Best, M., L., 1997; Models for Interacting Populations of Memes: Competition and Niche Behavior.
Journal of Memetics - Evolutionary Models of Information Transmission, **1**.
http://cfpm.org/jom-emit/1997/vol1/best_ml.html

Models for Interacting Populations of Memes: Competition and Niche Behavior

Michael L. Best
Media Laboratory,
Massachusetts Institute of Technology,
Cambridge, Massachusetts 02139, USA.
mikeb@media.mit.edu

The original version of this paper was published as Best, M.L. (1997). Models for Interacting Populations of Memes: Competition and Niche Behavior. In Husbands, P. and Harvey, I. (eds.) (1997). *Fourth European Conference on Artificial Life*, Cambridge, MA: MIT Press.
MIT Press have kindly given us permission to re-publish this new version.

Abstract

We make use of a set of text analysis tools, primarily based on Latent Semantic Indexing, to study the dynamics of memes on the Net. Our analysis discovers replicating memes within posts to the USENET News (or NetNews) system. We cluster the posts to NetNews into clouds within a conceptual sequence space; these clusters describe quasi-species. We then go on to study the pairwise interactions between these quasi-species by computing the cross-correlation between the interacting population's level of post activity. We analyze a particular corpus of posts to the soc.women newsgroup and argue that strong negative cross-correlations are examples of competition between the quasi-species. We find that high levels of competition occur more frequently among quasi-species who exist within a narrow ecological niche. We note that this phenomena also occurs in natural ecologies.

1 Introduction

Ideas do not exist in a vacuum. Neither does discourse, the interconnected ideas which make up conversation and texts. In this research we investigate the pairwise interaction between populations of ideas within discourse: Are our text populations in competition with each other? Do they mutually benefit each other? Do they prey on one another?

This work attempts to build models of *population memetics* by bringing together two disciplines: Alife and text analysis. Through techniques of text analysis we determine the salient co-occurring word sets, texts, and text clusters, and track their temporal dynamics. We then study the life-like properties of this human-made system by considering its behavior in terms of replicators, organisms, and species.

Richard Dawkins coined the term *meme* to describe replicating conceptual units (Dawkins [7]). In studying the population dynamics of ideas we consider the meme to be the largest reliably replicating unit within our text corpus (Pocklington & Best [27], Pocklington [26]). Through text analysis we identify memes within a corpus and cluster together those texts which make use of a common set of memes. These clusters describe species-like relationships among the texts.

The particular texts we study are posts to the popular USENET News (or NetNews) system. These posts form the basis of a new Alife environment, the *corporal ecology* (Best [4, 5]). In this ecology texts are the organisms, the digital system defined by NetNews describes an environment, and human authors operating within some culturally defined parameters are the scarce resource.

At the core of our study sits a large text analysis software system based primarily on Latent Semantic Indexing (LSI) (Furnas, et.al [16], Deerwester, et.al. [9]; Dumais [10, 11]). This system reads each post and computes the frequency with which each word appears. These word counts are then used in computing a vector representation for each text. A principal component analysis is performed on this collection of vectors to discover re-occurring word sets; these are our memes. Each post is then re-represented in terms of these memes. By grouping texts which are close to one another within this meme-space we cluster semantically similar texts into species-like categories or *quasi-species* (Eigen et.al. [12]).

We proceed to study the interactions between those populations which coincide temporally. For each cluster we compute a series which represents its volume of post activity over time, for instance how many texts of a given cluster were posted on a given day. Cross-correlations between each pair of time series are then determined. We find that some pairs have strong negative correlations and argue that these are examples of texts in competition. A number of examples of such competition are explored in depth. We argue that high competition is correlated with those text clusters which exist within a narrow ecological niche; this phenomena is also observed in natural ecologies (Pianka [24]).

Note that this is an unusual shift from the typical Alife environment. We are not synthesizing replicators, embodying them into agents, and observing their life-like interactions. Instead we are studying a pre-existing artifact. Through our analysis we *discover* replicators within organisms, and use computational techniques to observe their dynamics.

In this paper we first briefly overview the NetNews environment and describe the LSI based text analysis system. Next we describe the mechanism used to determine the temporal dynamics and cross-correlations given a corpus of posts. We then relate the cross-correlations to models of interacting populations. In the next section we examine in depth a couple pairs of post clusters with strong interactions. We then describe a theory of niches within the corporal ecology and note that narrow ecological niches are correlated with significant competition. We end with our conclusions.

2 The NetNews Corpus

Understanding our corpus requires a basic knowledge of the NetNews system. NetNews is an electronic discussion system developed for and supported on the Internet (Kantor & Lapsley [18]). Discussion groups have formed along subjects ranging from science to politics to literature to various hobbies. The collections of messages are organized into particular subject groups called *newsgroups*. The newsgroups themselves are organized in a tree-like hierarchy which has general top-level categories at the root and moves to more specific topics as you progress towards the leaves. 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, ``soc.religion'' is the name of a newsgroup concerned with social issues around the world's religions and ``soc.religion.hindu'' is a more specific group devoted to Hinduism.

Texts sent to NetNews, the *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 attach a number of additional fields to posted messages including a timestamp and the user name of the person who created the post.

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.

NetNews today has grown considerably from its beginnings in the late 70's and 80's. With over 80,000 posts arriving each day, it provides an excellent dataset for the study of cultural microevolution.

```
From: mikeb@media.mit.edu
Newsgroups: soc.religion.hindu
Subject: Angkor Wat
Date: 26 Jan 1997 02:17:05 -0700

Is Angkor Wat, the magnificent temple complex in the jungles of
Cambodia, considered a Buddhist or Hindu shrine?
```

Figure 1. A fictitious example post sent to the soc.religion.hindu newsgroup along with some of its header information.

3 The Text Analysis Method

We analyze a corpus of posts to NetNews to distill their salient replicating unit or memes, and to cluster together posts which make common use of those memes. We do this by employing a large system of text analysis software we have built. The techniques employed are based on the vector space model of text retrieval and Latent Semantic Indexing (LSI).

