Identifying Cross Copora Document Associations Via 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

Abstract

This talk will focus on our recent work in the identification of related documents from different corpora or discipline areas. The purpose of this work is to help a researcher or program manager to identify fruitful cross-disciplinary research areas. The talk will discuss some preliminary results that have been obtained using a small Science News (around 1200 articles) dataset. The work is predicated on new visualization environment to facilitate the exploration of the multi-class interpoint distance matrix. This work is joint with the Algotek Team in general and Edward J. Wegman of GMU and Avory Bryant of NSWCDD specifically.

Keywords

cross corpora data mining, automated serendipity, minimal spanning trees, graph theory, visualization

Introduction

A current researcher or research coordinator is faced with an explosion of information with which they must have a modicum of familiarity in order to adequately evaluate the relevance/novelty of a proposed new research agenda or of results obtained as a product of a new research agenda. This task is made particularly difficult in that said idea or results may be discussed using different terminology by researchers in different academic/engineering disciplines. An easily understood example of this is the following. Suppose that a researcher in the statistics community creates a "new clustering" algorithm. This researcher is now interested in evaluating the novelty of his new approach. In order to effectively do this he needs to canvass not only the statistics literature but also the computer science, pattern recognition, and mathematics literature. In fact there has been novel clustering work by the chemistry and bioinformatics community. One can certainly imagine

that an identical clustering algorithm might be described in a language that is sufficiently different in various discipline areas as to majorly obfuscate their identical nature.

A somewhat easier problem, which is the focus of our current discussions, is when one is interested in locating two articles from different discipline areas that are really discussing the same topic or ultimately may be detailing similar approaches to the same problem. The ability to mine a collection of articles from different academic disciplines might allow a user to semi-automatically locate interesting or serendipitous relationships across scientific disciplines. A classic example of such a discovery is the relationship between mathematical optimization, a concept within the mathematical community, and the improvement of biological fitness, a biological concept. A cogent researcher or group of researchers, the story varies depending on who is telling the story, made the connection between these two ideas and the field of genetic algorithms was born. Ultimately we would like to facilitate the discovery of such connections by human operators.

There are numerous steps that are necessary in order to facilitate this process. First one must be able to encode the semantic content of the document. This is necessary in order that the computer may be able to proceed forward with additional calculation. Given a semantic description of the documents one then needs a way to compute similarity or distance values between the documents. This produces an interpoint distance matrix that describes the relationship between the observations. It is important to note that we are really assuming that we have some previously obtained categorization of the documents that are also going to be part of the analysis process. This distinction is important in that the serendipitous discovery process can be described as seeking articles across discipline areas. The pre-categorization of the articles could be obtained from a human or could be obtained in a semi-automated manner via another algorithm.

At this point in time one has a labeled interpoint distance matrix. Subsequent analysis is aimed at allowing the human user to identify interesting relationships between the articles in the various discipline areas. Any approach to this should allow the human to not only make an initial identification of related articles but should also allow the human to drill down to ascertain the exact relationship between the articles.

Approach

Our developed system uses 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. 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 said 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.

Given an encoding of the documents as BPMs one needs a method to compute

distances between these documents. Martinez compared roughly 15 different ways of calculating distances between the BPMs. He studied the advantages of each of these approaches. Some are predicated on first computing a similarity measure between the documents and then converting this similarity measure into a distance. As discussed in the results section we proceed forward using a similarity measure that is easily transformed into a distance. Our choice for which similarity/distance method to use was based on the methodology that performed well in the dimensionality reduction/clustering work done by Martinez.

So now we have a labeled interpoint distance matrix. The first thing that one could do is merely find the closest documents residing on either side of a discriminant boundary that separates the corpora where we have chosen two corpora at a time. One might posit the fact that these article pairs would represent good candidates for serendipitous relationships. This approach was initially implemented and it did provide some interesting results which will not be described here. We were interested however in being able to extend this concept in order to discover other serendipitous relationships between articles.

Our previous work [Solka and Johannsen, 2002] had demonstrated how one could employ a minimal spanning tree computed on the interpoint distance matrix in order to characterize classifier complexity. In this case one computes the minimal spanning tree on the interpoint distance matrix that has been computed on the two class data. One then uses the number of minimal spanning tree edges that cross between disparate classes as a surrogate for classification complexity. We realized that these edges between disparate classes are exactly the edges that might indicate a serendipitous relationship between observations in the disparate classes. The minimal spanning tree captures the relationship of the observations to the discriminant boundary. We can look for serendipitous relationships among the observations based on their relationship to the discriminant boundary.

