Visual Data Mining With Minimal Spanning Trees

Jeffrey L. Solka
Code B10
NSWCDD
Dahlgren, VA 22448

Avory C. Bryant
Code B10
NSWCDD
Dahlgren, VA 22448

Edward J. Wegman
Department of Applied and Engineering Statistics
George Mason University
Fairfax, VA 22030

Introduction

This paper details some of our recent work in the application of graph theoretic techniques to facilitate the data mining process. We are particularly interested in the case where one is presented with a set of n observations in p dimensions space along with a set of provided class labels. These class labels can be provided by a human operator or may be provided based on some other automated classification strategy. In either case we are interested in employing these strategies to facilitate two goals. The first goal is the discovery of unusual relationships that may exists between the observations in one class and the observations in another class. The second goal is that we are interested in being able to cluster the set of observations and ultimately to weigh the clustering results against the provided class labels. In order to illustrate the success of our formulated approaches we will provide results based on mining two small sets of text documents. We do wish to emphasize the the discussed graph theoretic approaches and visualization frameworks are not predicated on the data type being text data but are equally appropriate for other data types. In fact one must merely be able to compute some sort of interpoint distance measure between the observations. This interpoint distance measure does of course define a complete graph on the set of vetices (observations) and we can proceed forward from there with our analysis.

Approach

As mentioned in the introduction, we will be using the text data mining problem as a representative application of our developed methodologies. We must first decide upon a way to capture the semantic content of the documents, extract features from the documents, and then we must be able to compute interpoint distance measures on these features. Once we have the interpoint distance matrix in hand then we can proceed forward with our analysis.

We have chosen to use the bigram proximity matrix of Martinez [Martinez, 2002] as a means for capturing the semantic content of the documents. The bigram proximity matrix (BPM) encodes the document as a n-words by n-words matrix

where an entry in the ij position of this matrix indicates that the ith word in the multi-corpus word list is proximate to the jth word in the multi-corpus word list in one of the sentences in the current document of interest. It is important to point out that stopper or noise words have been removed from the document prior to the computation of the BPM. The BPM entries that were studied by Martinez were either 0 or 1 to indicate the presence or lack of a particular word pair relationship or else a count of the number of times that a particular word pair appears in the document. The important point to note is that the research of Martinez strongly supported the claim that the BPM adequately captures the semantic content of documents using either coding scheme. He tested the semantic capture capability of the BPM and TPM via supervised learning testing, unsupervised learning testing, and three different hypothesis tests.

Given the set of BPMs that describe our document set we need a means of measuring the distances between the documents. Martinez compared over ten different ways of measuring distances between the BPMs. We have chosen to use the Ochai measure given bellow.

$$S(X,Y) = \frac{|XandY|}{\sqrt{|X||Y|}}. \tag{1}$$

This similarity is then converted into a distance using

$$d(X,Y) = \sqrt{2 - 2S(X,Y)}. \tag{2}$$

We have chosen this similarity/distance measure in part because of the favorable results obtained by Martinez using them

At this point in time we have an interpoint distance matrix where hopefully articles that are close semantically will have a small value for that row column pair in the matrix and articles that are far apart semantically will have a high value for that row column pair in the matrix. We also have a previously obtained categorization of the articles. So how do we explore/exploit this interpoint distance matrix in order to facilitate the discovery of interesting relationships and clusters? Let's first consider the problem of interesting relationship (serendipity) discovery. The simplest thing that one could do is to find the closest articles between each pair of categories. We initially tried this and although the results were interesting we found them less than fully satisfying. Alternative exploration strategies are suggested by considering the discriminant boundaries that separates the articles from the two categories. One might, for example, consider looking for those articles that are along the boundary that separates the two categories. Taking a step up from this "first order approximation," one could look for articles that have the same "relationship" to the discriminant boundary. Finally one might wish to provide the user with the capability to actually be able to select a local "focus of attention," to make a more in-depth study of the local geometry based on this focus of attention.

So what is required is a means to ascertain the relationships of the observations to the discriminant boundary. The key to understanding this relationship is to realize that if we view the articles as vertices in a graph, the interpoint distance matrix defines a complete graph on this set of vertices. Exploration of the complete graph is impossible so we need a subgraph of the complete graph that captures the information in the complete graph and in particular the relationship of the observations to the discriminant boundary. One natural candidate to capture this information is the the minimal spanning tree (MST). The MST is loosely defined

as a tree that traverses all of the nodes in the complete graph with a minimal sum of edge weights.

We [Solka and Johannsen, 2002] had previously used methods based on the MST, as computed on the interpoint distance matrix, to characterize discriminant surface complexity. This led us to believe that this would be an appropriate tool to facilitate the exploration of cross corpora (class) relationships in the document set. Our previous efforts had suggested that the number of edges that travel between observations of different classes is a good surrogate measure for classification complexity. This would suggest that one would have a convenient means to provide a human operator with the capability to explore the cross corpora relationships by providing him with a tool that facilitated an efficient exploration of the MST as computed on the two class set of documents along with the associated cross-copora edges.

The only missing piece in the puzzle is a way to provide a user with a "drill down" capability so that they can focus their attention on a set of nodes rather than the full MST. We chose to accomplish this by allowing the user to choose a subset of observations (nodes) and a depth exploration parameter k. This parameter specifies how far, in terms of number of edges, we wish to travel from our original node to include in our subsequent analysis. For example the user might select nodes 14 and 2 for his focus of attention and selected a search parameter of $k = 3$. The user selected nodes are colored red while the nodes within 3 links of these are colored blue. In the next stage of the analysis the user is presented with the complete subgraph. This is rather busy of course and it is necessary for the user to have the capability to remove some of the edges, based on edge strength, before they proceed forward with their analysis. We will have more to say about this when we discuss our software implementation of the user environment. Figure 1 presents a "cartoon" of how the selected nodes interact with the depth search parameter.

We were also interested in providing the user with the capability to explore clusters of observations (documents). Previous analysis in the literature had indicated that one can obtain all the possible single-linkage hierarchical clusterings of a set of observations merely by sequentially deleting the edges from the MST of the complete graph starting with the largest edge. There had been previous work appearing in the literature on the application of MSTs to geospatial data clustering [Guo et al., 2003, Guo et al., 2002] and gene expression analysis [Xu et al., 2001, Xu et al., 2002]. We choose to utilize a MST as computed on the full set of observations, all classes, as a starting point for our cluster analysis. We proceed forward cutting edges in the MST in order to obtain a clustering at a particular number of clusters c. We choose which edge to cut based on the edge strength divided by the mean of the associated edges of path length k adjacent to the chosen edge.

Let's now turn our attention to our software implementation of these algorithmic schemes. The BPM and similarity matrix calculations were implemented in C#. The MST calculations were implemented based on Kruskal's algorithm in JAVA. The visualization environment was implemented in JAVA and the graph layout was accomplished using the general public license package TouchGraph, www.touchgraph.com. TouchGraph supports zooming, rotation, hyperbolic manipulation, and graph dragging.

We originally implemented the software system with automated serendipity exploration in mind. Figure 2 presents the opening screen of our automated serendipity system. This screen presents the user with a webpage that provides links to all of the "choose 2" corpora within the text dataset. The user can easily navigate in order to study a particular comparison by first clicking the appropriate corpora

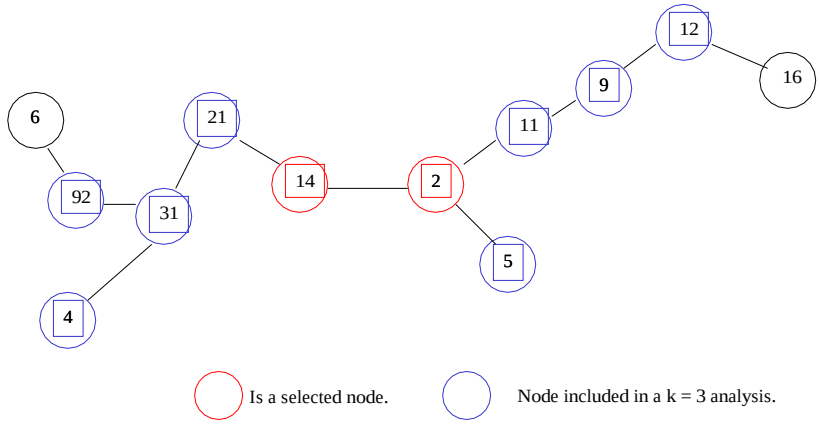

Figure 1: User specification of drill down focus region.

link at the top of the screen and then subsequently clicking which corpora that the user would like to compare the currently focused corpora to. Alternatively the user may simply scroll down to the point in the webpage that contains the association of interest.

Once a user chooses two categories to explore they are presented with a rendering of the MST based on a spring-based layout model. Figure 3 presents one example of this screen. We draw the users attention to the fact that we have provided an edge weight legend at the bottom of the plot. We have used a blue, green, yellow, orange, and red color scheme that is spread proportionally among the edge weight values. We have chosen to "reuse" the color map. We use it once to represent the intra-class edges and once to represent the inter-class edges. It is usually easily discernible to the user which situation that they are dealing with. We have also provided the user with a drop down menu on the lower left that has the edges sorted on inter-class strength from the weakest at the top of the list to the strongest at

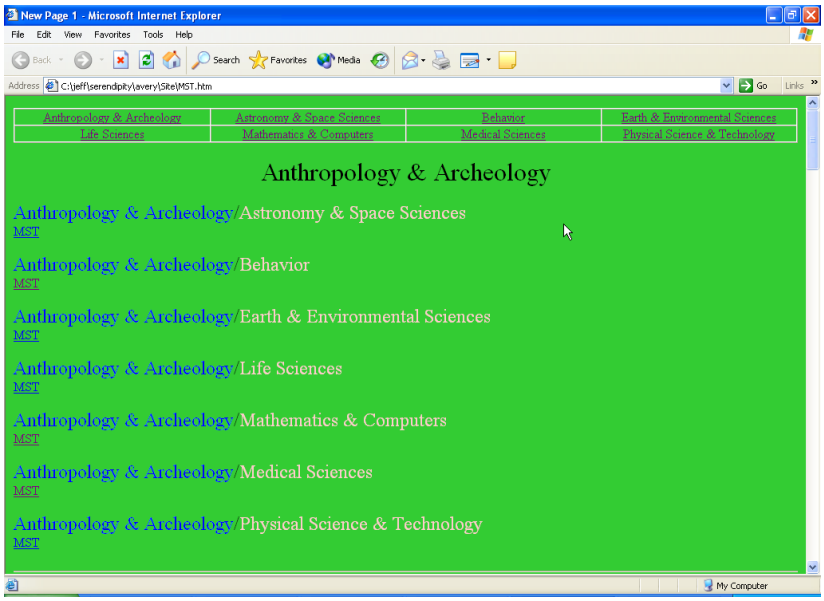

Figure 2: Opening screen of the automated serendipity search engine.

bottom of the list. In this way the user can rapidly explore the serendipitous type relationships. Experience also led us to change the symbols of the nodes as they are explored, In this way the user can keep track of his exploration of the graph.

Once a user has selected a particular edge to display then they are presented with a comparison file that highlights the word pairs from the BPM that the two articles have in common and also enumerates them, please see Figure 4.

The software system that provides a point picking capability and the software system that allows clustering are extensions of the basic framework that has been previously discussed. We think it prudent to say a few words about each of these in turn. First we will discuss the automated serendipity system that allows point picking. Figure 5 shows a sample screen from this version of the software. The main difference is that the user is provided with the capability to select two or more nodes for a focus of attention along with a regeneration parameter k and a button to pick to regenerate the complete graph with the vastly limited focus of attention.

Figure 6 shows the complete graph for the reduced focus of attention. It is quite apparent from a cursory examination of the figure that even the reduced complete graph is much too busy to effectively explore.

Figure 7 shows the complete graph for the reduced focus of attention where the user has chosen only to show those edge weights with a strength of orange or greater. Things are more manageable now. We will relegate the discussions of the clustering version of the software till the results section.

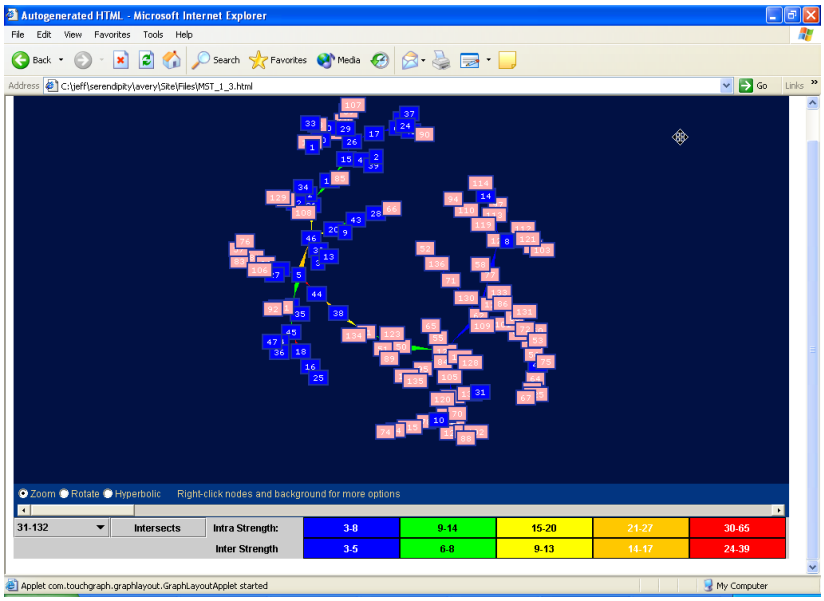

Figure 3: An example of the MST layout screen.