3.1 Vector Space Representation

We begin with a corpus composed of the full-text of a group of posts. We analyze the corpus and identify a high-dimensioned space which describes the conceptual elements within the texts. For each post we identify a point within this space which captures it semantically. This technique is known as a vector space representation (Salton & Buckley [28]; Frakes & Baeza-Yates [15]). Each dimension in this space will represent a *term* from the corpus where a term is a word that occurs with some frequency (e.g. in at least three posts) but not too frequently (e.g. the word ``not'' is dropped from the term list). The goal is to arrive at a set of terms which semantically capture the texts within the corpus.

Given the conceptual space described by this set of terms each post can be represented as a point within this space. We score each document according to the frequency each term occurs within its text, and assign each term/document pairing this *term weight*. The weighting we use for each term/document pair is a function of the *term frequency* (simply the number of times the term occurs in the post) and the *inverse document frequency* (IDF). Consider a corpus of m posts and a particular term, *j*, within a list of n terms. Then the IDF is given by,

$$IDF_j = \log(\mathrm{floor}(m - m_j) / m_j)$$

where m_j is the number of posts across the entire corpus in which term *j* appears. Thus, if a term occurs in 50% or more of the texts the IDF for that term will vanish to zero. But if, for instance, a term occurs in 10% of the documents the IDF will be log(10). In words, rare terms have a large IDF.

The term weight for a document, *i,* and term, *j*, is then defined by,

$$TermWeight_{ij} = w_{ij} = \log(TermFrequency_{ij}) \, . \, IDF_j.$$

Each term weight, then, is a function of the inter- and intra-document term frequencies.

Each post, *i,* is now represented by a particular term vector, $r_i = (w_{i1}, w_{i2}, ..., w_{in})$. The entire collection of *m* term vectors, one for each post, define the *term/document matrix*, *A*.

This set of steps, culminating in the term/document matrix, form the basis for much of modern text retrieval or filtering and are at the core of most Web search engines.

3.2 Latent Semantic Indexing

LSI is a technique used to distill high-order structures from a term/document matrix, consisting of sets of terms which re-occur together through the corpus with appreciable frequency. The re-occurring term sets are discovered through a principal component method called Singular Value Decomposition (SVD). While LSI was primarily developed to improve text retrieval, we are interested in its ability to find replicating term sets which act as memes. We will first overview the LSI technique and then discuss how it discovers memes.

LSI was originally proposed and has been extensively studied by Susan Dumais of Bell Communications Research and her colleagues (Furnas, et.al. [16]; Deerwester, et.al. [9]; Dumais [10, 11]). Peter Foltz investigated the use of LSI in clustering NetNews articles for information filtering (Foltz [14]). Michael Berry and co-authors researched a variety of numerical approaches to efficiently perform SVD on large sparse matrices such as those found in text retrieval (Berry [1]; Berry, et.al. [2]; Berry & Fierro [3]).

The SVD technique decomposes the term/document matrix into a left and right orthonormal matrix of eigenvectors and a diagonal matrix of eigenvalues. The decomposition is formalized as, $A_k = U V^T$.

The term/document matrix, A, is approximated by a rank-k decomposition, A_k; in fact the SVD technique is known to produce the *best* rank-k approximation to a low-rank matrix (Berry [1]).

We are interested in only the right orthonormal matrix of eigenvectors, V^T. Each row of this matrix defines a set of terms whose co-occurrences have some statistically salient re-occurrences throughout the corpus. That is, each eigenvector describes a subspace of the term vector space for which the terms are frequently found together. These *term-subspaces* describe a set of semantically significant associative patterns in the words of the underlying corpus of documents; we can think of each subspace as a *conceptual index* into the corpus (Furnas et.al. [16]).

For instance, an example term-subspace generated by analyzing a collection of military posts found three words as having significant re-occurrences, and therefore replicating together with success: ``harbor'', ``japan'', and ``pearl''. These term-subspaces make up our replicators and are our putative memes. Memes are not single re-occurring words but are made up of *sets* of re-occurring words.

Our final text analysis step is to ``compress'' the original term/document matrix by multiplying it with this right orthonormal matrix of eigenvectors (in other words we perform a projection). This, in effect, produces a *term-subspace/document* matrix. Each post is represented by a collection of weights where each weight now describes the degree to which a term-subspace is expressed within its post's text.

4 Meme and Quasi-species

4.1 Term-subspace as Putative Meme

We are looking for replicators within the corpus which are subject to natural selection. Elsewhere we have argued at length as to why the term-subspace captures the requirements of a true meme because its word sets act as a unit of selection within the corpus (Best [4, 5]; Pocklington & Best [27]). The strengths of this term-set as a replicating unit of selection are due to it meeting the following conditions:

- it is subject to replication by copying,
- it has strong copying fidelity,
- but not perfect fidelity, it is subject to mutation,
- it has a strong covariance with replicative success

(Lewontin [19]; Eigen [13]).

We will quickly review each of these points in turn.

SVD techniques exploit structure within the term/document matrix by locating co-occurring sets of terms. Clearly these term sets are replicating through the corpus since that is the precise statistical phenomena the SVD analysis detects. However it is not obvious that this replication is generally due to copying. Instances of precise copying occur when an in-reply-to thread includes elements of a previous post's text via the copying mechanism provided by the software system. Other instances of copying occur within a particular context or discussion thread when authors copy by hand words or phrases from previous posts into their new texts. More abstractly, replication occurs because certain memes are traveling outside of the NetNews environment (and thus outside of our means of analysis) and authors again act as copying agents injecting them into the corporal ecology. But, clearly, some re-occurrences are not due to copying but are a chance process where unrelated texts bring together similar words. The likelihood of such chance re-occurrences will be a function of the size and quality of our replicating unit. In summary, term-subspaces are instances of replication often due to copying.

