have been thoroughly treated do we recast the mechanism, in the following chapter, in biologically inspired language. We offer two modes of description because we believe that the text analysis techniques described herein will have practical applications to Net based technologies that are not directly related to our research into population memetics. We hope that readers will appreciate the text analysis techniques as potentially separable from the rest of the work.

3.1 System Overview

The system currently works off-line; that is, we begin with a collection of posts and our computations are only with respect to this fixed set of posts. In the future, we hope the system will support real-time on-line processing of posts.

We employ a number of fairly traditional text retrieval techniques which read and convert each post into a vector representation based on word occurrences. We then perform an eigen-analysis of these vector representations which distill from the corpus those most sta-
tistically relevant word co-occurrences. (By *co-occurrence* we mean simply that the words occur within the same document; in contrast, *collocation* often requires that words occur within some distance of each other in text.) Finally, we use a traditional clustering algorithm to group like-minded posts together. Figure 3.1 shows each of these main steps which we shall examine in turn.

![Figure 3.1: Major functional units of post analysis mechanism.](image)

### 3.2 Stop Words and Stemming

We begin with a pre-chosen set of posts to NetNews and each post is initially read and individually processed. The words from both the post body and the subject line are combined into a single word list. Our first step is to remove from this list those words that are so common in English as to carry little or no semantic content; in particular, we wish to remove function words, pronouns, and other common content words. Examples of such words are “and”, “is”, “I”, etcetera. This collection of common words is generally called a *stoplist* in the text retrieval community (Fox 1992). Our stoplist was developed in two steps. First we used the lexicon of Karttunen’s large morphological analyzer and part-of-speech tagger to identify the function words in English (Karttunen 1983). Next a list of high frequency words were assembled from a word frequency analysis of the Brown corpus, a well known and often studied collection of texts in English (Francis and Kucera 1982). The stoplist which resulted from these steps contains 270 common words.
Those words from a post that make it past the stoplist are then passed through a stemmer. The process of stemming attempts to conflate morphologically similar words into single terms by removing suffixes and prefixes and normalizing tense (so that “eating” becomes “eat” or “traveler” becomes “travel”) (Frakes 1992). The goals here are twofold - first, this step makes explicit the strong semantic link between the various forms of a root word. Moreover, this step reduces the total number of words that will need to be considered in subsequent analysis. We are currently using a very simple stemmer algorithm due to Porter (Porter 1980). The Porter algorithm contains a set of rules for locating and stripping suffixes (such as “ing”, “ed”, “s”). The algorithm does not attempt to correct word spellings after the suffix is stripped (for instance “running” will become “runn”). However, evidence from the text retrieval community suggest that as long as the stemming is applied in a consistent fashion across all words this should not be a problem or overly reduce the benefits of conflation (Smeaton 1995).

In the next step we combine the terms from all the posts in the corpus and lexicographically sort the resultant list, removing any duplicates. We simultaneously count up the number of times that each particular term occurs across the entire corpus of posts. In our continuing efforts to shorten the list of words, we remove any words which do not occur in at least some minimum number of posts; this minimum threshold is currently set at 3. This step helps to remove misspellings, nonsense words, and words that are so obscure as to have no discrimination value. This final set of sorted, stemmed, and pruned words makes up what is called in the literature the term list.
3.2.1 Example
Consider the post shown in Figure 3.2.

From: joe@foo.com (James Smith)
Newsgroups: sci.military.naval
Subject: Re: wrecks to locate?
Date: 26 Sep 1995 04:08:22 -0700

bill@osl.or.gov (Bill Smith) writes:

> Should all of the ships sunk at Pearl Harbor and
> Tranto been left there because they went down
> with casualties aboard?

Arizona HAS BEEN left at Pearl Harbor...

The others were salvageable and the US used them to fight the Japanese.

James Smith
Cypress, CA

Figure 3.2: Text of a fictitious post to sci.military.naval. The emboldened text represents those parts used by the text analysis software.

This message is in-reply-to an original posting by Bill Smith and the first three lines of the text body are quotes from the original post of Bill’s (the quoted text can be identified by the ‘>’ symbols). Only the subject line and the text of the posting are passed through the stoplist and stemmer. The resultant list of terms after this analysis is:

<table>
<thead>
<tr>
<th>aboard</th>
<th>locat</th>
</tr>
</thead>
<tbody>
<tr>
<td>arizona</td>
<td>other</td>
</tr>
<tr>
<td>bill</td>
<td>pearl</td>
</tr>
<tr>
<td>ca</td>
<td>re</td>
</tr>
<tr>
<td>casaulti</td>
<td>salvag</td>
</tr>
<tr>
<td>cypress</td>
<td>ship</td>
</tr>
<tr>
<td>fight</td>
<td>smith</td>
</tr>
<tr>
<td>harbor</td>
<td>sunk</td>
</tr>
<tr>
<td>jam</td>
<td>went</td>
</tr>
<tr>
<td>jpanes</td>
<td>wreck</td>
</tr>
<tr>
<td>left</td>
<td>write</td>
</tr>
</tbody>
</table>

This makes up the entire term list for this post.

3.3 Vector Space Representation
The term list is the key to creating a vector representation of each post. Consider a list of n terms, where n is the number of terms for the entire corpus. Each post is represented
within this \( n \)-dimensional vector space as a length \( n \) term vector. This vector space representation has been used extensively within the text retrieval community (Harman 1992; Salton and McGill 1983). We are left with selecting a method to compute the values for each vector; that is, how should we define the term weighting. The text retrieval community has shown that both the frequency of the term within a particular document as well as the frequency of the term across all the documents is useful in determining the weight assigned to each element in the vector (Salton and Buckley 1988).

The within document frequency, or term frequency, acts as a good measure of the salience and relevance of a term within a post. Simply put, the more frequently a word occurs in a document the more relevant the word seems to be. As is generally the case, we determine the term frequency by simply counting up the total number of occurrences of a term in a post. However in our system, those words from the subject line are given a higher weighting then those from the post body.

It is also common practice to include information about the overall frequency of the terms, across the entire corpus, in the weighting. If a term is extremely frequent in the entire corpus of posts it is unlikely to have high discrimination value even though it may be of high frequency in a particular post of interest. For computing the corpus-wide term frequencies we use the well-known inverse document frequency (IDF) (Salton and Buckley 1988). Consider a corpus of \( m \) posts and a particular term, \( j \), within the list of \( n \) terms. Then the IDF is given by,

\[
\text{IDF}_j = \log\left(\frac{m - m_j}{m_j}\right). 
\]

where \( m_j \) is the number of posts in which term \( j \) appears. The term weight for a document, \( i \), and term \( j \) is then defined by,

\[
\text{TermWeight}_{ij} = w_{ij} = \log(\text{TermFrequency}_{ij}) \cdot \text{IDF}_j. 
\]
Each term weight, then, is a combination of the inter- and intra-document frequencies.

Each post, \( i \), is now represented by a particular term vector,

\[
\mathbf{r}_i = (w_{i1}, w_{i2}, \ldots, w_{in}).
\]  

(3.3)

The entire collection of \( m \) term vectors, one for each post, define the \textit{term/document matrix}, \( A \),

\[
A = \begin{bmatrix}
\mathbf{r}_1 \\
\mathbf{r}_2 \\
\vdots \\
\mathbf{r}_m
\end{bmatrix} = \begin{bmatrix}
w_{11} & w_{12} & \cdots & w_{1n} \\
w_{21} & w_{22} & \cdots & w_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
w_{m1} & w_{m2} & \cdots & w_{mn}
\end{bmatrix}.
\]  

(3.4)

### 3.3.1 Example

We will construct a small example to illustrate how the term/document matrix is computed. Consider a corpus of three posts with the following very small term lists. (The number next to each term is the term frequency, the number of times the term occurred in the post.)

<table>
<thead>
<tr>
<th>post 1</th>
<th>post 2</th>
<th>post 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>harbor</td>
<td>harbor</td>
<td>left</td>
</tr>
<tr>
<td>japan</td>
<td>pearl</td>
<td>locat</td>
</tr>
<tr>
<td>pearl</td>
<td>salvag</td>
<td>write</td>
</tr>
<tr>
<td>ship</td>
<td>sunk</td>
<td>2</td>
</tr>
</tbody>
</table>

It should be easy to verify that the term list for the entire corpus has nine entries: harbor, japan, left, locat, pearl, salvag, ship, sunk, write. Due to the stemmer, some of these terms are not English words. (Note that for this example we are not requiring a term to occur in more then one post to be included in the term list.) Equation (3.1) above describes how to compute the IDF for a term. The resultant IDF for each term (where we consider \( \log(0) = 0 \)) is:
For each post we compute the term weights by multiplying the above IDF values with the integer floor of the base 2 logarithm of the term frequencies. The resultant term/document matrix is given by:

\[
A = \begin{bmatrix}
  r_1 \\
  r_2 \\
  r_3
\end{bmatrix}
= \begin{bmatrix}
  0 & 0 & 0 & 0 & 2 & 1 & 0 & 0 \\
  0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
  0 & 0 & 1 & 0 & 0 & 0 & 0 & 2
\end{bmatrix}
\]

Here the rows represent the nine terms alphabetically from left to right. Admittedly, the size of our corpus produces a term/document matrix with little clear discrimination power.