Results

Datasets

Now that we have briefly outlined the visualization framework, let's examine it's application to two different text multi-copora document sets. The first document set that will be discussed consists of 1117 documents that were published by Science News (SN) between 1994-2002. They were obtained from the SN website on December 19,2002 using the LINUX wget command. Each article is roughly 1/2 – 1 page in length. The corpus html/xml code was subsequently parsed into straight text. The corpus was read through and categorized into 8 categories by one of us, AB. The 8 categories were based on the fact that a very small subset of the articles had already been placed in these categories on the Science News website. The categories and their respective article counts are as follows: Anthropology and Archeology (48), Astronomy and Space Sciences (124), Behavior (88), Earth and Environmental Sciences (164),Life Sciences (174), Mathematics and Computers (65), Medical Sciences (310), Physical Sciences and Technology (144). The second document set is based on the Office of Naval Research (ONR) In-house Laboratory Independent Research (ILIR) Program. This program is the primary way that the Office of Naval Research funds the Naval Warfare Centers to conduct fundamental research. This set of documents was provided to us and consists of 343 documents each of which is about 1/2 page in length. Avory Bryant and Jeff Solka jointly read through these documents and classified them into previously defined ONR Technology Focus Areas. The number of documents in each of the ONR Focus Areas are as follows, Advanced Naval Materials (82), Air Platforms and Systems (23), Electronics (0), Expeditionary/USMC (0), Human Performance/Factors (49), Information Technology and Operations (18), Manufacturing Technologies (21), Medical Science and Technology (19), Naval and Joint Experimentation (0), Naval Research Enterprise Programs (0), Operational Environments (27), RF Sensing, Surveillance, and

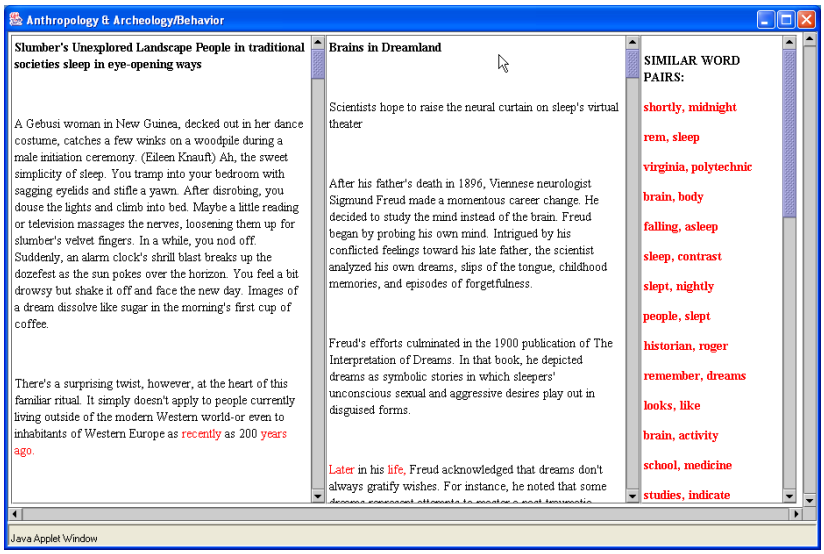

Figure 4: A sample comparison file.

Countermeasures (27), Sea Platform and Systems(38), Strike Weapons (0), Undersea Weapons (0), USW-ASW (5), USW-MIW (17), and Visible and IR Sensing, Surveillance and Countermeasures (17).

Feature Extraction

Both sets of documents were processed to remove standard stopper words prior to the extraction of the bigram proximity matrices. The remaining words were not stemmed in that the current version of the software does not readily support stemming. The BPMs were then extracted for each document. The Ochiai similarity measure and associated distance was then computed on each document pair using the equations described above.

Automated Serendipity Extraction on the Science News Data Set With No User Driven Focus of Attention

Let's begin our analysis of these document sets by first considering the Science News corpus. We will first analyze this set of data using the automated serendipity software system that does not support point picking.First we choose Mathematics and Computer Sciences as compared to Physical Sciences and Technology. Given the strong overlap in these two discipline areas we would expect that there might be strongly associated articles across the discipline areas. Figure 8 shows the MST for these two categories. We have selected the edge that represents the strongest association between the two categories. Right clicking on this link and choosing the view comparison file leads us to Figure 9. There are numerous associations between these two articles. Some of these are highly relevant such as (allow, electrons), (electronic, devices), (circuit, components), (atomic,scale), (biological, molecules), and some may be either identified as spurious or as encoding "meta associations" that may or may not be relevant such as (santa, barbara), (packard, laborties), (palo,alto), (hewlett, packard) and others. There are even associations that are

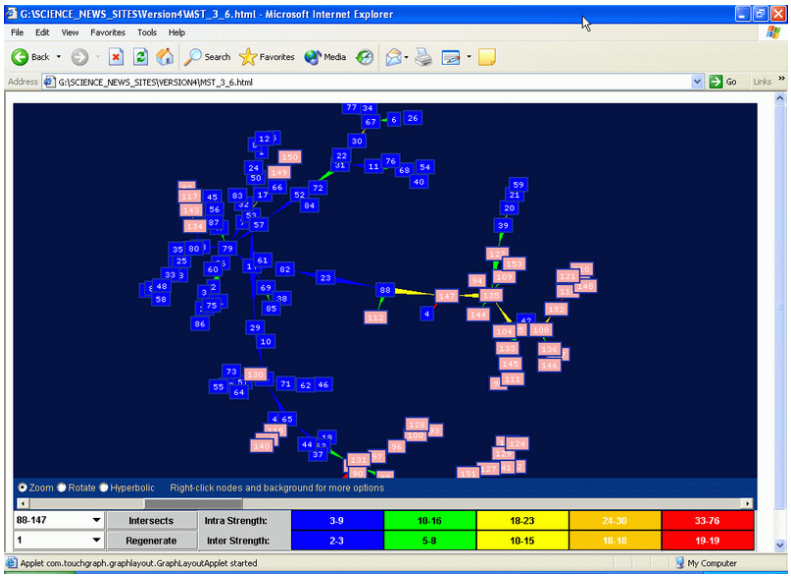

Figure 5: A example screen for the automated serendipity system with included point picking.

clearly "useless" such as (research, team) and (research, group). The "human identified" connection between the two articles is that fact that the first article from the Mathematics and Computer Science area is about computer chip design while the second article from the Physical Sciences and Technology arena is about materials science design at the nano level.

Next let's consider two discipline areas that are more conceptually distant such as Anthropology and Archeology as compared to Medical Sciences. Figure 10 shows the MST for this case along with the highlighted edge that indicates the strongest association between these two discipline areas. Right clicking on this link and choosing the view comparison file leads us to Figure 11. Once again the methodology has identified two articles with a large overlap in their BPMs and hence a large similarity in semantic content. A few of the very important word associations include (major, depression), (family, community), (mental, health), and (mental, disorders).Once again their are also several meta-level associations identified such as (delhi, india) and (anthropologist, veena). There are also some association that are spurious such as (nearly, years). A cursory examination of the two articles does indicate that the association is relevant. The first article from the Anthropology and Archeology category is about depression in people that are living in violent societies while the second article from the Medical Sciences category is about depression in individuals that suffer from disease such as AIDS that of the lead to many in society stigmatizing them.

Lets consider two more discipline pairs before moving on to the ILIR data.The first article pair comes from the Earth and Environmental Sciences and Physical Sciences and Technology categories. We have already provided enough examples of the MST pictures to forgo the display of these in the following example. There are some interesting things to say about the location of the associated articles in the MST but this type of analysis will not be performed within. We proceed exactly to the two strongest associated articles between these two categories, see Figure

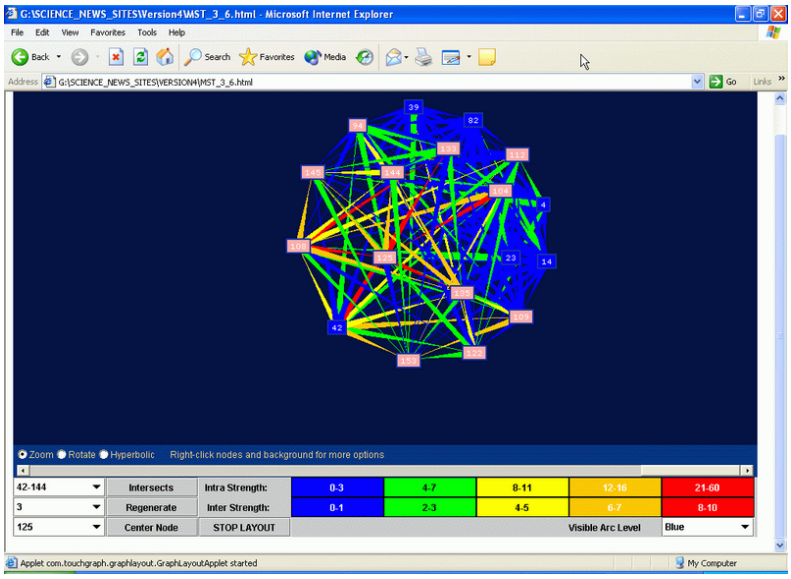

Figure 6: A example screen for the automated serendipity system with point picking where the user has chosen to limit the focus of attention based on k=3.

12. The first article from the Earth and Environmental Sciences category entitled "Power Harvests" is about wind generated power. The second article entitled "Oceans of Electricity," is about new methodologies for generation of power from ocean waves. These two articles are highly associated including numerous word pairs. It seems very reasonable to assume that those researchers that are involved in work on wind generated power would be interested in new technologies that may have been developed for ocean generated power. Ocean generated power research might include work in power conversion and battery storage. Both of these would be relevant to wind power research.

The last discipline area combination that we wish to consider is Anthropology and Archeology and Mathematics and Computer science. This might be considered the most interesting combination so far in that one would not expect that there would necessarily be a natural association between these areas. Once again proceeding directly to the comparison file between two articles as identified from a cursory examination of the MST. We point out the fact that this is not the strongest association but merely an interesting one, see Figure 13. The Anthropology and Archeology article is entitled "An artist's timely riddles," and is about the artist Marcel Duchamp while the Mathematics and Computer Science article is entitled "Visions of infinity tiling a hyperbolic floor inspires both mathematics and art," is about the tilings of the Poincare disk by M. C. Escher. The number of BPM commonalities between these two articles is not that large, 9, but there are highly relevant associations among the list of associations including, (different, perspectives), (euclidean, geometry) and (non,euclidean). It seems that Duchamp was also involved in explorations of non-Euclidean geometries as was Escher. The work by Escher is much well known but the authors had no idea about the Duchamp work.

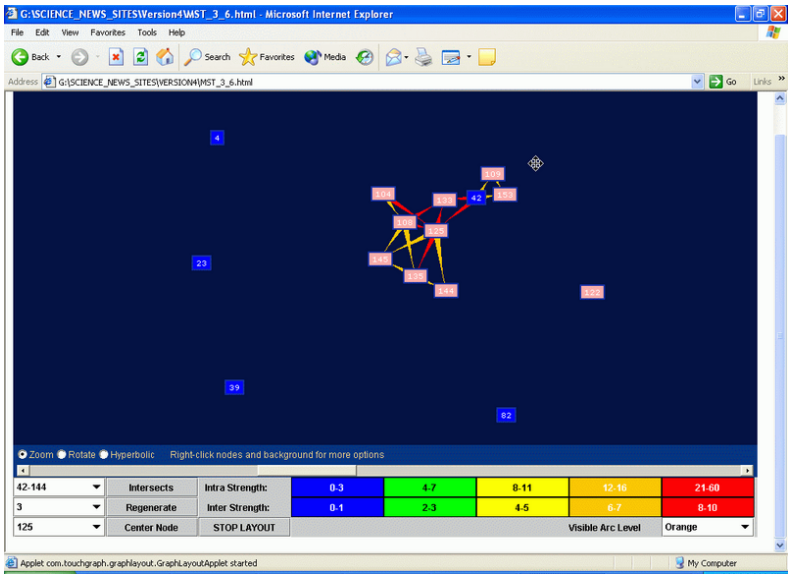

Figure 7: A example screen for the automated serendipity system with included point picking with point picking where the user has chosen to limit the focus of attention based on k=3 after the user has deleted those edges with a strength less than orange.

Automated Serendipity Extract on the ONR ILIR Data Set With No User Driven Focus of Attention

Next we turn our attention to an analysis of the ONR ILIR data using the same tool. The application of the tool to the ILIR articles was compounded by two facts. First the ILIR articles are much shorter in length than the Science News articles. Second the ILIR articles are a somewhat more orthogonal set of articles than the Science News articles. There are many more articles that come from the same type of discipline within the Science News dataset. This last fact may be in part due to the nature of the ILIR dataset and to the fact that there is just a shorter set of articles.

The first pair of articles that we will consider come from the Air Platform and Systems and the Sea Platform and Systems category. The two articles are presented in Figure 14. These two articles are about computational fluid dynamics as applied to the two discipline area. The list of commonalities in the BPMs is short but contains several meaning associations such as (navier, stokes), (high, reynolds), (reynolds, averaged), and (reynolds, number). The two projects were executed at different labs, NSWCDD and NAVSTO, during different years, FY01 and FY99/00. Clearly there would be benefits between collaborative discussions between the two investigators or the later investigator could benefit from some of the insights obtained by the earlier group.

The next two articles come from the Air Platform and Systems and the Sea Platform and Systems arena. These two articles are both about combustion physics, see Figure 15. One of the projects was executed at NSWCDD in FY99 and the other project was executed at NUWC in FY01. Each project was conducted by a different investigator.