The copying fidelity of a term-subspace is also a direct outcome of the SVD statistical analysis. But importantly, the copying fidelity of re-occurring term sets is not perfect across the entire corpus; the term sets will co-occur with some variation. These mutations are both changes *designed* by human authors and chance variation due to copying errors. In either case the mutations are *random* from the vantage of selection; in other words, human authors are not able to perfectly predict the adaptive significance of their inputted variations. These mutations work ``backwards'' into the actual term-subspace representation for a post organism. That is, a random mutation at the post level will actually result in a random mutation in the vector subspace representation (the memotype) for the post organism. In this way, the memes as represented in the memotype are subject to mutation.

Finally, we have elsewhere shown there can be a strong covariance between the replicative success of a cluster or thread of posts and the degree to which they express certain term-subspaces (Pocklington & Best [27]). In other words, a group of posts can increase its volume of activity over time by increasing the degree to which it expresses certain term sets within its post's text. This, then, is a covariance between the *fitness* of a population of posts and the expression of a particular *trait* as defined by a term-subspace. The demonstration of this covariance is critical to establishing that a replicator is subject to natural selection.

4.2 Quasi-species

If the term-subspace is a reasonable model for the meme then the term-subspace vector representation of a post is a good model of the post's *memotype*. Much as a genotype describes a point within genetic sequence-space for each organism, the memotype describes a point within conceptual sequence-space. By *sequence-space* we mean any of the search spaces defined by a replicator undergoing selection. Examples of sequence-spaces include the gene space, protein spaces under molecular evolution, and the meme space defined within a corporal ecology.

The notion of a *quasi-species* is due primarily to Manfred Eigen (Eigen, et.al. [12]; Eigen [13]). 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 [13]). One organism is a *mutant* of another if it is particularly close to the other in sequence-space.

We wish to group our posts into quasi-species. This requires finding groups of memotypes which are centered together within the conceptual sequence space. To do so we employ a simple clustering algorithm, the Nearest Neighbor Algorithm (Jain & Dubes [17]). We first normalize each post memotype to unit length; this amounts to discarding text length information and representing only the *relative* strength of each meme within a text. The clustering algorithm then considers each post memotype in turn. The current memotype is compared to each memotype which has already been assigned to a cluster. If the closest of such vectors is not farther than a 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.

This process assigns each post to a quasi-species defined as those posts which are close to one another in conceptual sequence-space.

The overall aim in grouping organisms is to bring to light certain evolutionarily significant relationships. Clearly, our quasi-species clustering method is ahistorical; that is, it does not directly account for decent when grouping together text organisms. The extent to which such groupings are effective when studying the relatedness of natural organisms is a matter of continued controversy as can be seen in the debates of the cladists versus evolutionary systematists versus pheneticists. While we are currently agnostic to this controversy we do agree with an original claim of the pheneticists: the more traits used when assessing the relatedness of individuals the more accurate are the groupings (Mettler, Gregg & Schaffer [23]).

We are in the happy situation of clustering based on the complete memotype for each of our organisms. The result is that under empirical verification our clusters exhibit extremely strong historical relatedness. We have found that the vast majority of texts clustered together come from the same in-reply-to thread and thus are related by decent (Best [5]). But our clustering method has the added benefit of grouping related texts even when the in-reply-to mechanism is not used and, alternatively, breaking up texts that are within the same thread but are not semantically related. This is of value since many posters to NetNews use the in-reply-to mechanism to post unrelated texts or, alternatively, post follow-up texts without bothering to use the in-reply-to facility. Thus, we claim that our clustering mechanism, due to its access to hundreds of traits, is actually superior at grouping together both related and descendent texts then would be a simple reliance on the threading mechanism. The clustering method meets our goal of illuminating evolutionarily significant relationships.

4.3 Comparison to Natural Ecologies

We are describing phenomena within a corpus of texts in terms of population ecology and population genetics. This is not simply a metaphorical device; we believe that interacting populations of texts and their constituent memes are evolving ecologies quite exactly. However there are clearly a number of interesting differences between genes and memes (as here operationally defined), natural organisms and texts, ecologies and corpora. Important differences include the driving forces behind mutation within the texts and the role of self-replication and lineage within the corpora. We leave to future work a more complete analysis of these differences.

---

5 Models for Interacting Populations

We now turn to studying the interaction between quasi-species of posts. We have so far only studied the pairwise interactions between post quasi-species. Similar pairwise interactions have been widely studied within theoretical ecology. Consider two interacting populations: one population can either have a positive effect (+) on another by increasing the other's chance for survival and reproduction, a negative effect (-) by decreasing the other population's survival chances, or a neutral (0) effect. The ecological community has assigned terms to the most prevalent forms of pairwise interaction, in particular:

- Mutualism (+, +)
- Competition (-, -)
- Neutralism (0, 0)
- Predator/prey (+, -)

(Pielou [25]; May [22]).

Our goal is to study the pairwise interactions of quasi-species within the corporal ecology with the hope of discovering some of these interaction types.

5.1 Time Series

To study how the interactions of populations affects growth rates we must define a method to measure a quasi-species' growth over time. Recall that a quasi-species describes a collection of posts which are close to one another in sequence-space. Each of these posts has associated with it a *timestamp* identifying when that text was posted to the system; in effect, its birth time and date. (Note that a post organism has something of a zero-length life-span; it comes into existence when posted but has no clear time of death.)

A histogram of the timestamp data is created with a 24 hour bucket size. That is, for each quasi-species we count how many member texts were posted on one day, how many on the next, and so forth through the entire population of texts. The datasets currently used span on the order of two weeks and consist of thousands of posts. So for each day a quasi-species has a volume of activity which can range from zero to 10's of posts. This rather course unit, the day, has been chosen to neutralize the strong daily patterns of post activities (e.g. activity may concentrate in the afternoons and drop off late at night, different timezones will shift this behavior and thus encode geographic biases). Thus the patterns of rise and fall in the volume of posts within a quasi-species when measured at the day level will, hopefully, reflect true changes in interest level and authorship activity rather then other external or systemic factors

5.2 The Test Corpus

Figure 2 is a typical graph for the volume of posts within a particular quasi-species over a period of ten days. This cluster was found within a corpus of all posts sent to the soc.women newsgroup between January 8, 1997 (the far left of the graph) and January 28, 1997 (the far right). In the figure the number of posts in a day is represented by the height of the graph. This particular cluster of texts exhibited an initial set of posts, a few days worth of silence, then a rapid building up of activity which then declined precipitously at the end of the dataset. The entire corpus used consisted of 1,793 posts over the same ten day period. The clustering mechanism arrived at 292 quasi-species the largest of which contained 103 posts.