### 3.4 Eigen-text

The previous set of steps, culminating in the term/document matrix, are the basis for much of modern text retrieval or filtering (Salton and Buckley 1988; Frakes and Baeza-Yates 1992). In our system we take a further step wherein we attempt to distill the high-order structures of this matrix. That is, we find the statistically important associative relationships of term co-occurrences. This is done through an eigen-analysis process called singular value decomposition or SVD.

Matrix decomposition techniques, such as SVD, are employed generally for two purposes -- data reduction (compression) and data interpretation. We expect to be dealing with very large and quite sparse term/document matrices (upwards of tens-of-thousands of posts and terms) and therefore need some data compression to insure that our continued analysis remains computationally tractable. But more importantly, in our further analysis we will make use of the salient conceptual structures present in the term/document matrix.
Applying SVD to the term/document matrix (a technique we call eigen-text) both compresses the data and distills out the salient underlying semantic structures.

The use of SVD for text-retrieval applications was originally proposed and has been extensively studied by Susan Dumais of Bell Communications Research and her colleagues (Dumais 1992; Dumais 1993; Deerwester, et.al. 1990; Furnas, et.al. 1988). They refer to this technique as latent semantic indexing (LSI). Peter Foltz has actually investigated the use of LSI in clustering NetNews articles for information filtering (Foltz 1990). Michael Berry and co-authors have researched a variety of numerical approaches to efficiently perform SVD on large sparse matrices such as those found in text retrieval (Berry and Fierro 1995; Berry, et.al. 1993; Berry 1992). Our approach follows the LSI approach closely, though we use their methods in novel ways.

3.4.1 Overview of SVD

The following exposition tries to give a brief overview of the linear algebra behind SVD and follows, for the most part, the notation of Berry; the interested reader is directed to any of a number of introductory texts on the subject (such as Anton 1981).

Consider the following decomposition of the \((m \times n)\) term/document matrix, \(A = U \Sigma V^T\) (3.6)

Here, \(U^T U = V^T V = I_m\), where \(I_m\) is the \((m \times m)\) identity matrix; in other words, \(U\) and \(V\) are both orthogonal. Further, \(\Sigma = \text{diag}(\sigma_1, \sigma_2, ..., \sigma_n)\), is the diagonal \((n \times n)\) matrix. Such a construction, with both a left and right orthogonal matrix and a middle diagonal matrix, is referred to as a two-sided orthogonal decomposition. We will call each of the diagonal elements, \(\sigma_i\), a singular value, which along with its two corresponding column vectors make up a singular triplet, \((u_i, \sigma_i, v_i)\). Note that the singular values represent the nonzero square-roots of the eigenvalues of \(AA^T\), the column vectors of \(U\) are the orthonormalized eigenvectors of \(AA^T\), and the column vectors of \(V\) are the orthonormalized eigenvectors of \(A^T A\).
We are interested in the rank of matrix $A$; in other words, we are interested in the number of dimensions the data spans. We know that the dimensions of $A$ are $(m \times n)$. Consider the $m$ row vectors taken as a group,

$$A = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{bmatrix}. \quad (3.7)$$

We wish to find the subspace spanned by this set of vectors, that is, we wish to find the space $\mathbb{R}^R$ such that all possible vectors in $\mathbb{R}^R$ can be written as linear combinations of our $m$ row vectors, $r_1, r_2, \ldots, r_m$. Clearly, $(R \leq n)$; our vectors will span at most the entire space of $\mathbb{R}^n$. The linear space spanned by a collection of row vectors is referred to as the *row space*, and the dimension of this space is referred to as the *rank* of the matrix, $R = \text{rank}(A)$. (Note that the forgoing comments would hold if "column" were substituted for all occurrences of "row".)

An important aspect of the singular value decomposition described above is that it is said to be *rank revealing*, that is, the analysis determines an approximate value for the rank of the matrix. While we described in Equation (3.6) an exact decomposition into left and right matrices of dimensions $(m \times n)$, this is not what in practice occurs. Instead, we keep only the top $k$ column vectors which correspond to the $k$ nonzero eigenvalues of $A$, where $k < (R = \text{rank}(A))$. The $k$ we determine through SVD nicely approximates the true rank of the matrix $A$ and thus the set of column vectors should approximately span the entire original set of vectors. These resultant vectors are orthogonal and thus linearly independent forming an approximate *basis*.

The decomposition is now formulized as,

$$A_k = U \Sigma V^T = \sum_{i=1}^{k} u_i \cdot \sigma_i \cdot v_i^T. \quad (3.8)$$
where $A_k$ is the rank-$k$ approximation to $A$. Note that while we expect $A_k$ to be a good approximation to the original $A$ we do not want it to be exact since this process is used to remove noise from the original matrix while keeping the most salient data. In other words, we make use of the structure in the term/document matrix to re-express its data in a more parsimonious fashion -- one in which this structure is brought to the surface unobscured.

$$A_{m \times n} = U_{m \times k} \Sigma_{k \times k} V^T_{k \times n}$$

**Figure 3.3:** Decomposition of term/document matrix into rank-$k$ approximation

Considering the term/document matrix we are reminded that a relatively minuscule percentage of all of the terms in the corpus appear in any particular document. Therefore, we expect the matrix to be extremely sparse. Moreover, we anticipate the matrix to have a fair amount of underlying structure due to word co-occurrences within the posts. More formally, we expect the term/document matrix to have a very low rank relative to its dimensions, and therefore, $k < R = \min(m, n)$. In practice, the dimensions of $A$ are generally on the order of tens- or hundreds-of-thousands while the rank is on the order of just hundreds. The SVD is, therefore, *low-rank revealing* and these top $k$ vectors will provide a good approximation of the fundamental numerical subspaces present in matrix $A$.

But what exactly do these numerical subspaces represent? They are vectors that define linear combinations of either terms (for the right orthonormal matrix) or documents (for the left matrix). We are interested in only the right matrix which defines term-subspaces. Each term-subspace describes a set of semantically significant associative patterns in the terms of the underlying corpus of documents: we can think of each subspace as a *conceptual index* into the corpus (see Figure 3.4) (Furnas, et.al. 1988).
3.4.2 Example

We have examined a small corpus of posts for test purposes. The collection is comprised of 784 posts composed of 82 posts to sci.military.moderated, 490 to sci.military.navy, and 212 posts distributed amongst the groups sci.psychology, sci.psychology.theory, sci.psychology.personality, sic.psychology.psychotherapy, sci.psychology.journals.psycholoquy, sci.psychology.misc, sci.psychology.announce, and sci.psychology.research. The posts were made during the month of September, 1995. Of all the words in the documents, 5162 terms made it through the stop-list and stemming process. Thus, the final term/document matrix was 784 x 5162 elements in size. Each document, on the average, was composed of 74 terms; and each term on average appeared in 6 documents. The term/document matrix was input into the SVD software which decomposed the matrix into a rank-265 approximation. Considering just the right matrix, each column vector describes a set of associative relationships of words as a weighted linear combination of the original terms. Generally, each singular vector has only a small number of terms of significant weight. Thus, when describing each singular vector (term-subspace) we consider only those terms with weights above some particular threshold, normally 0.1 (recall that these vectors are unit normalized).

![Figure 3.4: Most significant weights in the vector (shaded) represent the salient terms.](image)

We will list the salient terms (with weights greater than 0.1) from five example term-subspaces of this corpus. That is, we will list those most significant terms for these singular vectors. Each singular vector or term-subspace describes a statistically significant set
of term co-occurrences. Remember that these documents were either posted to military or psychology newsgroups:

- harbor, japan, pearl
- food, maze, rat, reinforce
- ai, obsolete
- airforce, arsenal, tomahawk
- explode, meltdown, nuclear, russia, sub

### 3.5 Cosine Similarity Measure

Having performed the SVD on our term/document matrix, the next step is to use the resultant matrices to cluster related posts. Grouping documents based on semantic relatedness is a bread-and-butter problem of the text retrieval community (Harman 1992). The general model is to input a query from the user, build a vector-space representation of the query itself, and finally to find documents in the corpus whose similarity measure is closest to the query. We are interested in a problem related but not identical to this; we would like simply to group related documents together based on some similarity measure but without any query.