Another capability that our proposed automated serendipity approach possesses

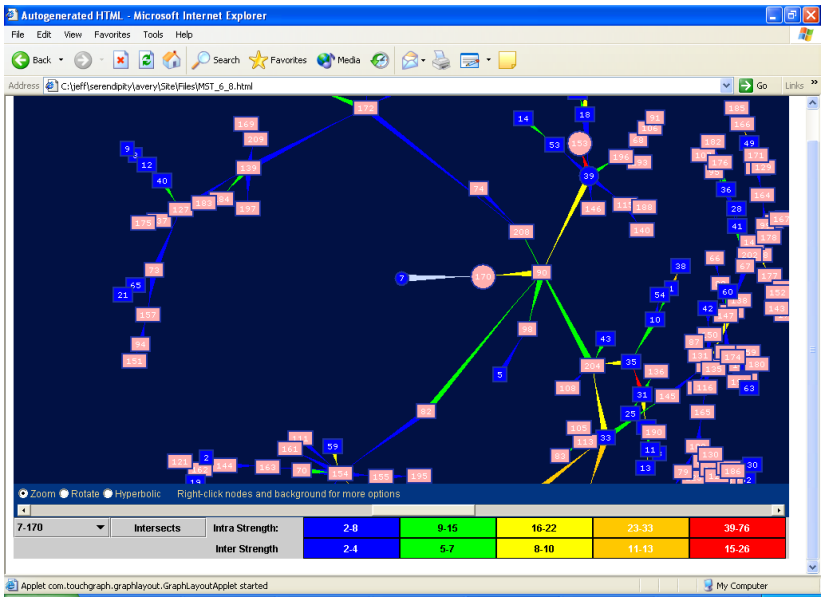

Figure 8: The MST for the Mathematics and Computer Sciences category as compared to the Physical Sciences and Technology category. The edge indicating the strongest association has been highlighted.

is anomaly identification, consider Figure 16. A project has clearly been entered into the database twice.

The next example illustrates another not quite so blatant anomaly, see Figure 17. These two abstracts come from two different projects executed by the same investigator. One project was executed during FY99 and the other was executed during FY01. The two abstracts seem to contain a section that is virtually identical but one might not have guessed this based on the titles of the two abstracts.

The next ILIR example that we will consider is perhaps one of the most curious ones that we uncovered in the database, see Figure 18. One of these articles was entitled "Direct adaptive, gradient descent and genetic algorithm techniques for fuzzy controllers," and one was entitled "Mission scenario classification using parameter space concept learning." A quick read of the titles of these two efforts would not lead one to believe that there would be any association between them but these two articles (abstracts) are virtually identical and perhaps even more curious was the fact that these two abstracts reflect two projects that were being executed by the same investigator, Venetsky, during the same years, FY99/00, at the same laboratory, NUWC. There is clearly something unusual going on here that while not nefarious is certainly curious.

We close this subsection with another cross focus area curiosity, see Figure 19. These two articles are entitled "AAWCREW performance as a function of induced thermal stress," (the Human Performance Factors article), and "Mathematical model of female thermoregulation," (the Medical Science and Technology article). Both of these articles seem to be about effects of temperature changes on female AAW crew members . Both of the efforts are being funded by the same lab, NAVSTO, in the same funding cycle (FY99/FY00)), and being conducted by the same researcher, Shender. The two articles are virtually identical. So one would surmise that this represents a two effort thrust that is being funded in the same

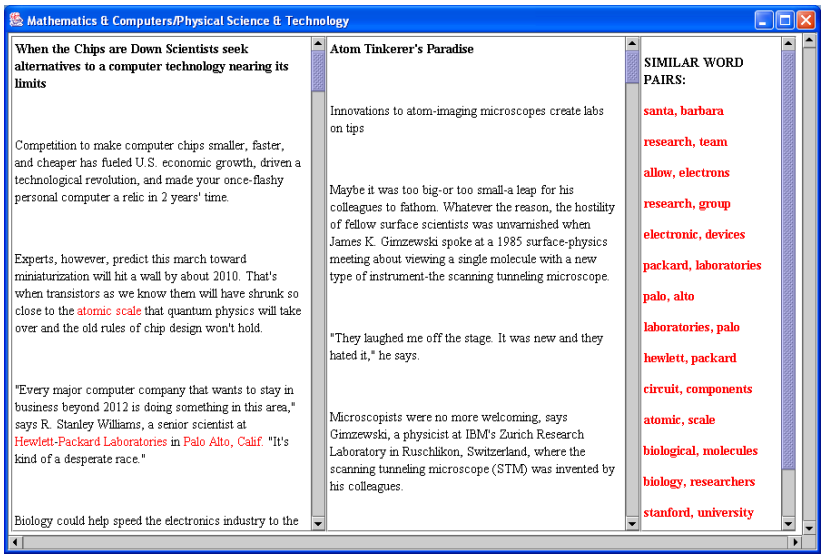

Figure 9: The BPM comparison file for the strongest association between the Mathematics and Computer Sciences category and the Physical Sciences and Technology category.

discipline area. So we have effectively identified an artifact of the funding strategy of the organization.

Automated Serendipity Extract on the Science News Data Set With User Driven Focus of Attention

Now we will consider how the insights obtained using the automated serendipity system are changed by allowing the user to focus their attention on a subset of the full set of nodes within the MST, specify a depth around these nodes to explore and finally examines the complete graph surrounding these nodes for interesting serendipitous types of relationships. We will illustrate this methodology via an analysis of the Science News dataset. Let's consider the Earth and Environmental Sciences vs. the Medical Sciences category. First the user must choose the nodes to focus attention on, see Figure 20. Figure 20. These two articles are entitled "Common pollution undermines masculinity," and "New concerns about phthalets ingredients of common plastics, may harm boys as they develop," see Figure 21. The two articles have 32 commonalities in their BPM and the first article details studies concerning the effect of phthalates on animals while the second paper details similar types of studies in humans. After specifying a depth parameter of 7 one obtains the following complete graph, Figure 22. We remind the reader that nodes 196 and 137 have been removed prior to the redndering of the complete graph. The complete graph is rather busy, looking much like a ball of rubber bands. Let's now, using the tool, remove those edges that are weaker than a strength of orange, see Figure 23. Now some structure is clearly discernible in the graph. There is a central cluster and several singleton groups surrounding this central cluster. Let's take a moment to discuss the structure of the central cluster. There are nodes (70, 217) with 17 associations. Node 70 comes from the Earth and Environmental sciences group and is entitled "Banned pollutant's legacy:lower IQs," while node 21

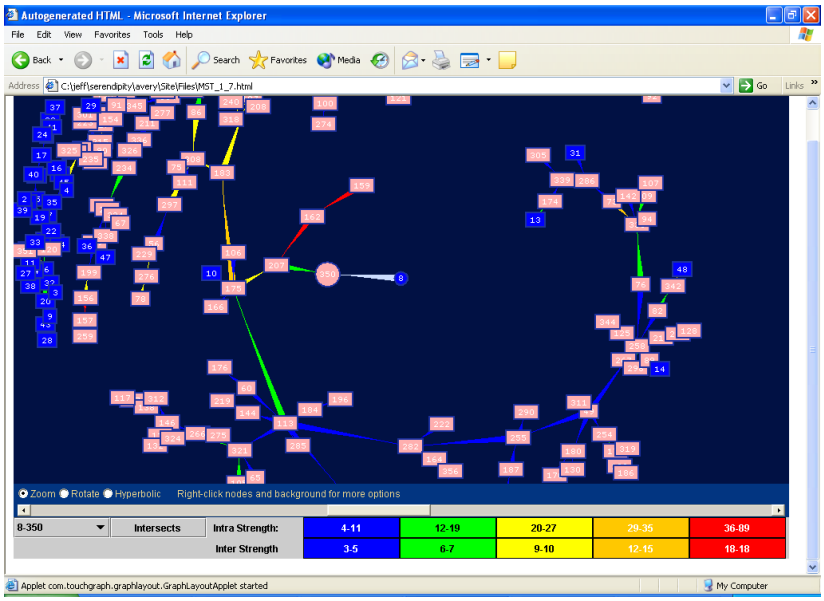

Figure 10: The MST and highlighted strongest edge between the Anthropology and Archeology discipline area and the Medical Sciences discipline area.

comes from the Medical Sciences category and is entitled "Lead therapy won't help most kids." Node 70 details the effects of PCBs on IQ while node 217 discusses the effect of lead. Node 70 also shares numerous (12) associations with node 231. The article corresponding to node 231 is entitled "Burned by flame retardants," which comes from the Medical Sciences category. Node 231 discusses PCBs in flame retardants. Node 231 also shares 12 associations with node 90 from the Earth and Environmental Sciences area. The article corresponding to node 231 is entitled "The rise in toxic tides," and discusses various sorts of biologically produced toxins killing land animals (cows) and marine animals (fish). Finally we point out an 12 associations between nodes 162 and 369. Node 162 is entitled "Estrogen's emerging manly alter ego," and details DDTs mimicking estrogen while node 369 is entitled "Does light have a dark side? Nighttime illumination might elevate cancer risk" discusses that fact that night lights may effect estrogen related maladies such as breast cancer.

Now that we have briefly outlined some of the interesting associations in this central cluster, let's take a little time to look a little closer at some of the more interesting associations. First we will consider the node (70, 217) pair, see Figure 24. There are numerous associations between these two articles some of which are clearly spurious such as (national, institute) and some of which are highly relevant such as (environmental, health) and (heavily, exposed). There are even meta associations such as a common researcher (Walter, Rogan). The totality of these associations is sufficient in order to make a meaningful connection and in fact the authors posit the idea that it might be fruitful for investigators to examine the synergistic effect of PB and PCB exposure.

It is worthwhile to more closely examine one other pair, (162,369), see Figure 25. Linkages between that articles based on topic environmental health and estrogen exposure were sufficient to lead to a meaningful association. These two might lead one to consider synergistic effects between DDTs lowered estrogen exposure due

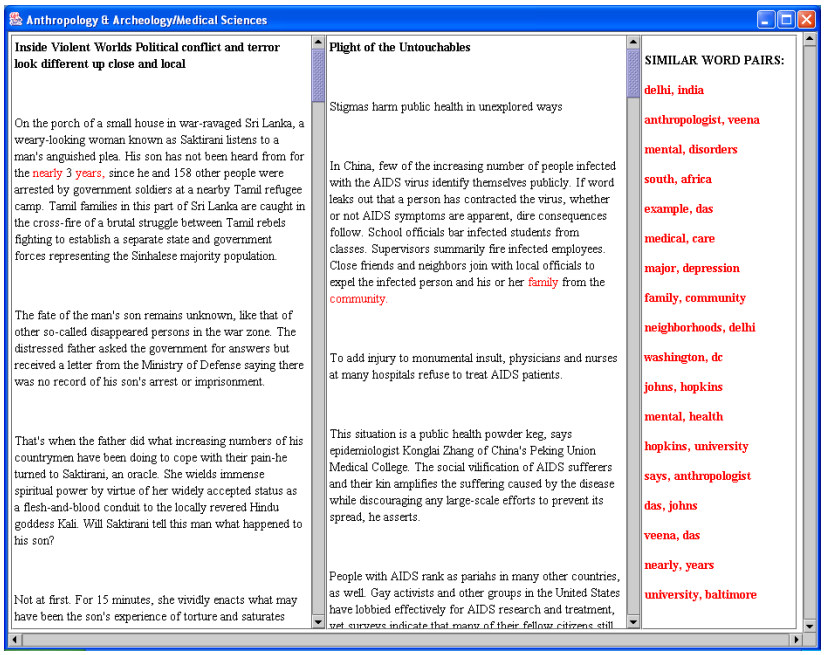

Figure 11: The BPM comparison file for the strongest association between the Anthropology and Archeology category and the Medical Sciences category.

to DDT substitution and night time illumination or EM fields (increased estrogen exposure).

Let's conclude this session by performing a similar analysis on the categories behavior and mathematics/computer science. Figure 26 illustrates the MST for this combination along with our nodes of interest (88,147). The comparison file is presented in Figure 27. The first article from the Behavior category is entitled "Marginal groups thrive on the internet," is about the existence of marginal groups in cyber space while the second article from the Computer Science and Mathematics category is entitled " The social net," and is about the application of social network theory to network based interactions.The two articles share many meaningful associations. Choosing these articles as the focus of attention along with a number of links parameter of 7 and then removing those links that are weaker than a strength value of orange yields Figure 28. Let's take a moment to discuss the articles that remain in the central group. Nodes (10,108) possess 10 associations and are entitled "Simple minds, smart choices," an article from the Behavior category about "bounded rationality", and one entitled "The soul of a chess machine," an article from the Mathematics and Computer Science category about computer chess programs. We will have much more to say about this article pair shortly. The next article pair is (80,133). Article 133 from the Mathematics and Computer science category is entitled "Playing your cards right," is about a computer poker program. The next article pair is (80,135). Article 135 from the Mathematics and Computer Science category is entitled "Agents of cooperation," and is about intelligent agents. Let's take a closer look at the node (80,108) associations, see Figure 29.