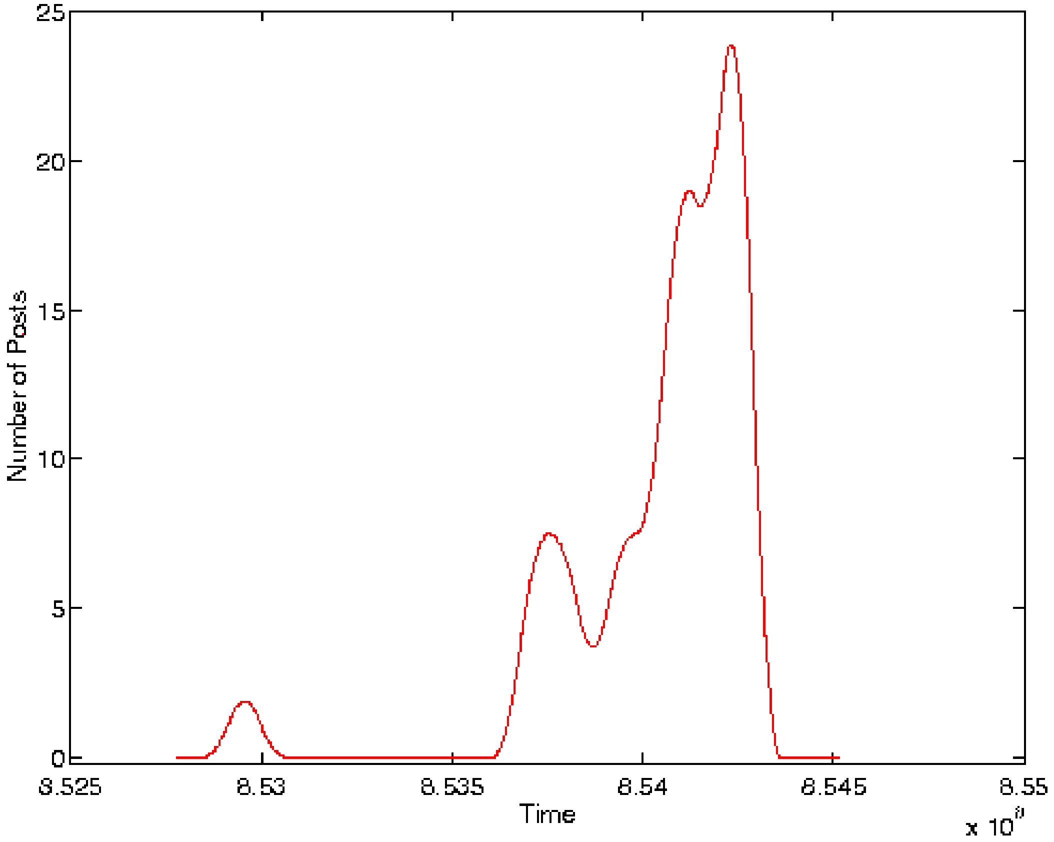

Figure 2. Typical time series of posts to quasi-species. Time axis is measured in seconds since Jan. 1 1970.

5.3 Time Series Cross-correlation

To study the relationship between the time series of two populations of posts we use the cross-correlation function. The use of the cross-correlation to study bivariate processes, and time series in particular, is well known (Chatfield [6]). Each time series is normalized to be of zero mean and unit standard deviation; that is, we subtract off the mean and divide by the standard deviation. In this way, the cross-correlations will not be dominated by the absolute volume of post activity within some cluster and instead will be sensitive to both large and small sized clusters.

We assume the readers are familiar with the regular covariance and correlation functions. Then the cross-correlation for two time series, *X* and *Y*, is given by,

$$\square_{xy} = \square_{xy} / \text{sqrt}(\square_{xx}\square_{yy})$$

Here $\square_{xy}$ = Cov(*X*, *Y*) and $\square_{xx}$ and $\square_{yy}$ are the variance of *X* and *Y* respectively. Note this formulation only considers the cross-correlation for a zero time lag. That is, it considers how the two time series are correlated for identically matching points in time. With a nonzero lag the cross-correlation would study cases when the two series might have correlations offset by some fixed amount of time. Since we group our time data into day-long chunks the zero-lag cross-correlation will be sensitive to covariances which have a time offset as large as 24 hours; this builds into the time series an adequate time lag.

When the cross-correlation between two sets of data is significantly different than zero it suggests the two sets of data have some relationship between them. A positive value means an increase in one series is likely to co-occur with an increase in the other series. A negative value means an increase in one series is likely to co-occur with a decrease in the other series.