Before we can do that, we need to devise a scheme to relate the vectors from the singular value decomposition directly back to the set of original post vectors. The LSI approach proposes using the right orthogonal matrix directly when comparing documents for their similarity (Furnas, et.al. 1988). They compare two documents by measuring the similarity between the two document’s right orthogonal matrix row vectors. They make the following algebraic observation,

\[ A_i A_j^T = (U \Sigma V^T)^T U \Sigma V^T = V \Sigma U^T U \Sigma V^T = V \Sigma \Sigma V^T = (V \Sigma)(V \Sigma)^T. \]  (3.9)

This shows that two documents can be compared by measuring the similarity between their rows in the right matrix weighted by the singular values, \( V \Sigma \).
We, however, have chosen to compute the inner product of the original document row vectors with the term-subspaces directly, and to then measure similarity between these resultant vectors. To wit, we perform the following matrix multiply,

$$\hat{A}_k = AV.$$  \hspace{1cm} (3.10)

We transform the sparse \((m \times n)\) term/document matrix, \(A_k\), into a dense \((m \times k)\) matrix, \(\hat{A}_k\). Each document is now represented by a \(k\)-length vector wherein each element specifies the significance or salience of each term-subspace to that document (e.g. "harbor, japan, pearl" is important but "ai, obsolete" is not).

Before we can group each \(k\)-length document vector we must decide on a method of measuring distance between them. We employ a similarity measure well known within the text-retrieval community, namely the cosine measure (Harman 1992; Salton and Buckley 1988).

![Figure 3.5: The cosine measure](image)

The measure is aptly named since it computes the cosine of the angle between its two arguments. Consider two documents represented by their \(k\)-length vectors, \(f\) and \(g\). The
cosine measure is defined as,

$$\text{cosine}(f, g) = \frac{\sum_{i=1}^{k} (f_i \cdot g_i)}{\sqrt{\sum_{i=1}^{k} (f_i)^2 \cdot \sum_{i=1}^{k} (g_i)^2}}. \quad (3.11)$$

The numerator should be recognized as the well-loved vector inner product. The denominator is the product of the vector norms and thus serves as a length normalization term. Recall that while the right matrix, $V$, is orthonormal, the $k$-length row vectors of $\hat{A}_i$ are not. In fact, their vector norm is a measure of the actual length of the post. In general, we do not want to consider text length when determining the similarity between two documents; therefore, the additional normalization term of the cosine measure is important.

### 3.5.1 Criticisms of Cosine

A criticism we have of the cosine similarity measure is that it is not a metric in the formal sense. Recall that a metric, $\Delta$, is a distance measure that meets the following properties:

$$\Delta(f, g) \geq 0 \quad (3.12)$$
$$\Delta(f, g) = \Delta(g, f) \quad (3.13)$$
$$\Delta(f, f) = 0 \quad (3.14)$$
$$\Delta(f, g) + \Delta(g, h) \geq \Delta(f, h) \quad (3.15)$$

The requirement of commutativity, Equation (3.13) above, is the only property maintained by the cosine measure. Because the term-subspace/document matrix may contain negative elements we cannot guarantee that all vectors will be in the first quadrant and that the cosine therefore will be positive. Since the cosine is then over $[-1, 1]$ the non-negative distance required of property (3.12) is not met. (In our analysis we currently are ignoring the
rare case when angles are greater than 90° and therefore produce negative similarity measures.) Nor can we guarantee mathematical admissibility as described by Equation (3.15). Moreover, since the smaller the angle between two vectors is, the larger their cosine, the “spirit” of Equation (3.15), that you can never un-do distance traveled, does not make particular sense. Finally, we note that the cosine of a vector with itself is not zero as required by Equation (3.14), but is one.

In our analysis we have found particular instances where the cosine measure has caused problems because it does not follow the above properties of a metric. We feel that the inverse cosine function would be a better similarity measure for our work. It should be easy to verify that the principal value of the arc cosine \(\text{acos}\) meets all of the requirements of a true distance metric set forth above.

A second criticism we have of the cosine metric (which also is addressed by the inverse cosine) is its lack of sensitivity to small angles. That is, the spacing of values along the circle is not uniform. This is not relevant if we only use the similarity measure in comparisons against itself. However, if we employ the measure for further analysis it may become an issue. The inverse cosine metric, we note, is indeed uniform across all angles.

**Figure 3.6:** The cosine measure is less sensitive to smaller angles. For an identical sized domain of angles the range of possible outputs is smaller.
We make use of both the cosine measure and arc cosine metric in our analysis and will specify which is being used at the appropriate times.

3.6 Nearest Neighbor Clustering
The final step is to cluster related documents for conceptual similarity. For this, we use their dense term-subspace vector and the cosine similarity measure. We employ a well-known technique called the Nearest Neighbor Clustering Algorithm (Jain and Dubes 1988). This algorithm requires the initial input of a distance threshold, \( t \). This threshold specifies the maximum allowable distance between two vectors assigned to the same cluster. Due to the properties of the cosine measure the distance between two vectors must be larger than the threshold to allow grouping within one cluster.

The algorithm is rather simple. It considers in turn each row vector from the dense matrix, \( \tilde{A}_k \). The current vector is compared to each vector already assigned to a cluster. If the closest of such vectors is not farther than the threshold distance then the current vector is assigned to that cluster. Otherwise the current vector is assigned to a new cluster. This continues until each and every vector is assigned to a cluster.

In our current datasets we have discovered an appropriate threshold, \( t \), through trial and error. In the future we would like to consider using techniques which do not require a user supplied threshold. Given a threshold distance, \( t \), and the cosine similarity measure, we note that each cluster traces out a multi-dimensional cone in \( k \)-space. That is, all lines contained within this cone will describe an angle with all other lines contained within the cone whose cosine is no larger than \( t \). We call such an object a hyper-cone.

We often represent a cluster of vectors by their centroid vector or center-of-mass. The centroid vector is trivially computed by averaging the value at each element for all of the
vectors in a particular cluster. We will consider the role of the centroid vector in the next chapter.

![Image of centroid vector and hyper-cone]

**Figure 3.7**: The centroid vector radiates from the cluster center-of-mass.

### 3.7 Direction Cosines

So far we have represented each post by its $k$-length document vector. Those vectors make up the rows of the term-subspace/document matrix, $\mathbf{A}_k$. We now will define an alternative and related vector representation. Consider the angles a vector makes with each of its axes. For instance, the vector in $\mathbb{R}^2$, shown in Figure 3.8, describes an angle $\alpha$ with the $x$-axis and an angle $\beta$ with the $y$-axis. These are called the *direction angles* of the vector and the cosine of such angles are called the *direction cosines*.

![Image of direction angles with x- and y-axes]

**Figure 3.8**: Direction angles with the $x$- and $y$-axes of a vector in two space.
In general, the direction cosine for some dimension, $i$, is computed with the following formulae,

$$\text{direction cosine}_i = \frac{f_i}{|f|} = \frac{f_i}{\sqrt{\sum_{j=1}^{k} f_j}}.$$  

(3.16)

There exists a very simple physical relationship between the term-subspace vector representation and the direction cosine representation. As can be seen from Equation (3.16), the direction cosine vector is simply the unit-normalized version of the term-subspace representation. This representation will be particularly useful when we wish to remove from consideration the length of each post. The cosine similarity measure automatically performs this normalization (since it too has the vector norm in the numerator). Thus, the direction cosine representation is only of use when we do \textit{not} employ the cosine similarity measure.
Chapter 4

The Biology of Eigen-Text

*Ontology recapitulates philology.*

James Grier Miller (quoted in Quine 1960)

4.1 Biological Connections

The mechanisms described in Chapter 3 identify related documents within a corpus of posts. While these techniques are useful for text retrieval, we have built them for the purpose of studying corporal ecologies and population memetics. In our description of the text analysis system we did not make any biological connections; we now will draw some of these connections quite explicitly. Note that this is necessarily a discussion *in-principle*, in future chapters we demonstrate ways to build empirical and experimental evidence for our claims here.

4.1.1 Ecological mosaic

The general definition of an ecology places a collection of interacting organisms within some environmental setting such that the organisms compete for some limited resource. In our corporal ecology each post is an individual organism, the newsgroups of NetNews define the environment, and the human poster is the scarce resource. As mentioned before, we never explicitly model the human poster. But implicit to our understanding of selection amongst the post organisms is an assumption that competition for a limited number of human posters is under way. It is through the recruitment of human posters that organisms and their constituent memes are replicated and survive.
Note that newsgroups describe a rather heterogeneous environment. It follows, then, that the type of organisms that proliferate in each environment will differ. Environments in the natural world vary both over space (lowlands give way to foothills) and time (erosion slowly advances a beachhead). The same is true with the environment of NetNews. Geographic space is described by the tree-hierarchy of newsgroups; and environmental conditions vary from group to group. Over time the conditions within a single newsgroup change; these changes are a function of the external world, cultural shifts outside of NetNews, the comings and goings of human posters and so forth. As an example, spring break has a tendency to decrease available human-poster resources across much of NetNews.