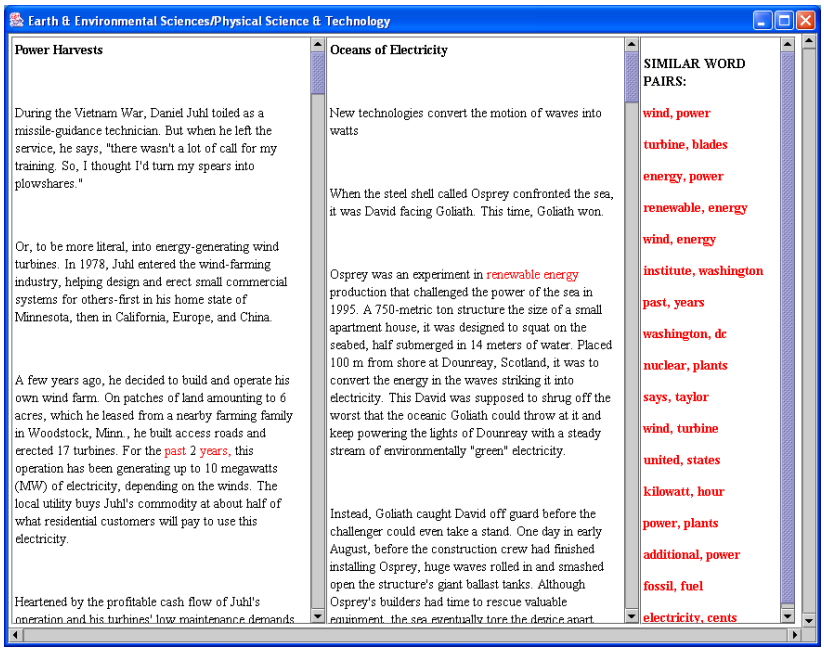

Figure 12: The BPM comparison file for the strongest association between the Earth and Environmental Sciences category and the Physical Sciences and Technology categories.

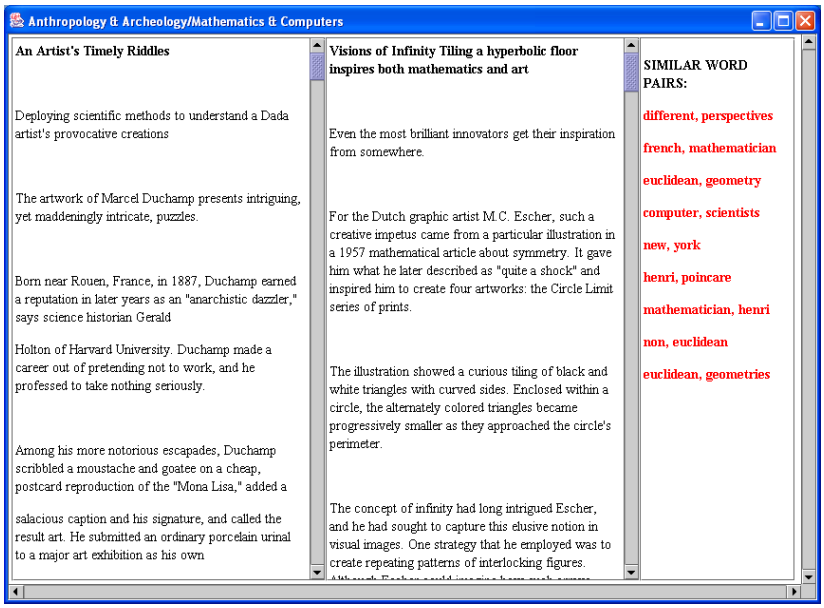

Figure 13: The BPM comparison file for an interesting association between the Anthropology and Archeology and Mathematics and Computer Sciences category.

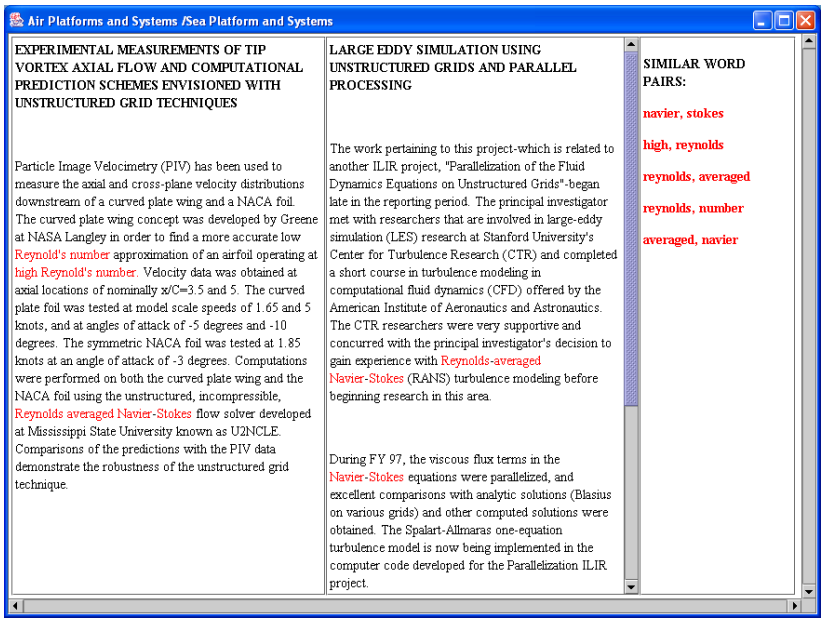

Figure 14: The BPM comparison file for a computational fluid dynamics association between the Air Platform and Systems and the Sea Platform and Systems classes.

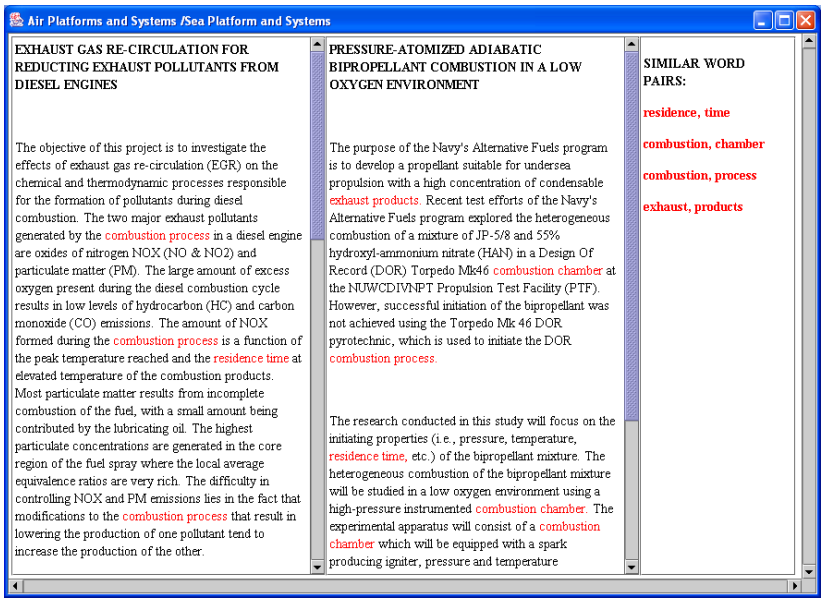

Figure 15: The BPM comparison file for a combustion physics association between the Air Platform and Systems and the Sea Platform and Systems classes.

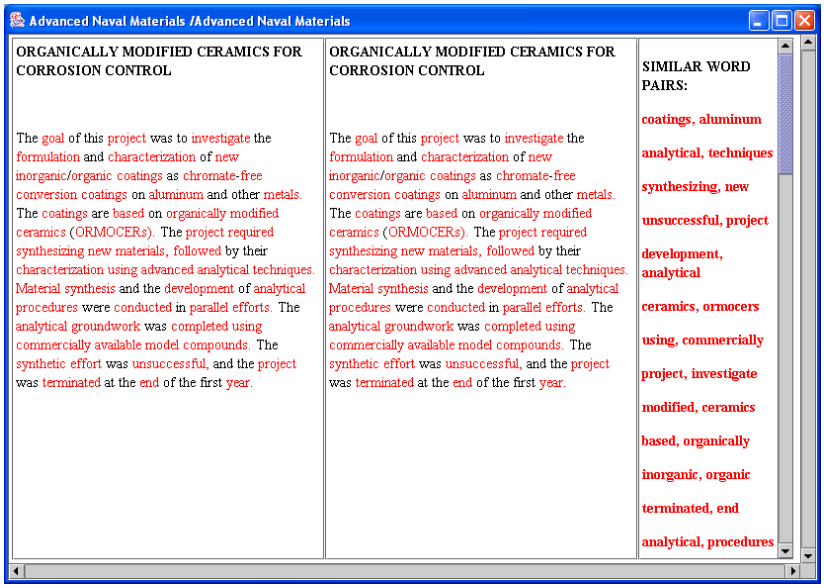

Figure 16: An apparent duplication entry in the Advanced Naval Material category.

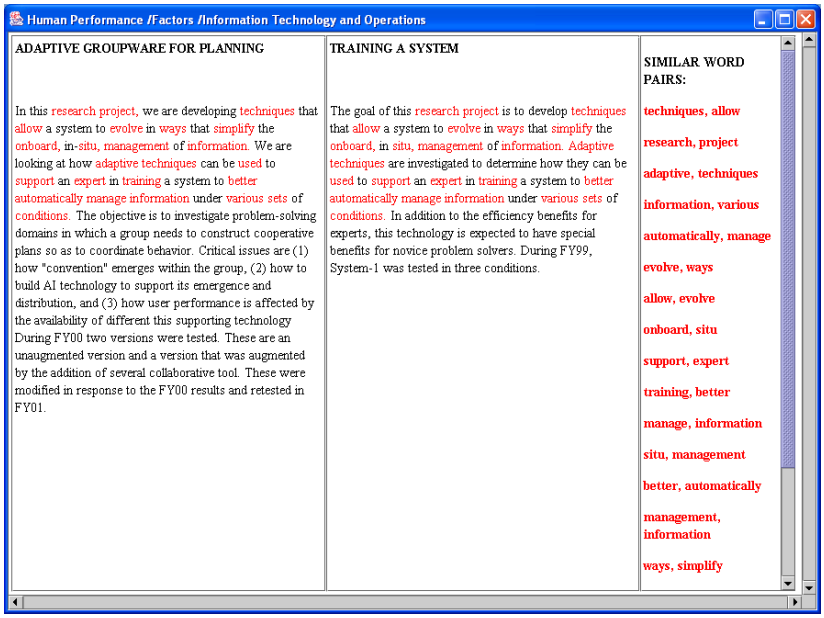

Figure 17: An unusual association across the Human Performance Factors and the Information Technology and Operation focus areas.

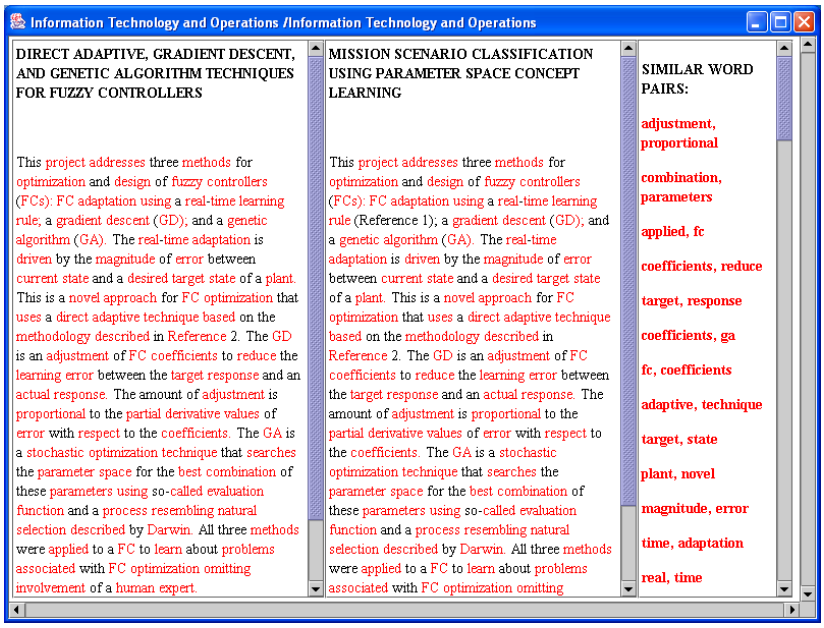

Figure 18: A highly unusual association between two articles in the Information Technology and Operations focus areas.

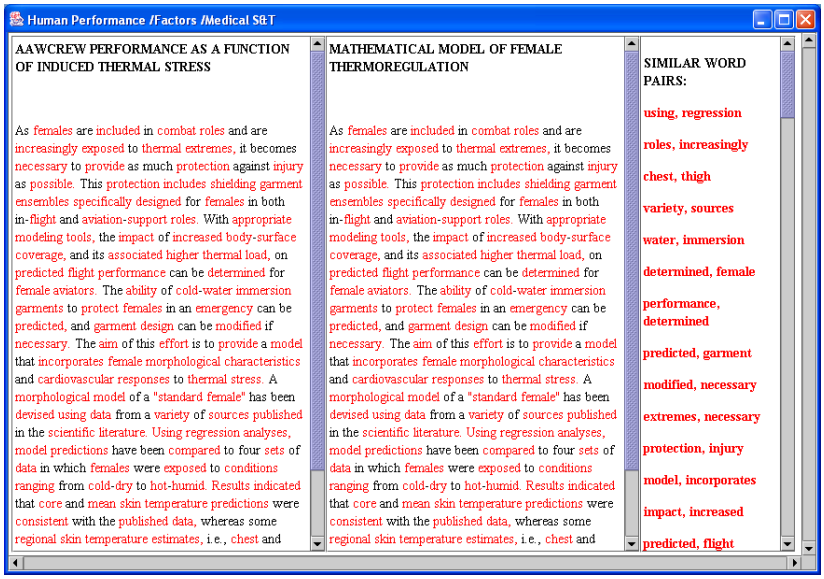

Figure 19: A highly unusual association between two articles in the Human Performance Factors and the Medical Sciences and Technology focus areas.