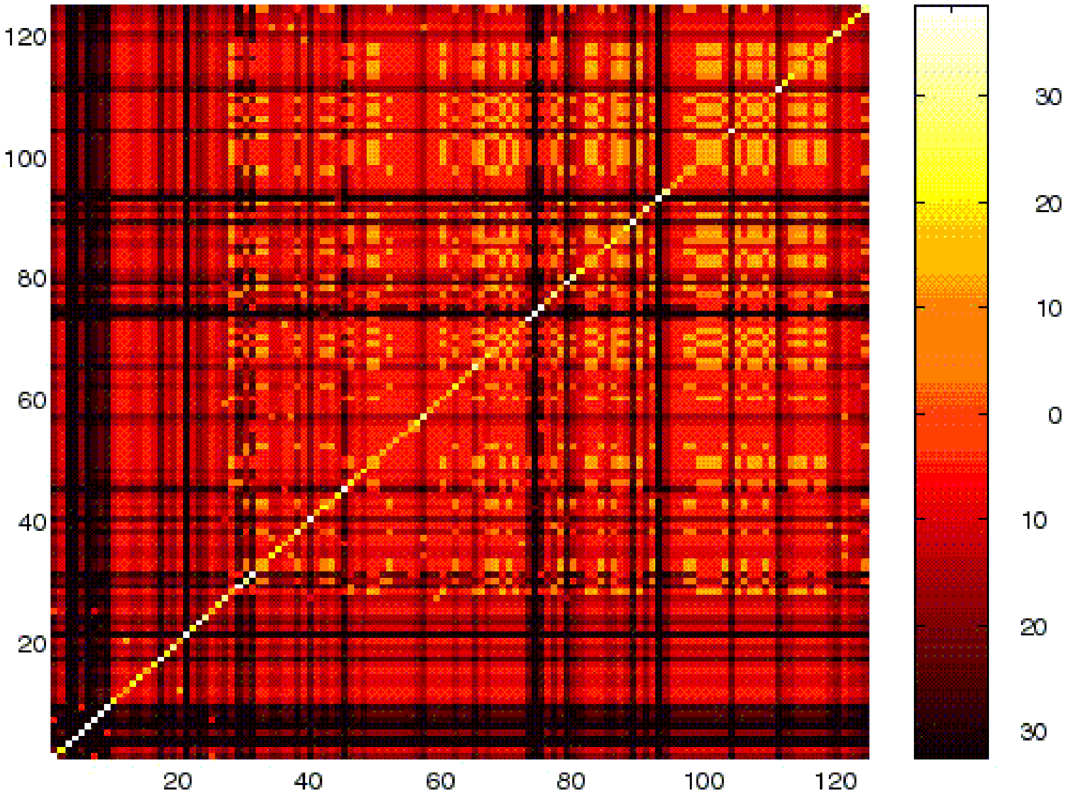

Figure 3. The pairwise time series cross-correlation for 125 largest quasi-species clusters.

Figure 4 shows the pairwise cross-correlations for the 125 largest quasi-species clusters within our corpus. The diagonal represents the cross-correlation between a time series and itself which, as expected, is identically one. Note that the matrix is symmetric about the diagonal. The off-diagonal values range from near one to -0.26. The mean cross-correlation is 0.3. This value is quite high, indicating that most of these post clusters are somehow positively related. We suspect this high average cross-correlation is at least partially due to external or systemic affects which were not removed by the day-long bucket size. For instance, our analysis would be sensitive to patterns due to the Monday-Friday work week common in the West. Further, some of this correlation may be due to a high level of mutualistic interactions amongst the posts. Clearly, the ideas conveyed within the soc.women newsgroup often share similar contexts.

In our analysis this overall high correlation does not particularly matter since we are concerned with the *relative* cross-correlation -- that is, those that are the largest and those that are the smallest.

6 Negative Cross-correlations: Competition versus Predator/Prey

We have primarily studied those pairs of quasi-species with relatively strong negative cross-correlations; to wit, those where ▯$_{xy}$ <= -0.2. Note that in all such cases (there are 42) P < .001 suggesting that with extremely high probability the correlations are not due to chance. Figure 4 and Figure 5 plot two such interactions, both fairly characteristic of this population.

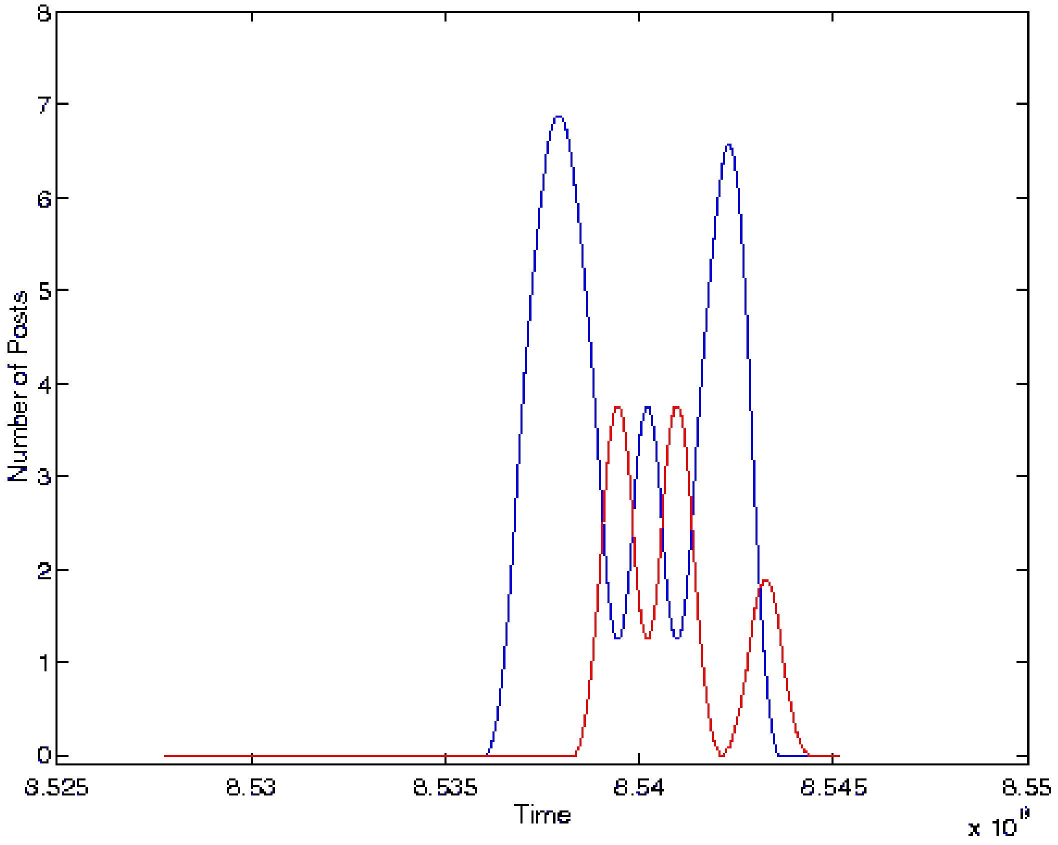

Figure 4. Volume of activity for two quasi-species. The cross-correlation between these two series is -0.26.