While some temporal shifts in environmental conditions may seem relatively "smooth" and make up a temporal gradient, the spatial shifts clearly are discrete -- there exists no "middle-ground" between one newsgroup and another. Such a condition is generally referred to as a patchy environment or a mosaic (Emlen 1973). Patchy environments exist in the natural world as well as in cyberspace; for instance, we recognize them as fields interspersed with wooded areas or even abruptly changing soil types.

4.1.2 Polymemes and memotypes

We claim that the post text represents the expressed morphology or phenotype of some underlying structure. It probably comes as no surprise that the particular underlying structure with which we are concerned is the term-subspace vector. This vector defines a memotype -- identical to the genotype of the natural world.

We define the meme as the most atomic replicating unit (like the gene) of the conceptual space. We consider each term from the original term/document vector to be an individual meme. However, as with genes, expressible traits are often made up of meme complexes. Such complexes we refer to as polymemes as they are directly related to the gene complexes called polygenes. Polymemes are described by the term-subspace vectors:
they are collections of many meme loci which all combine together to describe some specific morphological trait. As is generally the case with polygenetic traits in natural biology, polymemetic traits are metric. That is to say, they can take on values along some particular continuum (this should not be confused with a distance metric). Both our term-subspace and direction cosine vectors (two different memotype representations) are composed of weights along the continuum of the real line. Examples of metric traits in humans include height and weight while male pattern baldness is an example of a discrete trait.

4.1.3 Quasi-species and wild types

The notion of a quasi-species is due primarily to Nobel laureate Manfred Eigen (Eigen 1988, 1992). He states that the “quasi-species represents a weighted distribution of mutants centered around one or several master sequences. It is the target of selection in a system of replicating individuals that replicate without co-operating with one another (RNA molecules, viruses, bacteria),” (Eigen 1992, pp 160). One organism is a mutant of another if it is particularly close to the other in sequence-space. By sequence-space, we mean any of the search spaces defined within natural selection by a replicator. Examples of sequence-spaces include the gene space, protein spaces under molecular evolution, and the meme space defined within a corporal ecology.

The quasi-species is conceptually similar to a species insofar as it represents a collection of organisms (the master sequences and their mutants) which are close to one another in sequence-space. The result of this is that members of the same quasi-species share morphological traits. We claim that the populated hyper-cones computed in our nearest neighbor clustering step (Section 3.6) define quasi-species within our corporal ecology. In truth, the clustering operation finds organisms within meme-space which are suitably close together so as to be considered conceptually related. Since post organisms do not co-oper-
ate during replication (for instance there is no sexual reproduction) post clusters are quasi-
species quite exactly.

Eigen goes on to define a wild type as the “best-adapted genotype... of a species living in a natural environment and representing the majority of the individuals of the species. In a quasi-species, the wild type is the centre of the mutant distribution, and is defined by the consensus sequence.” (Eigen 1992, pp. 167). It should be clear that the centroid vector defined in Section 3.6 matches well this definition of the wild type.

4.1.4 Review

Our work explores the connection between theories of natural ecologies and population genetics and our models of corporal ecology and population memetics. Specifically, the main connections are:

<table>
<thead>
<tr>
<th>text analysis term</th>
<th>biological term</th>
</tr>
</thead>
<tbody>
<tr>
<td>post</td>
<td>organism</td>
</tr>
<tr>
<td>newsgroups</td>
<td>patchy or mosaic environment</td>
</tr>
<tr>
<td>post’s text</td>
<td>morphology or phenotype</td>
</tr>
<tr>
<td>term</td>
<td>meme</td>
</tr>
<tr>
<td>term-subspace</td>
<td>polymeme or meme complex</td>
</tr>
<tr>
<td>term weights or direction cosines</td>
<td>metric trait or allele</td>
</tr>
<tr>
<td>hyper-cone cluster</td>
<td>quasi-species</td>
</tr>
<tr>
<td>centroid vector</td>
<td>wild type</td>
</tr>
</tbody>
</table>

4.2 Natural Selection and Replication

We now will make the case that natural selection is actually taking place within our collection of posts and that eigen-text vectors form a reasonable unit of replication.

We again will rely on Manfred Eigen to formulate the necessary and sufficient conditions for natural selection to take place. He makes the following strong claim:

We find that the selection principle is neither a mystical axiom immanent in living matter nor a general tendency observable primarily in living processes. On the contrary, it is -- like many of the known physical laws -- a clear ‘if-then’ principle, that is, a principle according to which
defined initial situations lead to deducible behavior patterns.

Eigen 1992, pp 17-18

In the rest of this section we will discuss the three necessary components for natural selection according to Eigen: self-reproduction, mutagenicity, and metabolism.

4.2.1 Self-reproduction

According to Eigen, the first necessary condition for natural selection is that individual elements within the system be self-reproducing. In other words, there must be a fundamental replicator within the system. We have proposed eigen-text vectors as the replicator within our meme-space. Eigen-analysis falls within a general family of techniques which attempt to distill through statistical methods fundamentally concise representations of order within some system. In fact, our particular technique of singular value decomposition is known to produce the best rank-k approximation to a low-rank matrix (Berry 1992). Since our $k$ is less-than-or-equal to the true rank of the initial matrix we are computing the best approximation of the original data in a concise format. Concision need not be a property of our replicator. However, there is no particular reason to rule concision out. Moreover, we are philosophically attracted to the parsimony implicit with concision; this is a reasonable and simple starting point for a theory of replication.

It is useful to remember that identifying the fundamental unit of replication even within natural biology is a fairly messy business. The “gene” is a wonderfully useful concept and has proven its power over time. However it is not as pristine or rigorously defined as commonly believed. To be sure, chromosomes are composed of strings of nucleotide letters with genes defined as some portion of this chromosomal material. However one cannot claim with certainty where one gene begins, where it ends, and what other genes share its portion of the chromosome. In fact, the “gene” is often burdened with as circular a definition as we here seem to propose for “meme” -- it is simply those complexes of chromo-
somal material that successfully replicate (Williams 1966). If we accept this definition of gene (and it is not really so very controversial) then eigen-text offers an identical definition for the meme. The outcome of the singular value decomposition is those word co-occurrences which are replicating with sufficient success such as to have statistical salience. In other words, eigen-analysis finds groups of terms within the corpus that frequently co-occur. Eigen-text vectors, then, can be restated as, “that which segregates and recombines with appreciable frequency” (Williams 1966, pp. 24), since that very much describes what these co-occurring words are doing across a collection of documents -- segregating and recombining with frequency. This quoted phrase from Williams, however, is one of the classic definitions for the gene.

We further note that this notion of replication, both for the gene and the meme, is true only in the moment. And that in both cases predicting the course of a replicator over significant periods of time is fraught with difficulties.

If Eigen is right, and natural selection is an if-and-only-if proposition of auto-catalysis, mutagenicity, and metabolism, then by demonstrating natural selection with the eigen-text vectors we can conclude that it is a replicator. The real difficulty is arguing that there are not other more “correct” or useful replicating unit. This is not a simple matter insofar as it amounts to arriving at the Mendel’s Law for population memetics. This is a step which Darwin himself did not take; he died without incorporating Mendel’s theories of replication into evolution. We will mention other possible replicators when we discuss future work in a later chapter.

4.2.2 Mutagenicity

Given a replicator, Eigen’s second requirement for natural selection is that the process of self-reproduction have the potential to produce errors or mutations; in other words, copies must be made with less than perfect fidelity. Empirical evidence (along with plenty
of individual allegorical evidence) suggest that the mutation rate when copying text is rather high. Even when the intention is for every copy to be perfect, mutants often occur. In recent history we have all seen how chain email messages often mutate over time even though email systems should quite simply allow for perfect copying. And in the case of NetNews, in-reply-to messages or other posts on some particular topic will normally exhibit mutation-by-design. After all, why else would someone post a new message unless they believed they were offering an incremental mutation on the discussion thread. The study of various manuscript traditions, such as Chaucer or old Norse tales, gives a striking demonstration of the mutation rate of textual discourse across copies (Robinson and O’Hara Forthcoming; O’Hara and Robinson 1993). Dawkins relates one of our favorite textual mutations, this one due to a translation error in the Old Testament of the Bible:

I suppose the scholars of the Septuagint could at least be said to have started something big when they mistranslated the Hebrew word for ‘young woman’ into the Greek word for ‘virgin’, coming up with the prophecy: ‘Behold a virgin shall conceive and bear a son...’

Dawkins 1989, pp. 16

We note, however, an important property of selection amongst quasi-species referred to as the neutral theory (Kimura 1983). In standard evolutionary theory most mutations are bad for the organism while a small percentage actually afford some advantage. However, the neutral theory claims that in some cases many mutations are actually neutral to selection. That is, they afford neither a disadvantage nor an advantage to the organism. The result of this property is that the cloud in sequence-space associated with a quasi-species is large and contains a number of best sequences. A distribution of mutants emerge about the set of master sequences and this in turn guides evolution. Evidence of the neutral theory has been developed for molecular and viral evolution. We believe that the neutral
theory plays a role in memetic evolution as well, and therefore we suspect that memes evolve much more like viruses then like people.