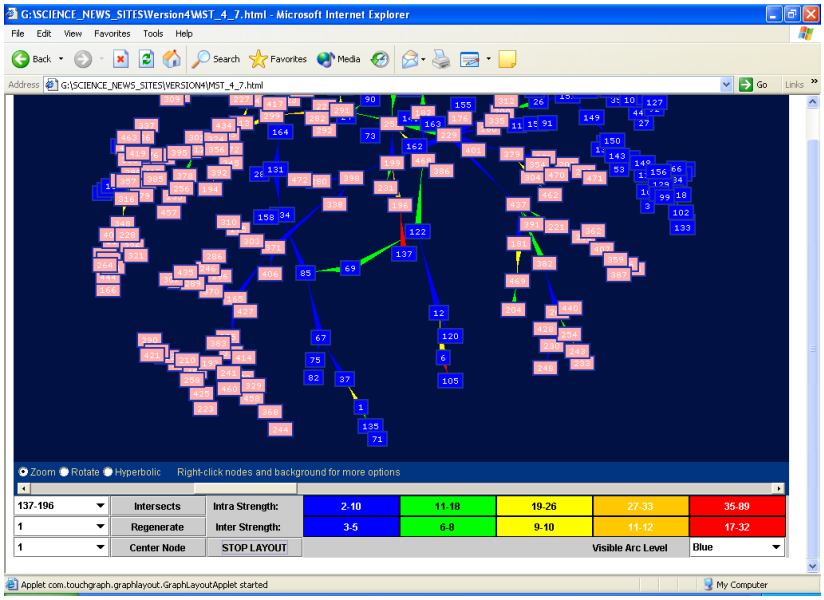

Figure 20: The user has selected nodes 196 and 137 as the "focus of attention."

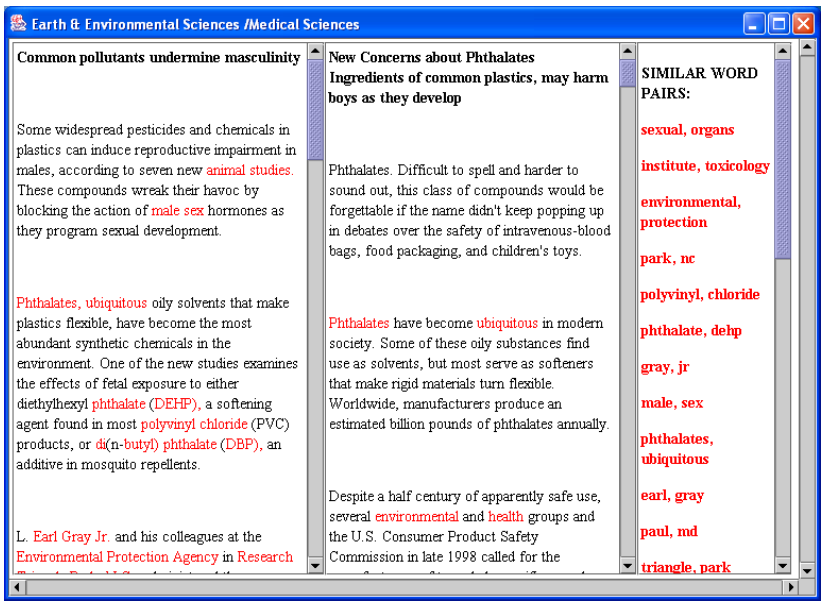

Figure 21: A comparison for the two articles selected for the"focus of attention."

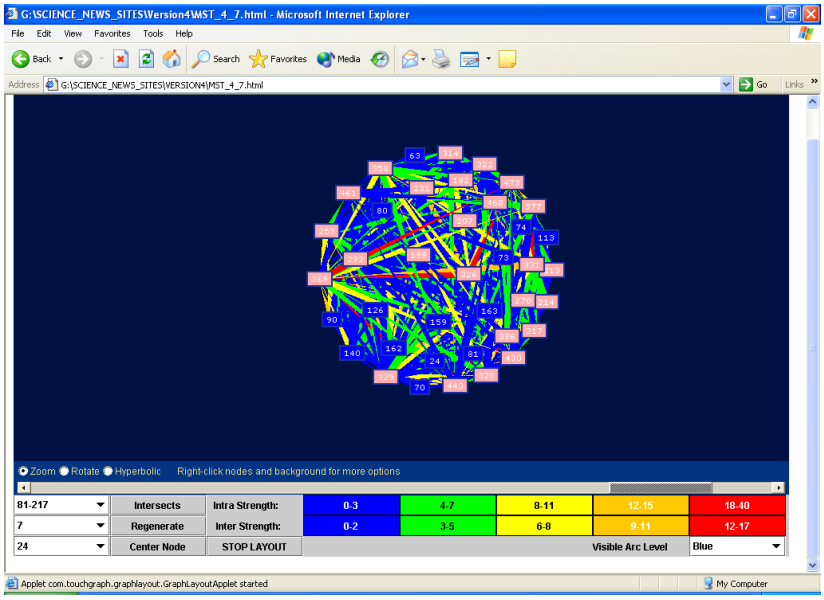

Figure 22: The complete graph based on those articles within 7 links of articles 196 and 197.

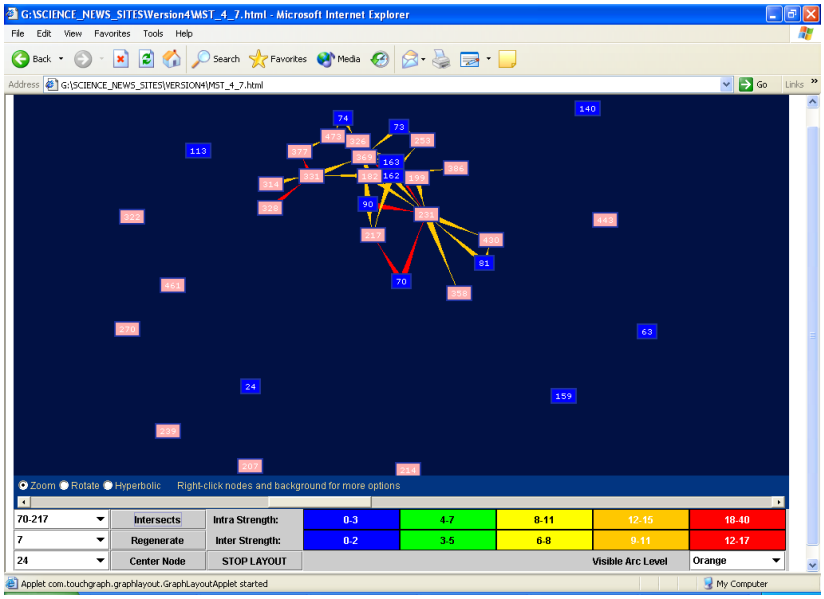

Figure 23: The complete graph based on those articles within 7 links of articles 196 and 137 with those edges with a strength less than orange removed. The graph has been centered on the edge extending between nodes 70 and 217.

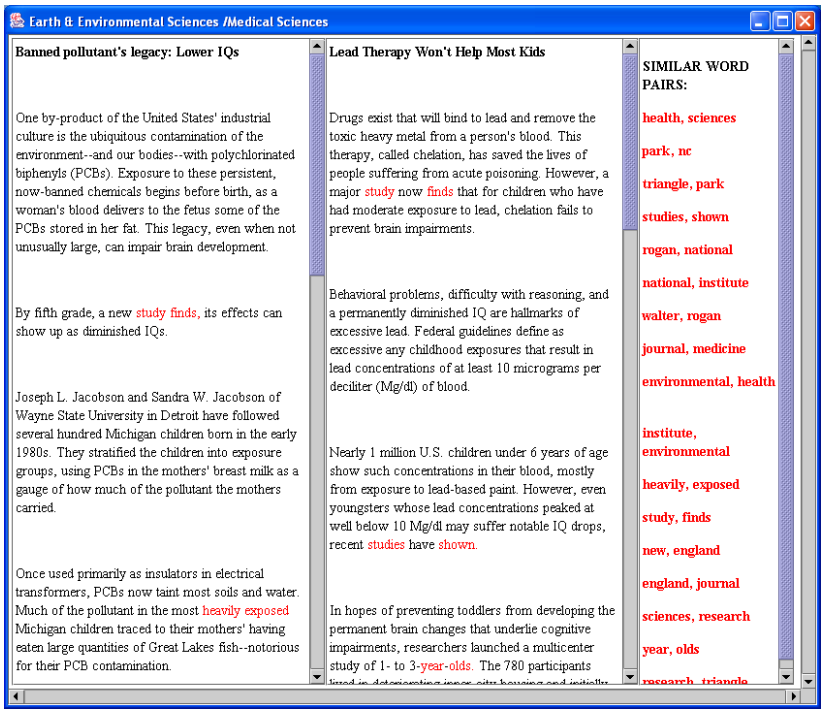

Figure 24: A closer look at the associations between nodes 70 and 217.

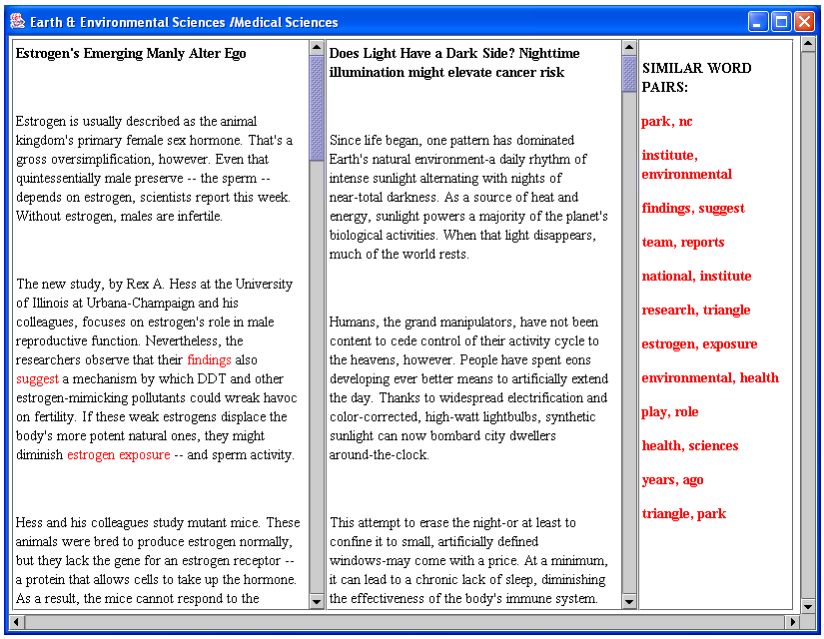

Figure 25: A closer look at the associations between nodes 162 and 369.

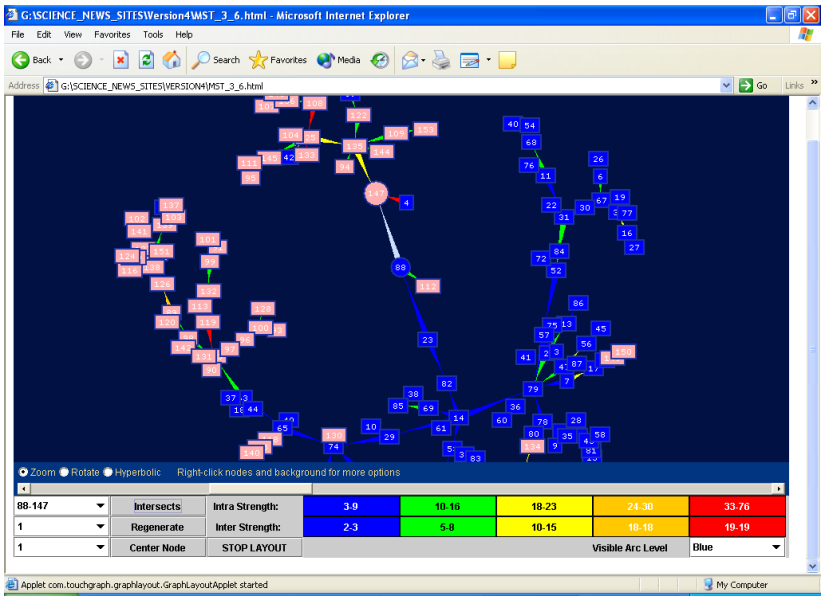

Figure 26: The MST and nodes 88 and 147 of the behavior and computer science and mathematics categories.

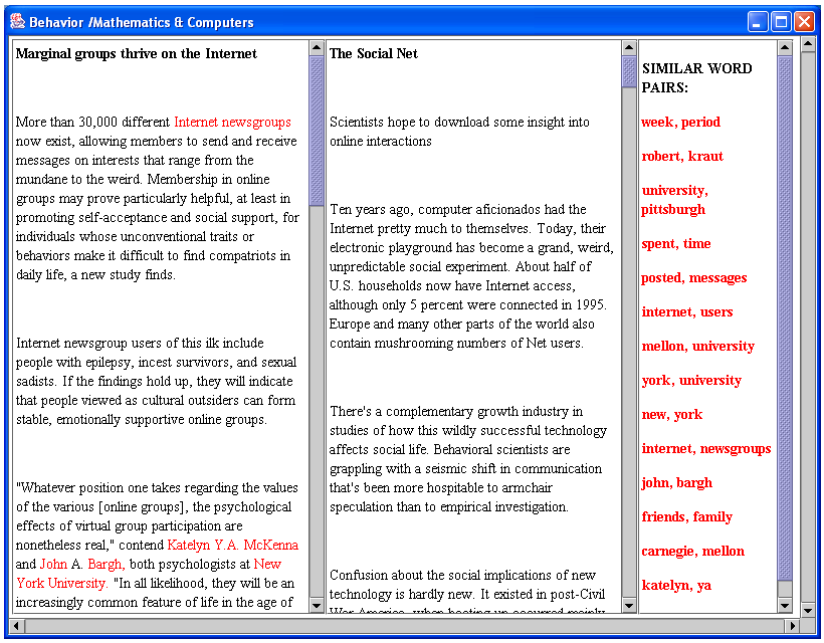

Figure 27: The BPM comparison file nodes 88 and 147 of the Behavior and Computer Science and Mathematics categories.