The last piece of the puzzle is how one lays out the minimal spanning tree in the plane. One can utilize the minimal spanning tree in order to obtain an ordering of the observations that preserves the spatial relationship of the observations. We currently have not employed this. Currently we layout the minimal spanning tree in the plane using a simple spring based model [Tollis et al., 1998]. This approach as described in the results section has been shown to work well and to facilitate the discovery of interesting serendipitous relationships between the articles from the disparate corpora. We now turn our discussions to a brief description of the implemented system.

Figure 1 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 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.

Let us continue with our exploration of the tool through the examination of one particular corpora comparison case. We now focus our attention on the comparison of the article in the anthropology/archeology corpora with those in behavior. Figure 2 presents the minimal spanning tree exploration screen which results from clicking on the link provided to study this association on the main page.

Figure 3 presents the document comparison screen. The left-most window presents the article from the anthropology/archeology corpus, the middle window the document from the behavior corpus, and the right-most window a list of word pairs that

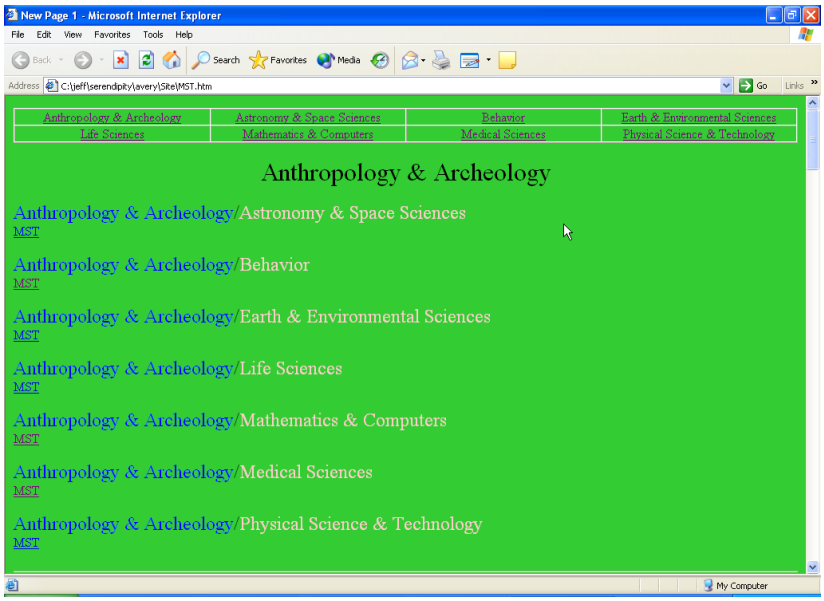

Figure 1: Opening screen of the automated serendipity search engine.

are in common to the two documents bigram proximity matrices. These word pairs have been rendered in red in the documents that are presented in the left-most and center windows. It is important to remember that the calculation of the documents' BPMs' occur after stopper word removal. In this manner two words that are red in the document windows may not actually be proximate to one another due to the removal of these stopper words.

Results

Now that we have briefly outlined the visualization framework, let's examine it's application to a text multi-copora document set. The document set that will be discussed consists of 1117 documents that were published by Science News 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, Avory Bryant. 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).

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 was then computed on each document pair using