4.2.3 Metabolism

Finally, Eigen’s third requirement is that the system be far from equilibrium and, to survive, capable of receiving inputted energy. In natural biology, the far-from-equilibrium state can be demonstrated by simply pointing to an organism’s metabolism as its means for receiving energy.

For our textual biology, we rely on information theory to make an equally straightforward set of arguments. At equilibrium a system can not originate information (Küppers 1990). So if we can identify any information (in the Shannon sense) within our corpus we will have demonstrated that the system is in a suitably far-from-equilibrium condition. There has been considerable work to build theories of the information content within connected text and we need not reargue those results here. Suffice it to say that the very fact that we can successfully perform an SVD is perforce evidence of Shannon information. However, we note that Küppers has analyzed the information theoretic content of text with the express purpose of evaluating it in the context of natural selection (Küppers 1990).

4.2.4 Summary

Following Eigen, the three main properties required for natural selection are: self-reproduction, mutagenicity, and metabolism. We have argued that eigen-text makes a concise replicator and demonstrates self-reproduction, that text clearly undergo mutations over copies, and that it is fairly straightforward information theory problem to demonstrate that a corpus of posts is not at equilibrium. Therefore, we can conclude that natural selection will necessarily take place within a corporal ecology given eigen-text.

4.3 Fitness Landscapes

The fitness landscape or surface has been one of the most influential tools of modern evo-
olutionary biology. The original idea, proposed by Sewall Wright, is simply to view the reproductive fitness of a wide range of possible gene combinations as a multi-dimensional graph (Provine 1986). While the visualizations are generally restricted to 3D, it is well understood that most sequence-spaces (genes, memes, or otherwise) are of far higher dimensionality. In the general case, the $x$- and $y$-axes represent the space of possible alleles for two particular traits while the $z$ dimension represents the fitness for each pairwise combinations of $x,y$ values. Fitness is generally measured as the relative success, in the presence of natural selection, of replicators across generations. Higher fitness is represented by values approaching positive infinity along the $z$-axis. In the case of eigen-text, the $x$ and $y$ dimensions represent two particular metric traits and larger values correspond to a more pronounced expression of that trait. What exactly the traits are is often not important; they can be thought of as simply two representative term-subspaces.

![Figure 4.1](image)

**Figure 4.1:** A traditional fitness landscape. The $x$- and $y$-axes represent the space of alleles for two traits. The $z$-axis represents the relative adaptive significance.

The fitness landscape provides a valuable conceptual tool for thinking about evolution within some fixed environment. Descent with modification can be thought of as a search through the space defined by the fitness surface. Since selection will favor the hills of the
surface (where fitness is relatively high), evolution's search of the landscape moves the population over time towards the mountain peaks.

How does the fitness landscape differ under our model of a corporal ecology? Figure 4.1 shows the peaks and valleys of the traditional fitness surface. However, the landscapes of our corporal ecologies do not look quite like this due to the particularities of the cosine similarity measure. The cosine measure produces radiating bands of iso-similar surfaces (Jones and Furnas 1987), which we have called hyper-cones. All points that lie upon any straight line emanating from the origin must share an identical fitness value since such points share identical angles with all axes. That is, all lines radiating from the origin of the graph must have the same value in the z dimension. The result is the warping of the traditional mountain landscape into a collection of hyper-cones radiating from the graph origin.

Consider the following two images. Allow the x- and y-axes to represent two particular term-subspaces; larger values on the x- or y-axis correspond to fuller expression of their particular textual properties. In Figure 4.2 the valley in the middle of the graph signifies an environment in which the mutual co-existence of the two traits is selected against. Perhaps
the x-axis is “james smith saint” and the y-axis is “james smith devil”. The environment is selecting for the use of one or the other but not both traits.

In Figure 4.3 we have an environment that selects for the opposite case. Here the adaptive surface demonstrates an environment which affords advantage to the co-occurrence of both traits but offers low fitness to organisms that exhibit only one or the other textual morphology in isolation.

Figure 4.3: An adaptive surface under the cosine metric. Here generality is selected for. The use of both traits in concert offers higher fitness.
Chapter 5

Fitness

We have implemented a software system which performs the text analysis detailed in Chapter 3 and are beginning to build empirical evidence for our models of Chapter 4. To date our modeling has concentrated on the static behavior of a preset corpus of posts, namely on the fitness and diversity. In future work we are interested in studying dynamic phenomena.

5.1 Dataset

Our static experimental dataset was composed of 11,758 posts sent to the “sci” family of newsgroups during the period of September 16 - 28, 1995. These posts were sent to a total of 813 different newsgroups within NetNews. Of those 813 only 179 where actually groups within the “sci” family; cross-postings accounted for the other newsgroups. More then half of the words found in the posts were dropped either by the stoplist or because they did not occur in the minimum number of posts. The total number of terms that made it through this process was 27,031; 14,949 where rejected because they did not appear in at least three posts. Thus the term/document matrix was 11,758 x 27,031. This matrix was sparse indeed; the total number of nonzeros was 897,351. Each post averaged 27 nonzero terms, each term on the average appeared in 76 posts.

We performed the singular value decomposition of this term/document matrix. The SVD code took approximately four hours running on an unloaded 150 Mhz DEC Alpha. Given the size of our original matrix we find this performance impressive. The rank of the resultant term-subspace/document matrix was 209. That is, we compressed the 27,031 term dimensions into 209 term-subspace dimensions. Note that the 11,758 x 209 term-
The subspace/document matrix was completely dense, it contained 2,457,422 nonzeros. Compared to the original number of nonzeros (897,351) it would appear as if the SVD step accomplished no data compression. We note that the sparse matrix format requires directly representing not just the value of each data element but also that element’s position in the matrix (since zeros are not explicitly stored). Taking this into account we see that the representations of the dense and sparse matrices are of nearly identical size. But in point of fact, the dense term-subspace/document matrix is sparser then would appear -- many of the elements in the matrix are near zero and could be well-approximated as zero. This point was made above when we mentioned the small number of truly significant terms in any one term-subspace. It is this preponderance of near-zeros which indicates the degree of data compression and noise reduction resulting from the decomposition.

Our next step was to run the nearest-neighbor clustering code. This took three hours on the same unloaded Alpha (in contrast to 27 hours to run the clustering operation over the original term/document matrix prior to the SVD step). The nearest-neighbor step produced 3,022 total clusters with each post assigned to exactly one cluster. The vast majority of clusters contained ten or less posts as members; only 166 hyper-cones were populated with more than ten posts, twelve were populated with 100 or more posts and the most highly populated cluster had 232 posts (Figure 5.1). This should appeal to an intuition about the traffic of posts over NetNews -- that the large majority of posts are isolated messages or perhaps are related to a rather small number of similar posts. These similarities may be the outcome of responses made to an initial post (in other words may constitute a traditional NetNews in-reply-to thread) or they may be a set of individual isolated posts which are nonetheless strongly related conceptually. Note further that the posts which make up an in-reply-to thread are not in every instance assigned to the same cluster. People say what they have to say, the result being that NetNews in-reply-to threads often deal
with multiple topics from post to post. The topics considered within a thread of posts will often vary considerably across the messages.

![Figure 5.1: Sorted size of the 3,022 clusters in our example dataset.](image)

Unlike the corpus discussed in Section 3.3.1, this corpus of posts spans a wider range of topics. For instance, newsgroups within the “sci” family range from sci.bio.food-science, sci.econ.research, sci.geo.meteorology, sci.polymers, to sci.virtual-worlds. The set of term-subspaces naturally captures a broader range of topics. Some examples include:

- algorithm, fuzzy, genetic, inference, neural
- drink, milk
- energy, solar
- chlorine, depletion, ozone, stratosphere

5.2 Fitness
Recall that the term *fitness* is generally used to describe the relative success, in the presence of natural selection, of replicators across generations. It is the ratio of a set of replicators (for example a genotype or memotype) in one generation to its numbers in the proceeding generation. Depending on the nature of the replicator, a number of fundamental properties will contribute to its fitness. For instance, the fitness of a meme complex,
given our definitions, is a function of its fecundity and copying fidelity. That is to say, the meme will be successful if it is able to replicate itself and if during replication it can hold down its mutation rate to an acceptable level.

In this chapter, however, we will consider the average population fitness within a static snapshot of NetNews, rather than the relative fitness of individual members. Recall that each post cluster represents a quasi-species and that we consider the quasi-species to be the target of selection. To a first order, the average population fitness of a quasi-species’ is given by its relative population size. Put succinctly, species population size is a good first approximation to population fitness. Population size, its weaknesses notwithstanding, has often been shown to be a satisfactory measure of average population fitness (Wallace 1981).