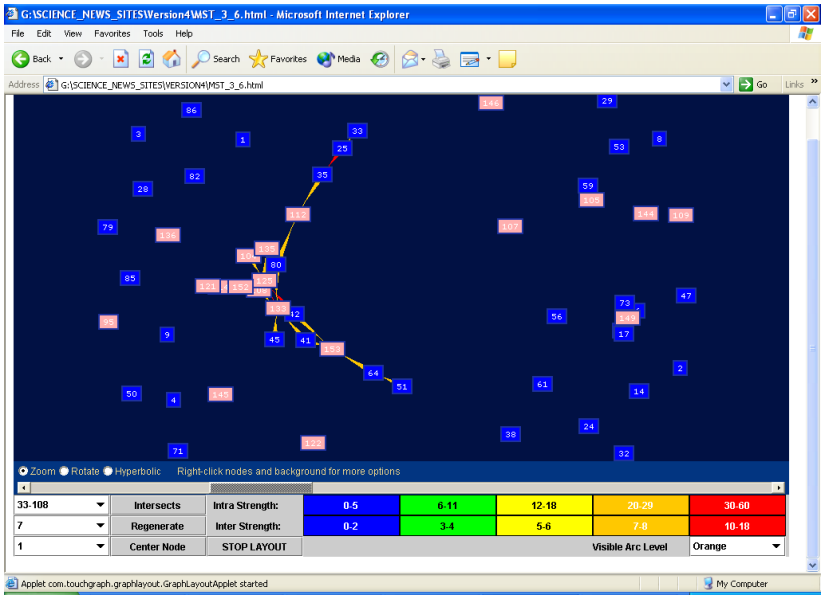

Figure 28: Those articles within 7 links of 88 and 147 with a strength of at least orange.

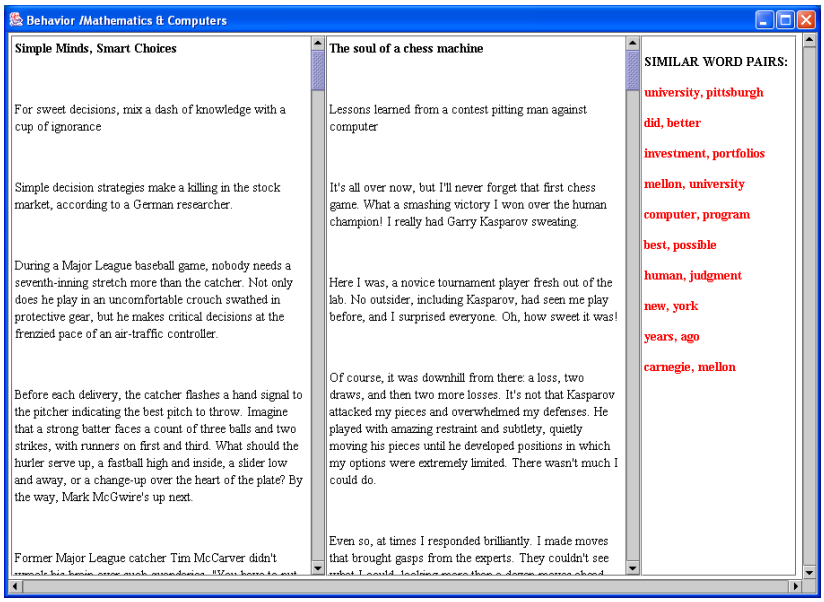

Figure 29: A closer look at the between article associations between article 80 a Behavior article and article 108 a Mathematics and Computer Science article.

There are 10 associations between this pair of articles however some of these are spurious. The set of non-spurious associations is small but relevant and includes (did, better), (investment, portfolios), (computer, program), (best, possible), and (human, judgment). When we originally formulated our research agenda in the automated serendipity program we thought of the field of genetic algorithms as one of those areas where an insightful researcher, numerous researchers claim credit in this case, discovered the relationship between the fields of evolutionary biology and mathematical optimization. We decided that whatever software that we developed would be "successful" if it could discover such an association. When we first encountered the above association we felt that perhaps we might have discovered such a "nugget." The first article, the Behavioral one, discusses the technique that humans use when they are faced with an overwhelming optimization problem for which they are presented with limited time to make a decision. The example provided in the article is a catcher in baseball. He must make a decision of the best pitch for the pitcher to throw while worrying about the runners stealing bases. He also has a limited time to make his decision. So one might wonder if the bounded rationality scheme might be relevant to formulating strategies for chess programs. Well one of the authors, JLS, has a little experience playing chess and one of his first instructions that his tutor offered him was that "human chess masters don't necessarily think as many moves deep as a computer but they have the ability to rapidly evaluate numerous positions that are several moves deep." So the instructor indicated that human beings employ bounded rationality. It is not much of a stretch to believe that it might be appropriate for a computer to employ bounded rationality in it's approach to chess. The potential applicability of the concept of bounded rationality to other areas is supported by a consideration of the comparison files between nodes (80, 133), see Figure 30. The comparison file between these two includes the word pairs (artificial, intelligence), (real, world), (world, situations), (computer, program), (best, possible), (decision, making), and (probability calcula-

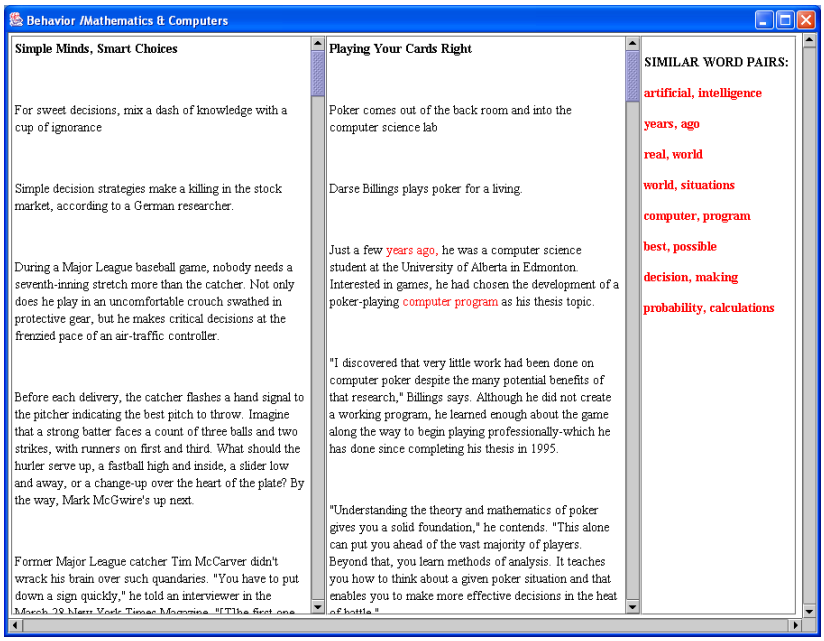

Figure 30: A closer look at the between article associations between article 80 a Behavior article and article 133 a Mathematics and Computer Science article.

tions). Once again we see that the article discussing bounded rationality has been linked to one discussing computer game strategy.

We may have saved some of the strongest evidence of the relevancy of our discovered association to the last. Dr. Stephen Lubard, the Director of the Office of Naval Research, made a presentation at the Dahlgren Division of the Naval Surface Warfare Center in May of 2004, [Lubard, 2004]. During this briefing he alluded to research underway aimed at applying the concepts of "bounded human rationality" to the Naval Command 21 Knowledge. We consider the fact that the Navy was already looking into the incorporation of bounded rationality into it information theoretic warfare strategies as the strongest endorsement of our auto serendipitous "discovery."

Clustering Results on the ONR ILIR Dataset

As indicated above we postponed our full treatment of the clustering software until now. We have chosen to cut the minimal spanning tree into subtrees at a particular edge based on a criteria determined by the strength of the associated edge divided by the mean strength of those edges connected to the nodes of the edge in question and of a path length of no more than k edges. In terms of implementation details the user first runs a JAVA program that creates the associated node lists based on a user specification of the maximum number of clusters, the minimum number of points per cluster, and the parameter k. Once this computation is performed the user is presented with the following opening screen, see Figure 31. The user can select how many clusters to break the data into and these clusters are presented to the user in a sequence of tab pane windows.

In Figure 32 we present an overview of the ILIR cluster structure along with the associated color map. A quick visual scan of the clusters does not reveal much

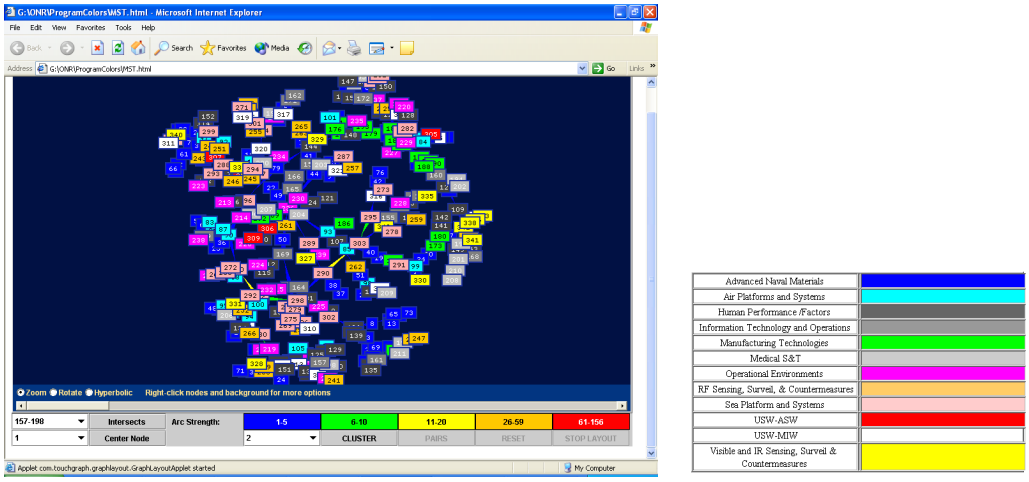

Figure 31: The opening screen for the clustering program when loaded with the ILIR data and associated color key.

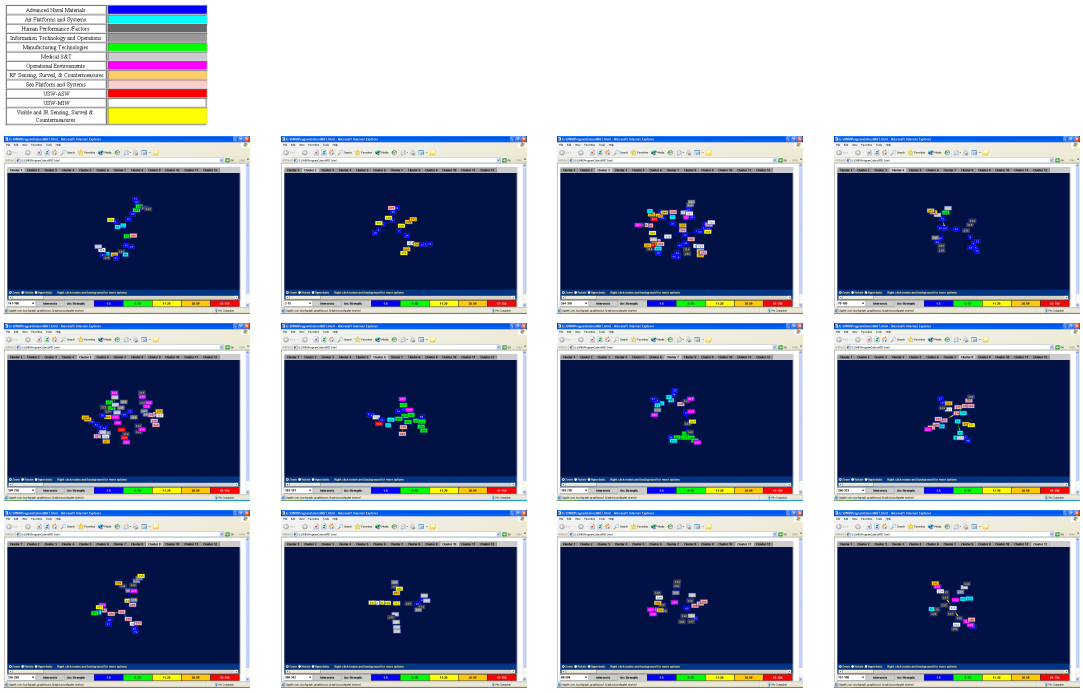

Figure 32: Overview of the ILIR cluster structure.

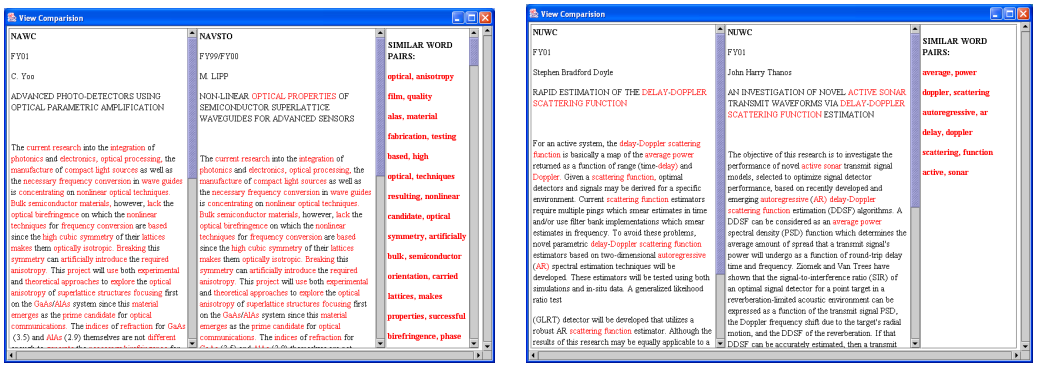

Figure 33: Two interesting article pairs from cluster 2 of the ILIR dataset.

uniformity in color among the clusters. We believe that this is in part because $S(X,Y) = 0 \rightarrow D(X,Y) = \sqrt{2}$ for many of node pairs. When we run our algorithm and ask for 8 cuts these $\sqrt{2}$ numerator terms dominate our decision as to where we wish to make our cuts and hence we effectively make 8 random cuts on the MST. It turns out that cutting the tree into 8 pieces does effectively help us to make a better exploration of the tree and we can identify a few other anomalies that were not evident in the tree originally.