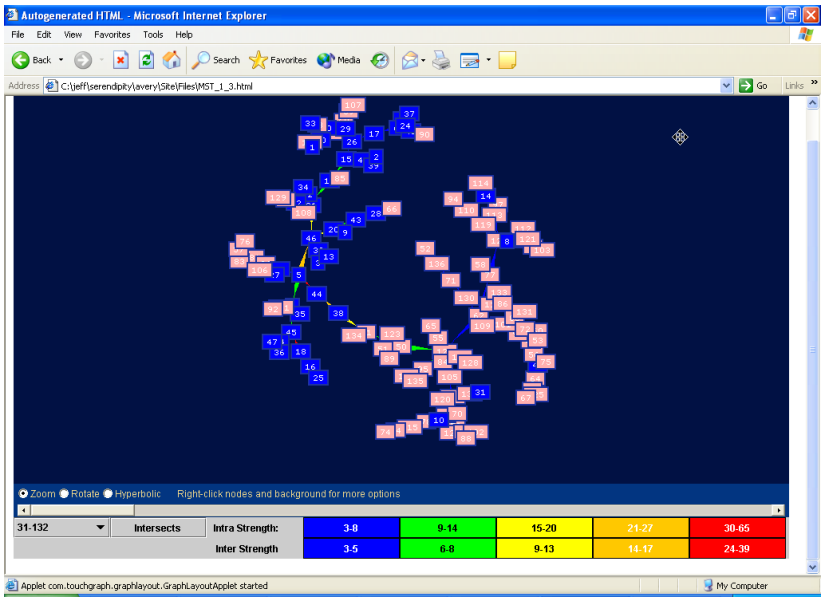

Figure 2: Minimal spanning tree exploration screen for the comparison of the anthropology/archeology and behavior corpora.

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

This similarity is then converted into a distance using

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

At this point in time we now have an interpoint distance matrix between all of the articles in the Science News collection. This interpoint distance matrix defines a complete graph where the set of vertexes correspond to the articles in the collection and the set of edge lengths are determined by the distances. As described above the minimal spanning tree was computed for each choose two corpora subgraph of the full corpora graph.

We will now discuss a few preliminary interesting cross corpora associations that we have discovered utilizing our tool. First let us explore relationships between the astronomy corpus and the physical sciences corpus. We will illustrate the exploration process with these two categories first in that one would expect that the two disciplines share common terminology. Figure 4 shows a particular branch of the minimal spanning tree where a strong cross corpora association is indicated between documents 49 and 128. This article pair represents the strongest association pair between the two corpora with 47 similar word pairs between the two documents. This number represents the numerator in the similarity calculation.

Figure 5 presents the comparisons file for this particular test case. The astronomy article which appears on the left is about manufacturing artificial black holes. The physical sciences and technology article which appears on the right is about modern cosmology theories. The two articles are very closely related and one can easily imagine a situation in which a different reviewer may have placed these two articles in the same category.

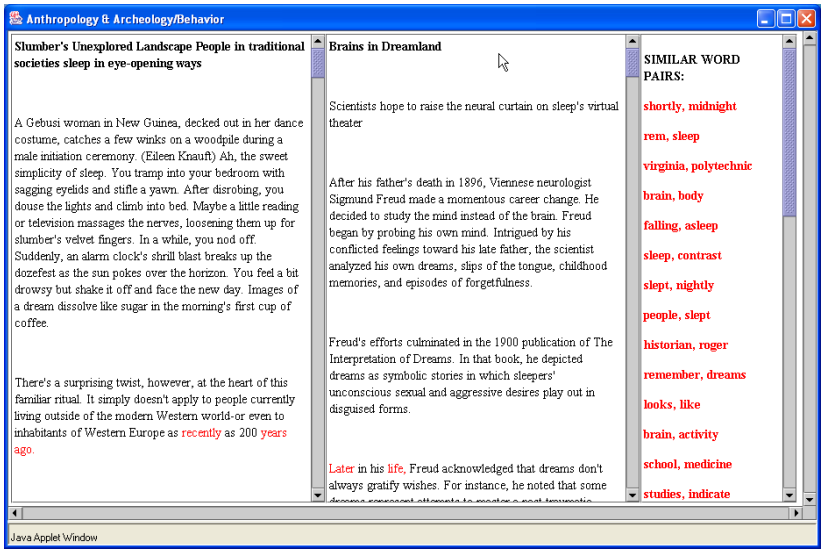

Figure 3: Document comparison screen for a two highly associated anthropology/archeology and behavior documents.

The final example that we will consider is based on a comparison of the earth and environmental sciences corpora to the medical sciences one. Figure 6 shows a particular branch of the earth and environmental sciences vs. medical sciences minimal spanning tree. In this case we are focusing our attention on node 137 (earth and environmental sciences) and node 197 (medical sciences). These two articles contain 32 common associations.

Figure 7 presents the comparison screen for the two articles. The first article is about the effects of phtlate plastics on the development of masculinity traits in animals while the medical article discusses similar studies in humans. A research coordinator or researcher would certainly be interested in the commonality of the findings in the animal studies and the human studies.

Conclusions

We have presented, within this paper a new multi-corpora text exploration tool that allows a human operator to explore the relationship between documents residing in disparate corpora. This system has currently been tested on a 1117 document 8 copora set of roughly 1/2 to 1 page articles extracted from the Science News website. Our preliminary analysis indicates that a trained human operator can successfully find meaningful associations between the corpora.