We assume that fitness is causally linked to the selection biases of certain textual morphologies or traits. In other words, if a quasi-species has a large population it is fit; and if it is fit this is due to the expression of certain well-adapted textual traits. We claim that a trait is particularly adaptive if it is accumulating usage beyond what would be expected a priori (e.g. due to chance). In the following sections we identify the quasi-species in our corpus with the highest population size and thus the best average population fitness. We find that the most populated cluster of posts relates to a flame-war around a prolific poster. We then identify the most significant term-subspaces for that quasi-species cluster. As we shall see, one of the most salient term-subspaces calls the prolific poster a “Nazi”.

5.3 Champion Quasi-species
Armed with these simplifying assumptions we are able to determine which quasi-species are most fit and which traits are most adaptive. We will study at some length the most fit cluster from our test dataset. The most populated cluster received 232 posts over the seven
day period from September 20 to 27. In total, 31 newsgroups received one or more posts from this cluster. However, the majority of posts went to a much smaller set of newsgroups primarily made up of alt.folklore.science, alt.astrology, alt.magik, sci.astro, sci.skeptic, sci.history, and sci.med.pharmacy. Almost all of the 232 posts were part of six large in-reply-to threads. However, the system was able to correctly cluster those posts that were not in-reply-to one of these threads but nevertheless were on this topic. Moreover, the system was able to group this entire collection of posts even though the conversation mutated somewhat over time and the threading and subject line changed fairly dramatically.

The main thread of discussion originated with a set of posts under the subject line “Homeopathy for Dummies” and “Homeopathy in TIME Magazine”. The discussion did indeed center at first on homeopathic remedies but the 232 posts which populate this particular cluster deal with a flame-war centered around an individual poster who had been involved in the original discussion of homeopathy. This individual is apparently a wildly prolific author of posts and his style is considered by some to be confrontational and insulting. After time the cluster mutated its subject line so as to make more explicit its concentration on this individual. In fact the largest number of posts in the cluster use the subject line “Is James Smith a Nazi?” (we are using a pseudonym). This subject line also mutated over time; interestingly, one such mutation read, “A Plea for Netiquette (was Re: Is James Smith a Nazi?)”.

It is interesting that this “champion” cluster of posts would be a flame-war. And, moreover, that it would center around an individual poster who was being criticized for authoring an exceptional number of posts himself and for cross-posting to multiple inappropriate newsgroups. While he may have angered other people, he had clearly been acting as an excellent agent in the replication of certain term-subspaces.
It should be noted that in this example, as in other tests we ran, the clustering mechanism did not perform flawlessly. Of the 232 posts assigned to the champion cluster, 21 did not have any readily discernible relationship to the main set of ideas. Some errors can be explained by noting that some of these posts were of such short length that they did not seem to have a sufficient set of nonzero terms to base a clustering on. But some other members of this erroneous group were sufficiently long and yet still mis-classified.

5.4 Adaptive Meme Complexes
We wish to determine which term-subspaces or meme complexes were most salient to this quasi-species. We claim that those most significant term-subspaces are adaptive. To show this, we find those textual traits that are accumulating more usage than would otherwise be expected. We chose to approximate the chance usage of each term-subspace by its mean accumulation of usage across a sufficiently sized corpus of posts. In other words, the mean usage across a collection of posts provides a good approximation to a random accumulation of usage. After subtracting off the mean usage from the term-subspace weights, any remaining significant usage is possibly an adaptation.

For this set of computations, we make use of the direction cosine vector representation. Recall that this representation normalizes the posts to unit length. In Figure 5.2 we see the mean direction cosine computed across the entire corpus for each of the 209 term-subspaces, plotted along with the direction cosine vector of the champion cluster’s wild type or centroid vector. As would be expected, the corpus-wide mean vector is nearly zero for most term-subspaces. A cosine of zero indicates a 90° angle for that axis. That is, it is completely orthogonal to that axis. From the figure we can see that the champion wild type is making heavy usage of a few particular meme complexes. Note also that the right-most term-subspace seems to enjoy considerable usage even from the corpus-wide mean vector. This is because that particular term-subspace corresponds to the first eigenvalue
and therefore is the vector of mean values for each element. We have chosen to ignore those first few term-subspaces in our analysis since heavy usage of these term-subspaces is something of a statistical artifact.

**Figure 5.2:** The mean direction cosine computed across the entire corpus (+) plotted with the champion cluster centroid vector (o).

Figure 5.3 is a histogram of the champion cosine vector minus the mean cosine vector. In other words, it shows how many term-subspaces have a direction cosine of zero, how many have a value of .05, and so forth. Note that the histogram appears to describe a gaussian distribution about 0.0. The preponderance of orthogonal components is an indication of the level of specificity for the cluster. Those few relatively congruent term-subspaces (with cosines as large as 0.2) represent the few most salient meme complexes and given a fit quasi-species are presumably adaptive. Note that the term-subspaces with *negative* direction cosines represent textual morphologies that are actually being depressed. In other words, they represent some textual morphologies that occur often throughout the
corpus but do not occur in this cluster of posts. We have so far ignored such negative textual morphologies but would like to examine this phenomena in the future.

![Figure 5.3: Histogram of the champion direction cosine minus the mean direction cosine. A zero cosine is orthogonal.](image)

What are these adaptive term-subspaces? We have studied the few most salient meme complexes of the champion quasi-species. Not every term-subspace has a readily discernible conceptual meaning. Some seem to have no single set of terms which are significant and may simply represent grammatical structure. Some term-subspaces group together words which seem at the face unrelated and these may represent simply chance statistical correlations. Of the 209 term-subspaces, 89 had no readily discernible conceptual meaning in our estimation.

The top four term-subspaces of our champion cluster had the following discernible meanings:

* faq
* ???
* motive, science
* james, smith, nazi
The “motive, science” term-subspace was relevant to the champion cluster and was also critical to a set of discussions around the Unabomber. The Unabomber discussion was, as it happens, the runner-up to the champion cluster with 217 posts. The term-subspace denoted with question marks had no readily discernible set of terms. The term-subspace, “james, smith, nazi”, has obvious conceptual significance.

5.5 Summary
We have developed techniques to measure the average population fitness of post quasi-species and to find those meme complexes which accumulate significant usage and are therefore adaptive. We studied a corpus of over 10,000 posts and found that a flame-war had the highest population fitness. The meme complexes which accumulated the most usage within that quasi-species included some harsh name calling -- in particular, a poster is called a “Nazi.”
Chapter 6

Diversity

6.1 Levels of Diversity
Diversity is an important concept for the natural ecologist. Within some environment, investigators want to know how many species there are and how the organisms within a species differ. Just so, we are interested in diversity within our corporal ecology. We would like to understand how many quasi-species there are, and within a quasi-species how diverse are the expressed morphologies.

We will consider diversity at two levels. First we consider *ecosystem diversity* by which we mean variation across organisms within one or a collection of environmental conditions. For instance, in the natural world, investigators have studied why rain forests support a much wider diversity of organisms than deserts. Second, we are interested in *population diversity*. Here we mean the variation within a given species; that is, the cloud within sequence-space occupied by that species. Natural biologist study how a species responds to changing environmental conditions by increasing its population diversity. By increasing population diversity a species is able to broaden the evolutionary space it searches. In the natural world an increase in population diversity is affected through increased hybridization, decreased copying fidelity, and other measures.

By investigating diversity within a corporal ecology we discover a number of interesting results. Primarily, we find that as the number of posts for a particular quasi-species increases, the diversity between those posts decreases. That is, as the number of posts on some topic increases, the topical diversity across those posts actually decreases.

The method we apply to compute diversity within our corpus is due to Mark Bedau and co-authors (Bedau, et. al. 1992). They define two types of diversity measures: the
within sequence-space measure and the between sequence-space measure. Within sequence-space variation means that given a particular loci in sequence space what is the variation within that location across a collection of organisms. It is computed as the mean squared deviation between the mean vector averaged over the whole population (or environmental condition) and each individual vector. In contrast, the between sequence-space diversity is the variation between the average value at a loci of an organism and the average value for all loci and organisms. It, then is computed as the mean squared deviation of the mean vector averaged over all populations from the average value averaged over all populations and all sequences. Given a term-subspace/document matrix of size \( (m \times k) \), these diversity measures are defined as follows:

\[
D_{within} = \frac{1}{mk} \sum_{i=1}^{m} \sum_{j=1}^{k} (w_{ij} - \bar{w}_j)^2, \tag{6.1}
\]

\[
D_{between} = \frac{1}{k} \sum_{j=1}^{k} (\bar{w}_j^i - \bar{w}_j)^2, \tag{6.2}
\]

where

\[
\bar{w}_j^i = \frac{1}{k} \sum_{j=1}^{k} w_{ij}, \quad \text{and} \quad \bar{w}_j = \frac{1}{mk} \sum_{i=1}^{m} \sum_{j=1}^{k} w_{ij}. \tag{6.3}
\]

Finally, total diversity, \( D \), is defined as,

\[
D = D_{within} + D_{between}. \tag{6.4}
\]

It is important to develop an understanding of these diversity measures within our corporal ecology. Both the between and within diversity measures are sensitive to variation in post lengths (i.e. the actual size of each document). In our analysis below we compute the diversity using both the term-subspace weight vector representation and the unit normalized direction cosine representation. This latter model removes from consideration post length.
The within diversity measure is also sensitive to meme complex variance across the post vectors. In other words, the within diversity will be large if there is a large diversity from vector to vector in what term-subspaces are most salient (Figure 6.1).