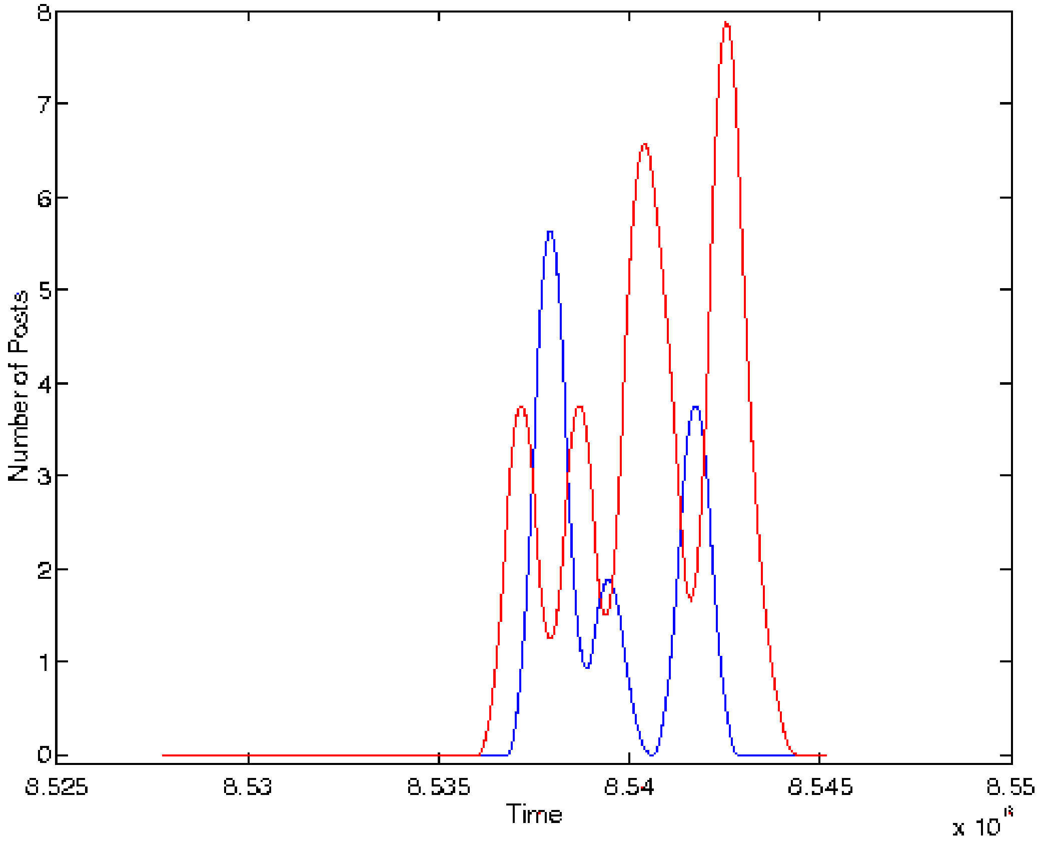

Figure 5. Volume of activity for a different set of two quasi-species. The cross-correlation between these two series is -0.23

Both of these figures demonstrate a clear negative covariance between the volume of activity of the two post clusters. This negative covariance is both statistically significant and visually compelling. But what do these graphs signify and can it be interpreted within the rubric of ecological interactions?

At first glance the interactions appear to be of a predator/prey variety; they have a (+, -) relationship to them. However, competition might also produce similar interaction phenomena if the competitors are operating close to some limitation or environmental carrying capacity. In such instances the relationship between population sizes will be a zero-sum game, when one goes up the other must come down. To be able to classify the interactions of Figures 4 and 5 we need to consider the qualitative details of these two interactions through direct study of the texts.

Recall that in the case of a predator/prey relationship, one population enjoys an increased growth rate at the *expense* of another population (e.g. one population feeds on the other). The presence of a relatively large population of predators will result in a diminished level of success for the prey (they get eaten up). Conversely, the relative absence of prey will result in diminished success for the predator (they have nothing to eat).

Now consider the case of competition. In competition two interacting populations inhibit each other in some way, reducing each other's level of success. This often occurs when the two populations rely on the same limited resource. Unlike the predator/prey relationship where the predator requires the prey for success, with competition the two populations would just as soon avoid each other all together.

This pressure towards avoidance is the source of much ecological diversity since it propels populations to explore new and therefore competition-free niches (Pianka [24]). An ecological *niche*, for some particular species, is simply that collection of resources the species relies on. Interspecific niche overlap occurs when two or more species share one, some, or perhaps all of their resources. When those resources are scarce, interspecific competition will result. The *width* of a niche is simply a qualitative sense of the variety and number of resources a population makes use of.

6.1 Competition and Niche Behavior

We have studied the set of posts which make up the four quasi-species shown in Figures 4 and 5 in an attempt to qualitatively classify their interactions. The quasi-species of Figure 4 are made up of posts within a single thread. The subject line for these posts reads, ``Men's Reproductive Rights''. In general, these posts are concerned with the responsibilities and rights of men towards their unborn children. The quasi-species displayed with a dashed line in the figure is centered around the use of contraceptives. It consists of a collection of posts wherein the authors debate who is most responsible, the women or the man, when using contraception. The quasi-species with a solid line deals instead with the use of abortion and whether the father has any intrinsic rights in deciding whether or not to abort an unborn child.