There are numerous ways to improve upon or extend this "first generation" system. One of the ideas that we are interested in pursuing is the use of the minimal spanning tree to aid in the spatial layout of the nodes. Another important upgrade would allow the user to specify a subset of nodes and then provide the user with a new view of the network that focuses attention on this subset of nodes and allows the user to make a more in-depth exploration of these nodes. Finally it might make sense to provide for some capability to use the minimal spanning tree to cluster the data. This might allow the user to discover interesting cross corpora subtopics.

Acknowledgments

The authors would like to acknowledge the support of the Defense Advanced Projects Research Agency.

References

[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*.

[Tollis et al., 1998] Tollis, I. G., Battista, G. D., Eades, P., and Tamassia, R. (1998). *Graph Drawing:Algorithms for the Visualization of Graphs*. Prentice Hall.

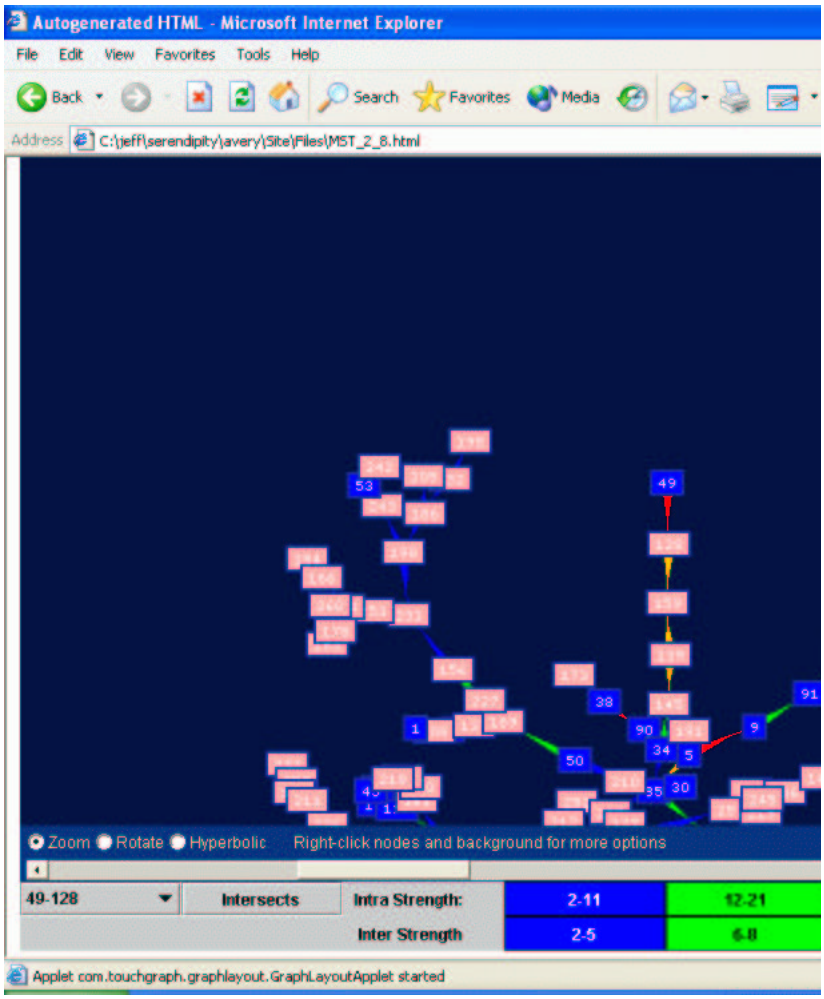

Figure 4: Minimal spanning tree branch containing the two documents, 49 (physical sciences) and 128 (astronomy), with a strong cross copora association.