Figure 6.1: The within diversity for these two vectors is high. Vector $f$ has a strong component only with the $x$-axis whereas vector $g$ is weighted towards the $y$-axis.

In contrast, the between diversity is sensitive to topic specificity. Given a particular vector, if most every term-subspace weight is about equal then this measure of diversity will be low (each trait will be expressed with nearly identical strength). However, if there are strong differences between a sizeable set of term-subspace weights then between diversity will be large (Figure 6.2).

Figure 6.2: Vector $g$ has high between diversity since its $x$ and $y$ components are very different. Vector $f$ has low between diversity since its components are nearly equal.

6.2 Ecosystem Diversity
We first will consider the diversity level across different ecological patches -- that is, from one newsgroup to another. We might expect diversity across newsgroups to remain rela-
tively constant: we have no particular reason to expect one newsgroup to have a much larger diversity than another. We have measured the total diversity, \( D \), for each of the 179 different “sci” newsgroups posted to from our corpus of 11,758 posts. In Figure 6.3 we see the diversity for the 179 different newsgroups sorted from least to greatest. Note that this is a semilog graph; the diversity across newsgroups spans four orders of magnitude. The newsgroup with the highest total diversity was sci.answers whereas the newsgroup with the lowest diversity was sci.metallurgy. It is appealing to construct an explanation for this -- sci.answers is naturally a more “general” newsgroup compared to sci.metallurgy.

![Figure 6.3: Sorted total diversity for each of the 179 “sci” newsgroups.](image)

Recall that our measure of total diversity includes post length diversity, topic variance, and term-subspace weight diversity. We would like to discover which of these particular properties are accounting for the wide range in observed ecosystem diversity. If we compute diversity using the direction cosine vector representation which is length normalized we can determine to what extent post length is contributing to this diversity variance. But, before computing diversity on the normalized vectors we are curious what actually is the degree of post length variance within the corpus. Figure 6.4 shows a histogram of the doc-
ument lengths for all 11,758 posts. This image maps very closely to the F probability density function, a member of the family of beta distributions. The typifying properties of this distribution are a sharp and somewhat normal initial density along with a very long tail to the right.

Figure 6.4: Histogram of document length for all 11,758 posts.

Figure 6.5 shows the sorted total diversity, $D$, for the normalized direction cosine representation. We see the dramatic reduction in diversity variance, from nearly four orders of magnitude to much less variance. Clearly, the process of unit normalization seems to
squeeze out much of the ecosystem diversity; most diversity is due to post length variance across newsgroup environments.

Figure 6.5: Sorted total diversity of the unit normalized direction cosine vectors for each of the 179 newsgroups.

6.3 Population Diversity
Recall that population diversity is a measure of the cloud within sequence-space covered by a population of some species. A larger value would correspond to a population which is searching a larger evolutionary space.

We measured the population diversity for all 3,022 quasi-species from our corpus of posts. However we did not measure the within sequence-space diversity and instead have concentrated on the between sequence-space diversity. The within diversity, recall, is sensitive to topical variance. But by definition, a quasi-species has low topical variance and therefore the within diversity for each cluster will necessarily be small. The between sequence-space diversity for a population is a measure of both the variance in post lengths for that quasi-species and a measure of the term-subspace weight variance, which is sensitive to topical specificity.
It is interesting to consider how population diversity varies with population size. That is, do the most highly populated (and thus fit) quasi-species demonstrate relatively high between sequence-space diversity? Figure 6.6 displays the cluster size versus the between diversity on a log-log plot. First we note that between population diversity varies significantly (nearly eight orders of magnitude) across the various quasi-species. Surprisingly, as population size increase, diversity seems to stabilize at a point midway between the total range of diversity values.

**Figure 6.6:** Between population diversity. As quasi-species population size increases, the diversity stabilizes.

We again would like to find out how much of this phenomena is due to the variance in post lengths. Figure 6.7 displays the between population diversity using the normalized direction cosine representation versus the quasi-species population size on a semilog plot. First off, we notice again the dramatic reduction in variance of diversity for the normalized dataset. We again can conclude that post length variance has a large role in overall diversity. In the normalized case, the diversity does not stabilize in the midpoint as was the case with the un-normalized data. Instead we see a clear reduction in diversity as the population
size increases. This may actually be counterintuitive. We might expect that as the number of organisms increases for some quasi-species, the diversity would also increase as the population took on a larger range of topics. However, the reverse seems to be the case; as the number of posts on some topic increases, the topical diversity across those posts actually decreases. One possible explanation is the “me too” character to some of these large clusters. As the population size increase, many of the posts seem to become *more* topically similar to their compatriots due to extensive quoting of previous messages and general restatements.

**Figure 6.7:** Normalized between population diversity. As quasi-species population size increases, the diversity decreases.
Chapter 7

Visualization Methods

As buds give rise by growth to fresh buds, and these, if vigorous, branch out and overtop on all sides many a feeble branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications.

Charles Darwin, 1859

7.1 Newsgroup Tree

We have built systems to visualize aspects of our corporal ecology. The first visualization tool is perhaps the most obvious. In it we image the tree defined by a collection of Net-News newsgroups. For example, in Figure 7.1 we see the tree made up of the 179 newsgroups within the “sci” family. The visualization software then reads post summary files generated by the text analysis system described in the third chapter. From these files the software can determine when a document is posted to any of the newsgroups being displayed and “flash” those nodes of the tree which correspond to the appropriate newsgroups. This visualization tool is effective for viewing global spatial patterns in the distribution of posts across newsgroups. For instance, a particularly active newsgroup or
concentration of newsgroups at some instance in time will flash dramatically.

Figure 7.1: Visualization system displays the entire tree of the “sci” family of newsgroups. When a post arrives for a particular group the display flashes.
If the newsgroup display in Figure 7.1 is flipped ninety degrees we arrive at the image displayed at the top of Figure 7.2. Note that we have replaced the text description of each newsgroup with a simple rectangle. In this visualization we again flash the newsgroup sites when a post arrives for that group. Below the tree image we track post volume for four specific clusters or quasi-species; each hashed circle represents one such cluster. While the user can select any of the clusters found in the corpus, the most populated and active clusters are generally the ones most interesting to track.

Figure 7.2: Visualization system display entire tree of the “sci” family as well as post volumes for four active quasi-species clusters.

The horizontal squiggly lines trace the relative population of each cluster within a given window of time. Consider a point midway along one of the horizontal lines. If we consider
that point time $t$, then a window is defined as being from time $t-n$ to time $t$ where $n$ is some preset constant. The number of posts to each cluster is computed within that window of time and the clusters are then ranked based on this post volume. This windowing method is designed to smooth out the display, it integrates the ranking of cluster volume over a window. It is this ranking that is displayed on the graph. The top of the graph area for each cluster corresponds to the largest cluster population over all windows and the bottom of the graph corresponds to a population size of zero. Time moves from left to right.

This visualization both allows you to see global spatial patterns of posts as well as track the volume of traffic for specific post clusters. When a post occurs for one of the particular clusters which is being tracked the corresponding newsgroup rectangle is flashed with a distinguishing color identifying the particular cluster. In this way, popular and active clusters can be tracked both in volume (with the lower portion of the display) and in their spatial properties (with the color-coded flashes above). To better narrow ones focus on a particular cluster, another mode of operation causes the system to only flash posts which are members of the particular quasi-species cluster of interest. In this way the spatial patterning for a particular cluster can be closely studied over time.

7.2 Phylogenetic Display
Systematics is the discipline concerned with classifying organisms according to their natural relationships. A taxonomy is one such classification which often describes a tree-like hierarchical relationships between a collection of organisms or species. These tree topologies are designed to group together organisms based on their relationships of structure, genotype, lineage, and so forth.

One useful way to visualize the relationships between the quasi-species of a corporal ecology is to fit them into such a taxonomy -- that is, to classify them based on their
mutual relationships. In this case we are relating the quasi-species based on the distances between them in the sequence-space defined by the memotype. Those quasi-species which are near in sequence-space will be close under our classification scheme.