Let's first explore cluster 2 a little, see Figure 33. The window on the left contains the BPM comparison file for two articles that contain virtually identical abstracts but yet represent projects that were executed by different investigators, Yoo and Lipp, at different labs, NAVSTO and NAWC, during different years, FY01 and FY99/00. The window on the right presents BPM comparison file for an abstract from the USW-MIW and RF Sensing, Surveillance, and Countermeasures categories.Both of these efforts were conducted during the same year, FY01, at the same lab NUWC, by different investigators. They seem to be closely related technologically and there clearly could have been fruitful interactions between the two investigators.

Let's next explore cluster 3 a little, see Figure 34. The window on the left contains the BPM comparison file for two articles that contain virtually identical abstracts representing work being executed at the same organization, NAVSTO, during the same year, FY99/FY00, with different authors and abstract titles. The window on the right presents the BPM comparison file for an abstract from the Advanced Naval Materials and Visible and IR Sensing, Surveillance and Countermeasures Categories. The two articles seem to be a continuing project for the investigator, S. Nee. It just so happens that the article title and our categorization of the article changed from FY99/FY00 to FY01.

There are at least two other cluster 3 article pairs that warrant additional discussion, see Figure 35. The window on the left presents an interesting association between an Advanced Naval Materials and Visible and IR Sensing, Surveillance and Countermeasures categories. These projects were executed by different investigators, Tran and Elson, at different labs, NAWV and NAVSTO, during different years, FY99/FY00 and FY01. The two projects are on finite difference time domain methods. Once again we can make a strong case for successful collaborations among the researchers. The window on the right presents the BPM comparison file for an abstract from the Sea Platform and Systems and the USW-ASW categories. These two articles represent projects that were executed at different labs, NUWC

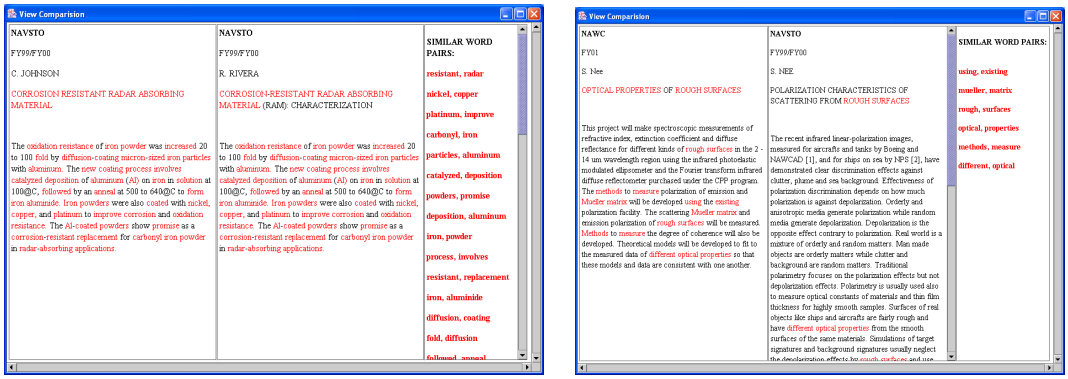

Figure 34: Two interesting article pairs from cluster 3 of the ILIR dataset.

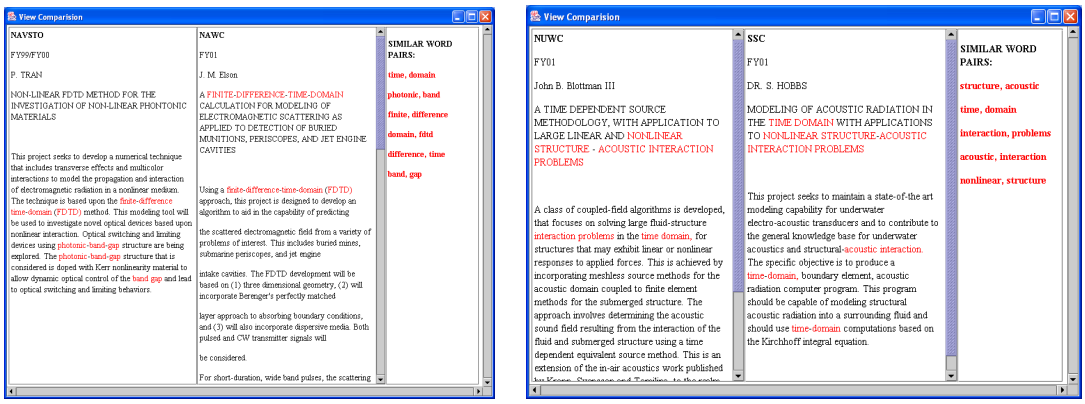

Figure 35: Two interesting article pairs from cluster 3 of the ILIR dataset.

and SSC, during the same year, FY01, by different investigators, Blottman III and Holmes. They both seem to be about nonlinear interactions in the acoustic domain. The commonality of topic might once again suggest fruitful interactions between the investigators.

Next we would like to point out some interesting cluster 5 article pairs, see Figure 36. The articles described in the left window describe two efforts at NSWCDD that were executed during FY01. Both of these articles come form the Sea Platform and Systems category. Both of these articles are on the application of Fourier-Kochin theory. One can only hope that the two investigators were in contact with one another during the execution of their respective projects. The article pair detailed in the right window both come from the Operational Environment category. There are many interesting associations between these two articles. The first one is entitled "Submarine officer of the deck knowledge elicitation and modeling," while the second article is entitled "Information requirements and information organization in submarine combat systems." The executor of the first project, Dr. Shobe, hopefully was aware of the work that had been previously performed by Dr. Carr.

We conclude this section with one last set of article pairs, see Figure 37. The article pair in the left figure, from cluster 7, represents an interesting association between an Advanced Naval Materials and Manufacturing Technologies article. The word pairs, (high, speed), (induced, reaction), (observed, high), (framing, camera),

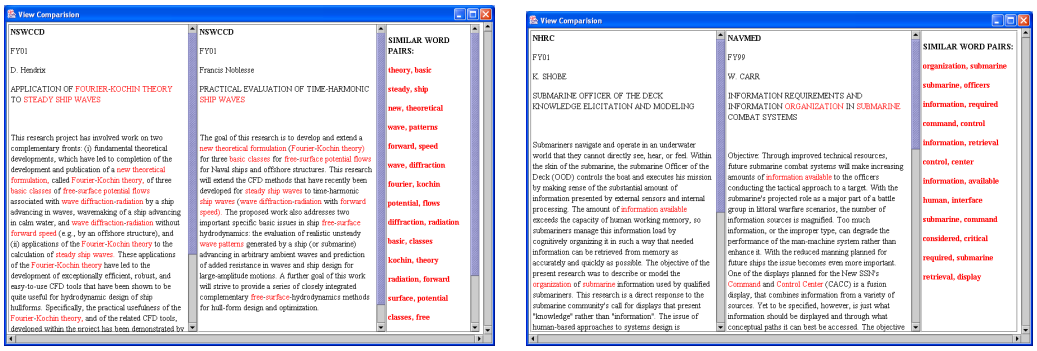

Figure 36: Some interesting article pairs as revealed in cluster 5 of the ILIR dataset.

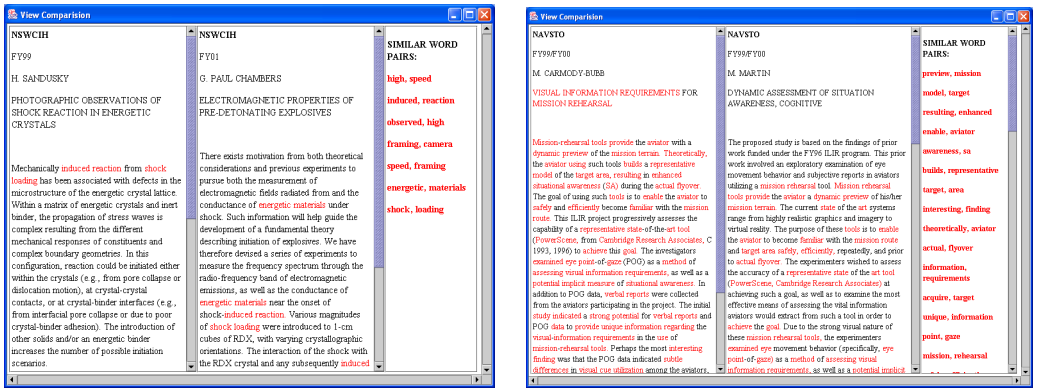

Figure 37: Some interesting article pairs as revealed in clusters 7, left figure, and 12, right figure, of the ILIR dataset.

(speed, framing), (energetic, materials), and (shock, loading) seem to indicate that both articles are studying fundamental properties of explosives. The article pair on the right, from cluster 12, presents a more interesting association. Both of the articles come from the Air Platforms and Systems category. Both of these projects were being executed during the same time frame, FY99/FY00 at the same lab, NAVSTO. It is surprising to see the very large number of associations between the two articles given the different titles. One would not have suspected such a strong association based on their titles, "Visual information requirements for mission rehearsal," and "Dynamic assessment of situation awareness, cognitive."

Clustering Results on the Science News Dataset

Next we turn our attention to the results obtained on the Science News dataset using the clustering program, see Figure 38. There seems to be a little more uniformity of color in the various clusters, as compared to the ILIR dataset. Let's take a moment to explore some of the substructure of some of the clusters. Let's first consider an interesting substructure in cluster 1, see Figure 39. This subcluster consisting of nodes 429, 483, 451, 509, and 543 is focused on animal behavior and sexuality.

Let's consider another interesting subcluster of cluster 1, see Figure 40. This subcluster consists of nodes 131, 47, 60, 102, 84, 167, 124, 105, 148, 132, 61. These nodes form the spokes of a wheel about the article 134, "Infrared camera goes the

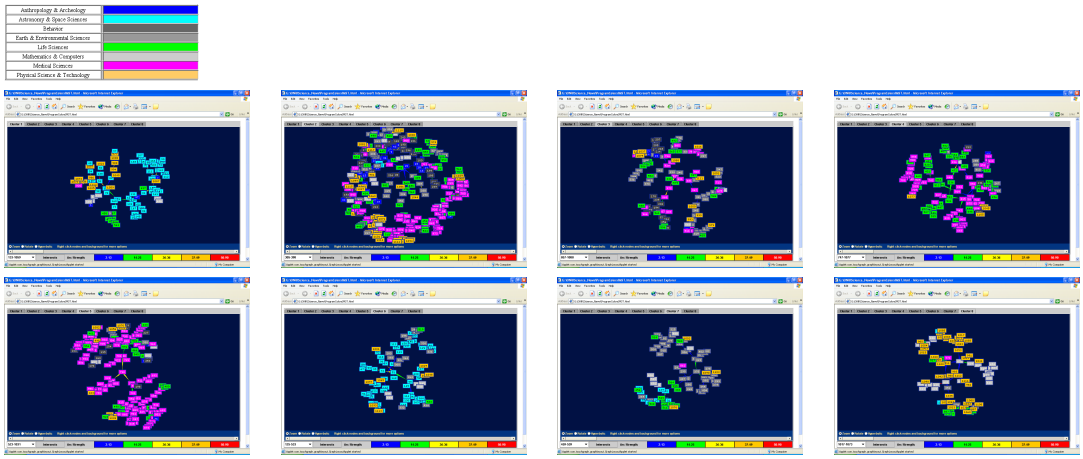

Figure 38: Overview of the Science News cluster structure.

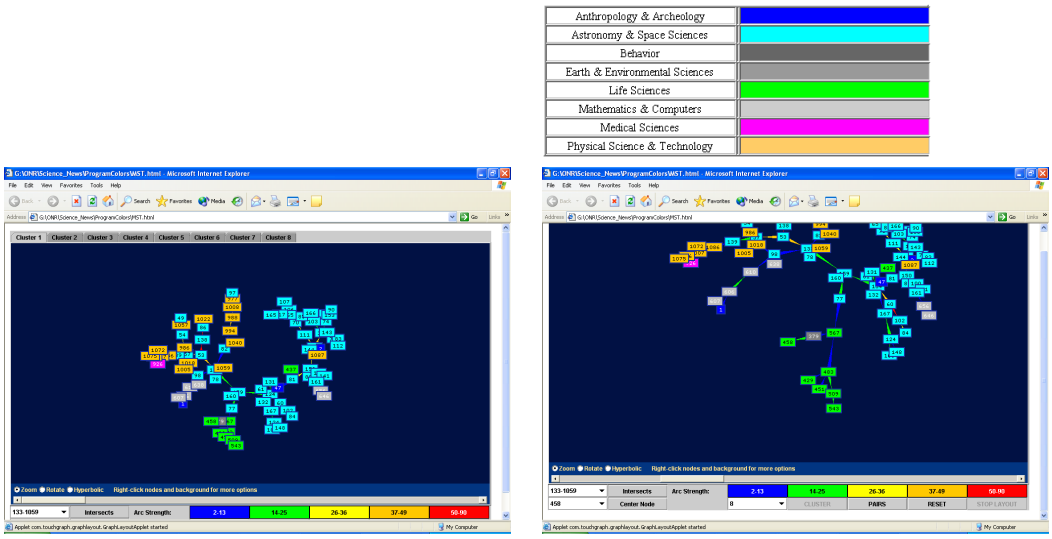

Figure 39: A subcluster of Science News cluster 1 that is focused on animal behavior and sexuality.