In Figure 5 these two quasi-species are also from a single thread. The subject line here reads, ``Unequal distribution of wealth?''. This particular thread of discussion was rather large. In fact there was a total of 365 posts to this thread which our text analysis tools broke up into a number of quasi-species due to significant bifurcations of the topic. In other words, many parallel discussions occurred all within a single in-reply-to thread. The cluster of discussion shown with the solid line in Figure 5 centered around a debate as to whether the US military was a ``socialist collective''. The quasi-species with the dashed line was a debate on the value of releasing the mentally ill from hospitals. Clearly, these two debates are quite dissimilar even though they span the same set of days and are posts to the same discussion thread.

The quasi-species of Figure 4 are different but related discussions. Those of Figure 5 are different and not clearly related. Still, we believe that both of these sets of interactions demonstrate elements of competition. Within the texts there is no evidence of predator memes; in Figure 5, in fact, the memes seem entirely orthogonal to one another. However, in both examples the memes are competing for the same collection of human authors who must act as their agents if they are to propagate and succeed. This seems even more likely when we consider that all these posts are to the same newsgroup which due to its narrow subject area supports only a limited supply of human posters. Moreover, each pair of interactions are confined to a single thread of discussion, which again has an even more limited set of potential human authors since users of the NetNews system often zero-in on particular threads they find interesting and ignore others. After inspecting most of the interactions which demonstrated strong negative correlations we observed no examples of predator/prey interactions but many instances which appeared to be examples of competition.

6.1.1 Statistical Artifacts

We computed the cross-correlation between 125 different clusters, arriving at 15,625 different correlations. It is possible, therefore, that the cross-correlations with large negative values exist simply by chance; they represent the tail of the distribution of correlations.

However, we believe that our qualitative analysis provides strong evidence that these negative correlations are *not* artifacts but are indeed due to an interaction phenomena between the two quasi-species. The two pairs of quasi-species described in detail above demonstrate this point. The likelihood that two quasi-species be brought together by mere chance *and* both be from the same thread (out of 324 threads within the corpus) seems vanishingly small.

6.1.2 Competition

We now will test our theory that these interactions are of a competitive nature. Again recall that competition is often caused by populations existing within the same (narrow) ecological niche. What makes up an ecological niche for a meme within NetNews? We argue that the newsgroups themselves make up spatially distributed ecological niches. Since there is relatively little interaction between newsgroups (save the phenomena of cross-posting) we would expect these niches to behave something like island ecologies -- they remain relatively isolated from each other. Within a single newsgroup (which is all we have studied so far) niches might be described by threads of discussions. As previously stated, we have found that individual posters to the system tend to become involved in particular in-reply-to threads which interest them. Thus the memes within a particular thread make use of a set of human resources which is smaller then the entire set of potential human resources available to the newsgroup. These resources define the niche.

We theorize that cross-correlations which approach -1 in our corpus are examples of competition, and competition will be more likely between populations which are posted to the same threads and thus have overlapping niches. The most direct way to test this theory is to see if negative cross-correlations between two quasi-species correlates with the degree to which they post to the same threads. For each of the 125x125 pairwise interactions we computed the number of threads each of the quasi-species pairs had in common and divided that by the total number of threads posted to by each quasi-species. For example, one quasi-species may contain posts which went to two different in-reply-to threads. Another quasi-species may have posts which span three different threads one of which is identical to a thread within the first group. So this pair of quasi-species would have posted to a total of four different groups one of which was shared. Their relative niche overlap would therefore be 0.25.

We calculated the correlation coefficient between the negative cross-correlations of Figure 3 and the percentage of thread overlap between these quasi-species pairs. We found this correlation to be -0.04. While this correlation is statistically significant (P < .001), it is not very pronounced. The negative sign, though, does indicate that as the level of competition increases (a negative cross-correlation) the percent of overlap of their niche also increases (a larger positive shared thread percentage).

This small correlation coefficient may be due to a small signal/noise ratio. Since most pairwise interactions result in small correlations, the relative number of large negative correlations is quite small. The number of interactions grows with the square of the number of quasi-species. We suspect that a simpler experiment which grows linearly with the number of quasi-species will have a better signal/noise ratio.

We have studied the correlations between the absolute number of in-reply-to threads a quasi-species is posted to and the average degree to which the quasi-species finds itself correlated with other clusters. Our hypothesis is that the absolute number of threads a quasi-species is posted to will be related to the average degree of competition the quasi-species experiences in its interactions. Since the variety of resources used by an entity defines it niche, if a quasi-species is posted to a relatively small number of threads then it exists in a narrow ecological niche. Should there subsequently be any interspecific overlap of these narrow niches, scarcity will result in competitive encounters. We computed the correlation coefficient between the total number of threads within a quasi-species and its *average* cross-correlation value. The correlation coefficient here is 0.25. Thus, as the number of threads within a quasi-species increases (the set of available resources is widened) the average level of competition diminishes (the mean pairwise cross-correlation also increases). This correlation is statistically significant (P < .001) and rather pronounced.

We further computed the correlation coefficient when the absolute number of threads was normalized by the size of the quasi-species. We might expect that the number of threads employed by a quasi-species would grow with the number of posts within that quasi-species. In other words, as a quasi-species gets larger the number of threads increases too. This might affect the analysis above such that instead of measuring niche width we were simply measuring quasi-species size. Dividing out the size amounts to computing the *average* number of threads employed by a post for a given quasi-species. When this set of values was correlated with the mean cross-correlation we arrived at a nearly identical coefficient as above and again clear statistical significance. Thus quasi-species size is not a major factor in level of competition.

7 Conclusions

We have described a set of text analysis tools, based primarily on Latent Semantic Indexing, which distill replicating memes from a corpus of text. We have trained this analysis system on a corpus of posts to NetNews. This makes up a corporal ecology where the posts are organisms, NetNews is the environment, and human authors are a scarce resource. We argue that this represents an important bridging of text analysis and the Alife research program. Further, it amounts to a novel shift for Alife research -- rather then synthesizing life-like agents we are analyzing a pre-existing environment and discovering life-like behaviors.