To compute the taxonomy we have employed software developed by Joseph Felsenstein called PHYLIP (Phylogeny Inference Package) (Felsenstein 1993). The particular software package we have used fits a collection of quasi-species wild types into a tree using the Fitch-Margoliash least squares method (Fitch and Margoliash 1967). This method requires that we first compute the pairwise distance matrix for the post clusters. Given a corpus of posts already grouped into quasi-species we start by computing for each quasi-species its wild type or centroid vector. It is these wild types which we will use to represent the clusters in the distance matrix. Next, for a set of \( n \) wild types, we compute the \( (n \times n) \) pairwise distance matrix. To compute these distances we are not using the familiar cosine distance measure but instead make use of the inverse cosine distance metric \((acos)\). This is because the Fitch-Margoliash software requires that the distance between an organism and itself be exactly zero \((acos(ff) = 0)\). Note that because the inverse cosine metric is commutative the distance matrix will have identical upper and lower triangles.

After having computed the pairwise distance matrix we are ready to call on the PHYLIP software to fit this data to a tree. The software tries to find the tree which minimizes the least squared difference between the computed and observed pairwise distances. Let \( D_{ij} \) be the computed distance between the wild type of clusters \( i \) and \( j \); in other words, \( D_{ij} = acos(i,j) \), which is the value of element \((i,j)\) in the distance matrix. Then the phylogeny software attempts to find a tree with a label for each edge such that the equation,

\[
\text{squared error}_{i,j} = \sum_{i=1}^{n} \sum_{j=i}^{n} \frac{(D_{ij} - d_{i,j})^2}{D_{i,j}^2},
\]

(7.1)
is minimized. Here the observed distance, $d_{i,j}$, is taken as the sum of the distance labels on the edges which make up the shortest path between $i$ and $j$ (Figure 7.3). The job, then, is to come up with both a tree and a set of edge labellings such that the squared error is minimized.

$$d_{i,j} = (0.7 + 1.2 + 0.6) = 2.5$$

**Figure 7.3:** The observed distance, $d_{i,j}$, is the sum of the labelled edges which compose the shortest path between node $i$ and $j$.

In Figure 7.4 we display one such phylogeny. The distances on each edge are not displayed and each leaf node represents a particular wild type. The English label we have assigned to each leaf node was hand generated and is designed to evoke the topic discussed by the particular cluster. All of the posts used here were to the “sci.stat” family of statistics newsgroups. Note that this phylogeny represents only 51 clusters generated from just 165 posts. This relatively small dataset is necessary in particular because the tree fit-
ting process is computationaly very expensive, $O(n^4)$. We tried larger corpora but were not able to find a tree after running the software for multiple days.

Figure 7.4: Taxonomy of 51 clusters. The labels were hand generated and hope to evoke the main topic of each cluster. Posts were to the “sci.stat” family.
Chapter 8

Conclusions and Future Work

Meme evolution is not just analogous to biological or genic evolution.... It is not just a process that can be metaphorically described in these evolutionary idioms, but a phenomenon that obeys the laws of natural selection quite exactly. The theory of evolution by natural selection is neutral... regarding the differences between memes and genes; these are just different kinds of replicators evolving in different media at different rates.

Daniel Dennett, 1995 (on Dawkins)

8.1 Conclusions

We have introduced the concept of the corporal ecology and have applied these theories to the analysis of posts to USENET News. In our ecology posts are organisms, the newsgroups of NetNews define an environment, and human posters situated in their wider context make up a scarce resource and describe a dynamic fitness landscape. Within the corporal ecology we have developed theories of population memetics. We described eigen-text, a method of statistical text analysis, and argued that this technique distills self-reproducing units of text from a corpus. Eigen-text arrives at suitable replicators because it discovers the word co-occurrences that segregate and re-combine with appreciable frequency. We argued as well that natural selection is necessarily engaging eigen-text because all of Eigen’s requirements for its occurrence obtain: self-reproduction, mutagenicity, and metabolism. We described a set of experiments performed on a static corpus of over 10,000 posts. In our experiments we investigated two fundamental elements of population ecology: fitness and diversity. Our study of fitness led to the unhappy discovery that a flame-war and an overly prolific poster is the king of the jungle. In investigating diversity we made the surprising conclusion that as the number of posts on a particular subject increases, the diversity between those posts actually decreases.
The statistical text analysis method we use, eigen-text, is closely related to latent semantic indexing, the text retrieval method due to Dumais and colleagues. We described at some length, and without initially making connections to a corporal ecology, this system and our departures from it. These methods have many practical applications and we will consider some of those possibilities in the next section.

Finally, we ended with a description of three visualization tools for our system. Two of them allow the user to examine the spatial patterns and activity levels of clusters of posts across the landscape defined by NetNews. The third system builds and displays a taxonomy of posts to NetNews describing a tree-like set of relationships between post clusters.

### 8.2 Future Work

As is often the case our work poses more questions then it answers, and there exists a wealth of potential future work.

One large arena for future work is to collect empirical evidence of natural selection within our ecology. The purpose of such work is both to (hopefully) build evidence of the strength of our theories as well as to develop a greater understanding of the qualities of discourse over NetNews. Though many of the most interesting questions are related to dynamic activity, we have so far only studied static properties of a preset corpus of posts. To effectively study post dynamics will require, in our estimation, a much more scalable set of analysis tools. The software system will need to handle one or two orders-of-magnitude more posts then the current system before we will be able to effectively study true dynamics. Moreover, the system will have to be installed online and made to function in realtime. Happily, all of the methods used to date (including SVD) can account for the online “folding-in” of new posts; so scaleablility is a small matter of programming. Some of the dynamic properties we would like to investigate include: speciation and extinction
of quasi-species, the evolution of meme complexes themselves, predation, replacement, migration, niche behavior, and colonization.

We would also like to study how our clustering and analysis methods compare to and predict elements of the in-reply-to threading as well as the layout of the environmental mosaic defined by newsgroups. We are concerned with possibilities of circularity in our work. That is, the meme complex as well as quasi-species are both statistically derived phenomena. Some of our analysis may indeed apply further statistical analysis to these elements -- applying statistics to what already are statistics. We have been cautious about avoiding such circularity so far; for instance, much of the diversity results are based on the environmental patches defined by the newsgroups and not on statistical calculations. We would like to further study both the in-reply-to threading and the newsgroup environmental mosaics as means of avoiding any problems with circularity, along with hopefully discovering new and interesting phenomena.

We also would like to explore areas which are not part of the current neo-Darwinian orthodoxy. For instance, we would like to see the ties between the evolution of memes and viral or molecular evolution strengthened and evidence built for Kimura's neutral theory. Further, we are interested in exploring ways in which the environment, development, or dynamics of the system might serve to constrain evolution. In other words, we would like to explore phenomena beyond selection which may control aspects of evolution.

Besides scalability, there are other improvements we would like to make to the current software system. For example, the current clustering algorithm is not completely satisfactory. We would like to explore clustering algorithms that automatically compute the threshold distance -- the limit of how far apart two posts can be and still be grouped in the same cluster. For example, the ISODATA method may be an adequate alternative (Ther-
rien 1989). Further, for reasons argued in section 3.5.1, we would like to use the arc cosine metric rather than the more traditional cosine measure for all distance computations.

While we have argued the strengths of the eigen-text method, we would still like to see work investigating other possible means of distilling replicators. SVD is but one possible method of decomposition and perhaps there exists others that would be more effective, for instance wavelet decomposition. We suspect that a statistical approach at replicator discovery, such as eigen-text, is destined to be the most effective. However, we would still like to see if higher level text analysis or understanding systems (semantic or pragmatic natural language systems) can discover replicators with success.

We have applied our theories of a corporal ecology to posts to NetNews. We would like to study other corpora as well. Areas of interest include email, chain letters, manuscript traditions, and the Web. The Web defines a relatively static set of information insofar as an arrow of time does not run through the collection of Web pages. Still, we believe the Web is an interesting area to study though our model may require adjustments to work in this domain.

As previously mentioned, we believe there are a number of practical applications for the eigen-text method, in particular when considering information on the Net. For instance, there currently exists a NetNews reader called “trn” which is designed to allow users to read posts based on their in-reply-to threading. Thus, one can read an entire thread at a time rather than then receiving the posts in the strict order of their arrival. We would like to build a NetNews reader called “crn”, an eigen-text cluster based news reader. Instead of basing it on threading, this system would present posts in semantically related groups. Eigen-text might also be an effective means to build up user profiles for information filtering and personalization. Since the term-subspace can be thought of as an automatically generated set of keywords, user modeling might be based upon discovering collections of
term-subspaces of interest to a user. Eigen-text, then, might helpful in information filtering, finding important or relevant documents, and personalizing information on the Net.
References


Frakes, W.B. and R. Baeza-Yates, eds. 1992. *Information Retrieval: Data Structures and