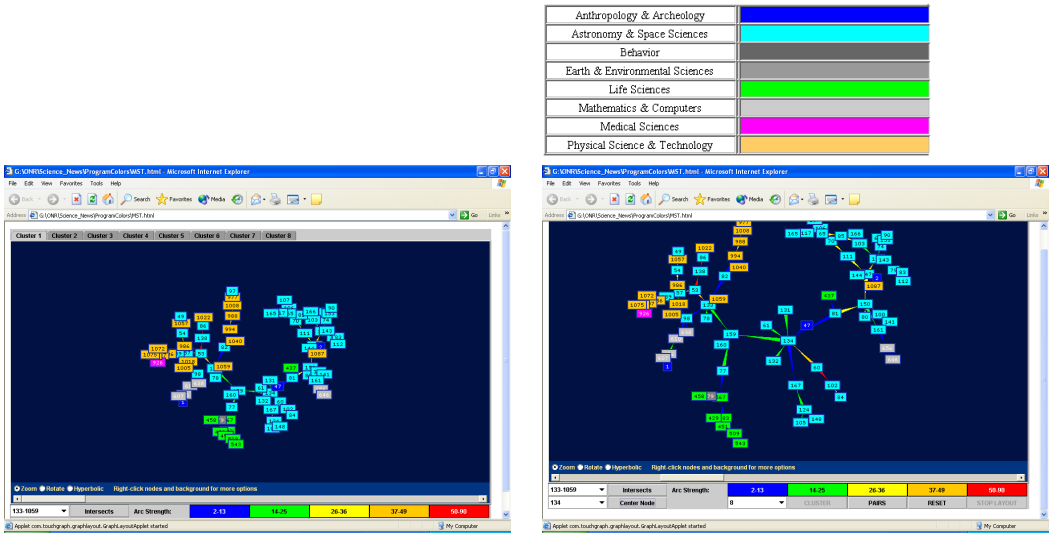

Figure 40: A subcluster of Science News cluster 1 that is focused on an infrared camera and its applications.

distance." All of these articles are about applications of infrared sensors to various problems in cosmology and other disciplines.

Let's next consider a subcluster of cluster 5, see Figure 41. This subcluster consists of nodes 933, 804, 855, 823, 911, 800, 710, 877, 755, 754, 956, and 1034 consists of articles about AIDs.

Next we will consider a subcluster of cluster 6, see Figure 42. This subcluster consists of nodes 62,114,130, 119, 332, 75,73,163,121,125,414, 533, and 164 consists of articles about solar activity.

Next we will consider a subcluster of cluster 7, see Figure 43. This subcluster consists of nodes 314, 508, 55, 71, 136, 171, 454, 269, and 116 consists of articles on the origins of life. We note that the node placement has been rendered using a slightly different solution to the spring equations than was presented in the cluster overview Figure 38.

Finally we will consider a subcluster of cluster 8, see Figure 44. This subcluster consists of nodes 618, 635, 643, 655, 621, 614, 645, 654, and 604 consists of articles on artificial intelligence (gray nodes). We note that the node placement has been rendered using a slightly different solution to the spring equations than was presented in the cluster overview Figure 38.

Conclusions

We have presented a new method for effective exploration of a labeled interpoint distance matrix. Our methodology supports the discovery of serendipitous relationships between the various categories encoded within the matrix. The methodology also support cluster structure exploration.

We have illustrated the application of our developed techniques with a Science News extended abstract dataset and with a short ONR ILIR abstract dataset consisting of roughly 1/2 paragraph length project abstracts. We have used these datasets to illustrate both the automated serendipitous discovery and the cluster

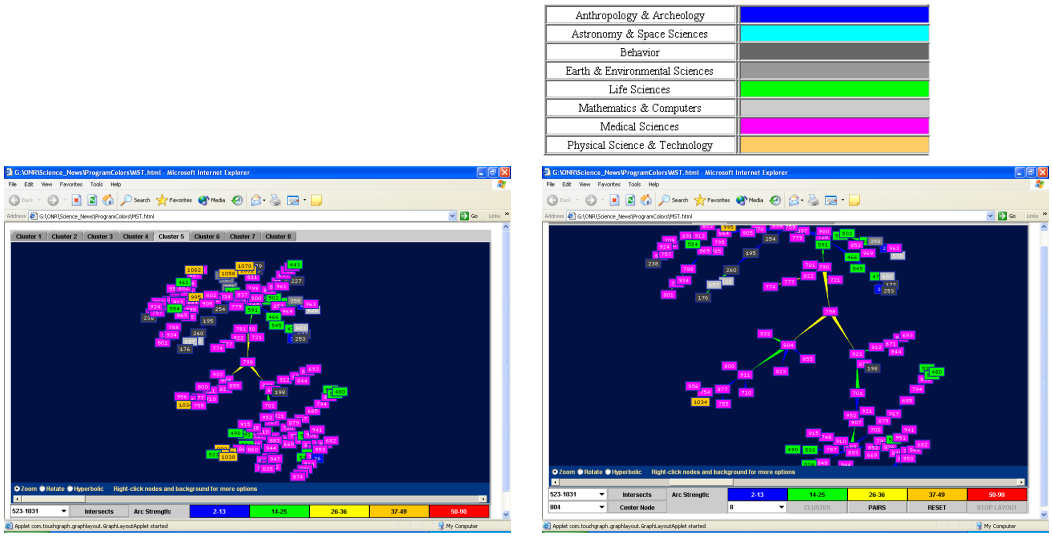

Figure 41: A subcluster of Science News cluster 5 that is associated with AIDs..

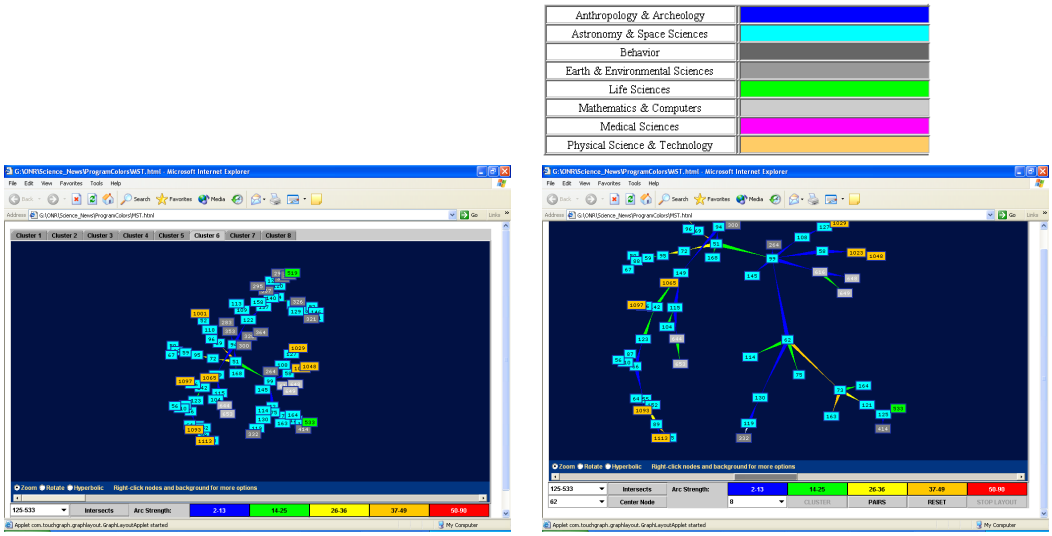

Figure 42: A subcluster of Science News cluster 6 that is associated with solar activity

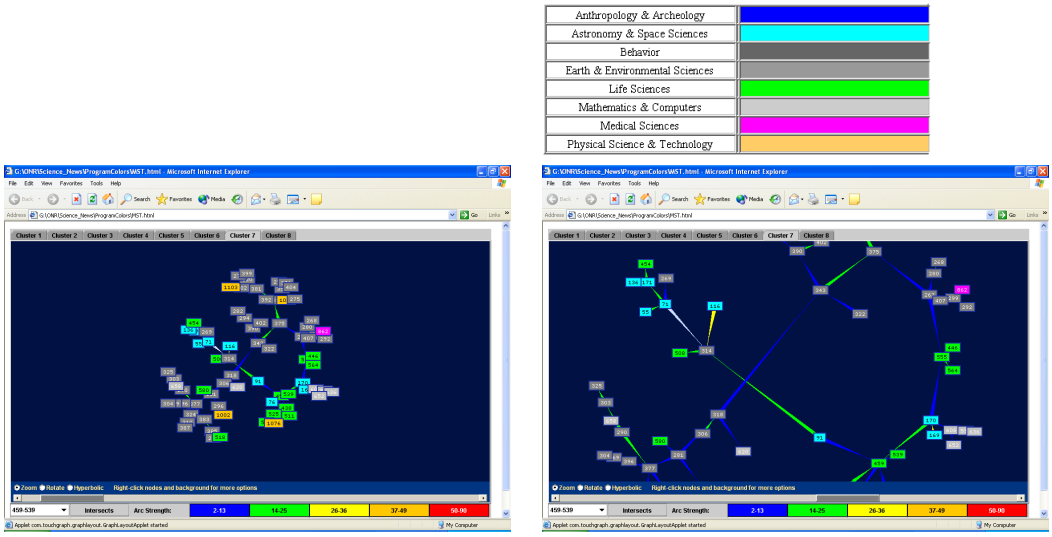

Figure 43: A subcluster of Science News cluster 7 that is associated with the origins of life.

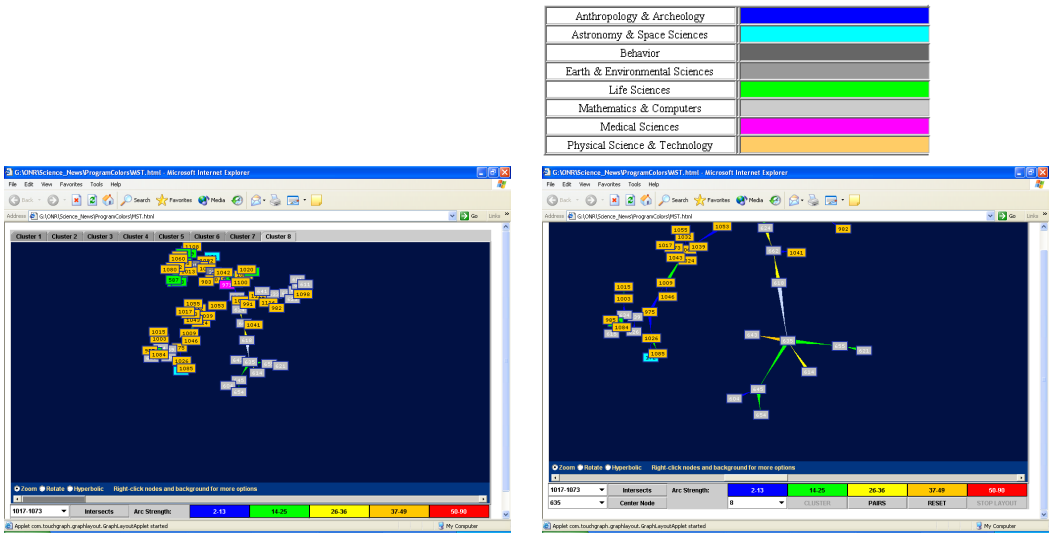

Figure 44: A subcluster of Science News cluster 8 that is associated with artificial intelligence.

exploration. We have discussed many examples using both datasets. We have shown one example, that we feel is particularly relevant, that links the fields of human bounded rationality and computer artificial intelligence for game theory. We have provided evidence to support this particular linkage via current Naval investment strategies in the application of human bounded rationality to shipboard decision making.

This work represents the tip of the iceberg of a new area that is not only of strategic importance to the United States but is also highly relevant to all who are currently conducting research in any discipline. The identification of effective research strategies is an area that is of paramount strategic importance as the number of possible beneficial connections that exists between various scientific disciplines increases and the amount of funding that can be allocated to these areas decreases.

Acknowledgments

The authors would like to acknowledge the support of the Defense Advanced Projects Research Agency. Some of the authors (JLS and ACB) would also like to acknowledge the support of the Office of Naval Research.

References

[Guo et al., 2002] Guo, D., Peuquet, D., and Gahegan, M. (2002). Opening the black box: Interactive hierarchical clustering for multivariate spatial patterns. In *Geographic Information Systems 2002*.

[Guo et al., 2003] Guo, D., Peuquet, D., and Gahegan, M. (2003). Iceage: Interactive clustering and exploration of large and high-dimensional geodata. *GeoInformatica*, 7(3):229–253.

[Lubard, 2004] Lubard, S. (2004). Briefing to the Naval Surface Warfare Center Dahlgren Division.

[Martinez, 2002] Martinez, A. R. (2002). *A Framework for the Representation of Semantics*. Computational sciences and informatics, George Mason University.

[Solka and Johannsen, 2002] Solka, J. L. and Johannsen, D. A. (2002+). Classifier optimizaztion via graph complexity measures. In *Proc. of the Army Conference on Applied Statistics 2002*.

[Xu et al., 2001] Xu, Y., Olman, V., and Xu, D. (2001). Minimum spanning trees for gene expression clustering. *Genome Informatics*, 12:24–33.

[Xu et al., 2002] Xu, Y., Olman, V., and Xu, D. (2002). Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics*, 18(4):536–545.