In results reported here we group together posts which make use of similar sets of memes. These groups, clouds within a conceptual sequence-space, describe quasi-species. For each quasi-species we compute its time-wise volume of activity by histogramming its daily post levels. We then study the pairwise interaction between quasi-species by computing the cross-correlations between their time series. In our corpus, strong negative cross-correlations signify conditions of competition between the interacting populations where the quasi-species are competing for a limited set of human authors. Furthermore, quasi-species with relatively narrow ecological niches, those which make use of a small number of in-reply-to threads, are more likely to be in competition with other quasi-species. This behavior is analogous to what is found in natural ecologies (Pianka [24]).

Why do these quasi-species compete? Qualitative analysis of the posts, such as those described in the previous section, shows that many competing quasi-species are posts sent to the same or similar threads. Competition is over the scarce authorship resources within these specific thread niches. Over time a particular thread of discussion may bifurcate into two or more internal themes which then proceed to compete for ``air-time'' within the thread.

Acknowledgments

We thank the two anonymous reviewers for their helpful critique of this paper. We gratefully acknowledge the significant input of Richard Pocklington to this work and his careful reading of a draft of this paper. Laurie Hiyakumoto also gave very helpful comments on a draft of this paper. We thank Bernd Schoner and Michael Patrick Johnson. Finally, the author is pleased to thank the sponsors of the MIT Media Laboratory for their financial support.

References

[1] Berry, M.W. (1992) Large-scale Sparse Singular Value Computations. *The International Journal of Supercomputer Applications.* Vol. 6, No. 1.

[2] Berry, M., T. Do, G. O'Brien, V. Krishna, & S. Varadhan (1993). SVDPACKC (Version 1.0) User's Guide. University of Tennessee Computer Science Department Technical Report, CS-93-194.

[3] Berry, M.W. & R.D. Fierro (1995). Low-Rank Orthogonal Decompositions for Information Retrieval Applications. University of Tennessee Computer Science Department Technical Report, CS-95-284.

[4] Best, M.L (1996). An Ecology of the Net: Message Morphology and Evolution in NetNews. Massachustes Institute of Technology, Media Laboratory, Machine Understanding Technical Report, 96-001.

[5] Best, M.L. (1997). An Ecology of Text: Using Text Retrieval to Study Alife on the Net. To appear *Journal of Artificial Life*.

[6] Chatfield, C. (1989). *The Analysis of Time Series An Introduction.* London: Chapman and Hall.

[7] Dawkins, R. (1976). *The Selfish Gene.* New York, Oxford University Press.

[8] Dawkins, R. (1982). *The Extended Phenotype.* San Francisco, WH Freeman.

[9] Deerwester, S. S.T. Dumais, G.W. Furnas, T.K Landauer, and R. Harshman (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science.* 41(6): 391-407.

[10] Dumais, S.T. (1992). LSI meets TREC: A status report. In *The First Text REtrieval Conference (TREC-1)*, ed. D. Harman. NIST Special Publication 500-207.

[11] Dumais, S.T. (1993). Latent semantic indexing (LSI) and TREC-2. In *The Second Text REtrieval Conference (TREC-2)*, ed. D. Harman. NIST Special Publication 500-215.

[12] Eigen, M.J, J. McCaskill, & P. Schuster (1988). Molecular Quasi-Species. *Journal of Physical Chemistry*, Vol. 92, No. 24.

[13] Eigen, M. (1992). *Steps Towards Life: A Perspective on Evolution.* Oxford: Oxford University Press.

[14] Foltz, P.W. (1990). Using Latent Semantic Indexing for Information Filtering. *Proceedings of the 5th Conference on Office Information Systems.* ACM SIGOIS Bulletin vol. 11, issues 2,3.

[15] Frakes, W.B. & R. Baeza-Yates, eds. (1992). *Information Retrieval: Data Structures and Algorithms.* Englewood Cliffs, New Jersey: Prentice Hall.

[16] Furnas, G.W., S. Deerwester, S.T. Dumais, T.K. Landauer, R.A. Harshman, L.A. Streeter, & K.E. Lochbaum (1988). Information Retrieval using a Singular Value Decomposition Model of Latent Semantic Structure. *Proceedings of the 11th International Conference on Research and Development in Information Retrieval (SIGIR)*. New York: Association for Computing Machinery.

[17] Jain, A.K. & R.C. Dubes (1988). *Algorithms for Clustering Data.* Englewood Cliffs, New Jersey: Prentice Hall.

[18] Kantor, B. & P. Lapsley (1986). Network News Transfer Protocol: A Proposed Standard for the Stream-Based Transmission of News. Internet RFC-977.

[19] Lewontin, R.C. (1970). The Units of Selection. *Annual Review of Ecology and Systematics*, Vol. 1.

[20] Lewontin, R.C. (1974). *The Genetic Basis of Evolutionary Change.* New York, Columbia University Press.

[21] May, R.M. ed. (1981A). *Theoretical Ecology Principles and Applications.* Oxford: Blackwell Scientific Publications.

[22] May, R.M. ed (1981B). Models for Two Interacting Populations. In (May 1981A).

[23] Mettler, L. E., T.G. Gregg, & H.E. Schaffer (1988). *Population Genetics and Evolution*, 2nd ed. Englewood Cliffs, NJ, Prentice Hall.

[24] Pianka, E.,R. (1981). Competition and Niche Theory. In (May 1981A).

[25] Pielou, E.C. (1969). *An Introduction to Mathematical Ecology.* New York: Wiley-Interscience.

[26] Pocklington, R. (1996). *Population Genetics and Cultural History,* Msc Thesis, Simon Fraser University, Burnaby.

[27] Pocklington, R. & M.L. Best (1997). Cultural Evolution and Units of Selection in Replicating Text. To appear *Journal of Theoretical Biology.*

[28] Salton, G. & C. Buckley (1988). Term-weighting Approaches in Automatic Text Retrieval. *Information Processing & Management.* 24: 5, 513-523.

Back to Issue 2