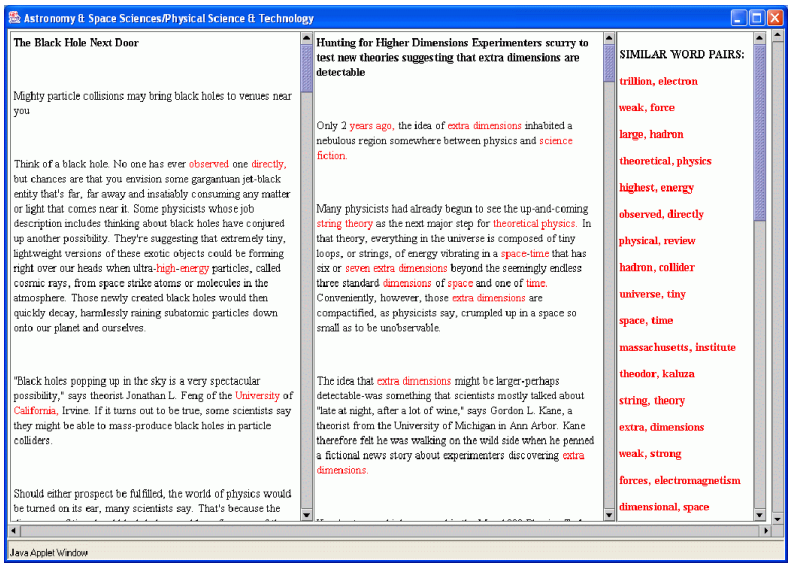

Figure 5: Comparison files for the two documents 49 (physical sciences) and 128 (astronomy). The long list of common associations suggests a strong relationship between these two articles.

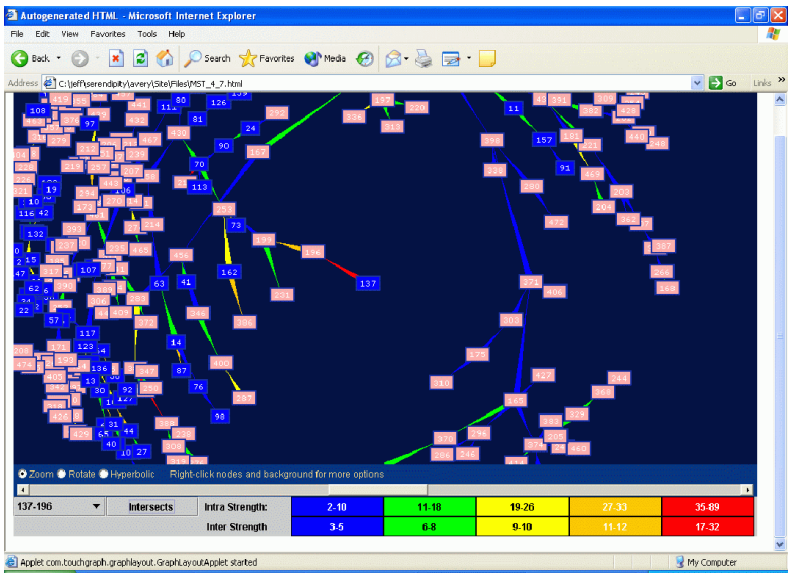

Figure 6: Minimal spanning tree branch containing two documents,137 (earth and environmental sciences) and 197 (medical sciences), with a strong cross copora association.

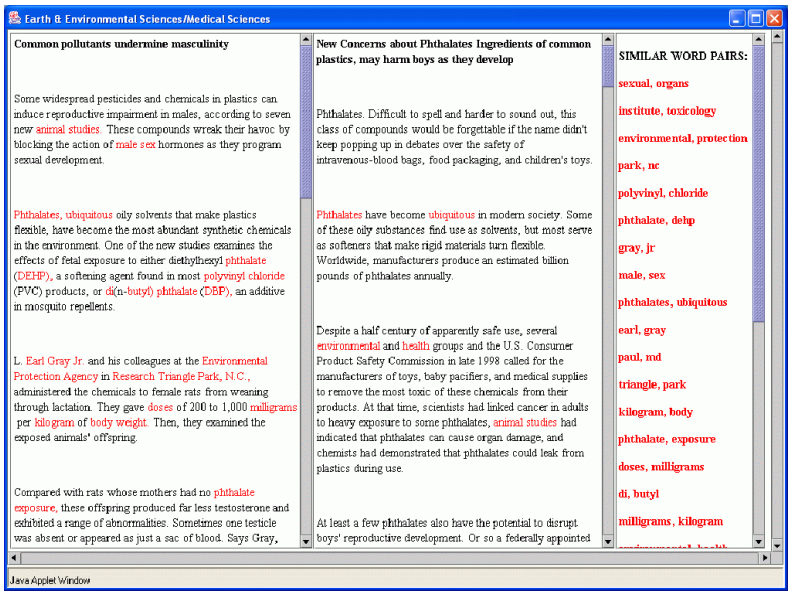

Figure 7: Comparison files for the two documents 137 (earth and environmental sciences) and 196 (medical sciences). The long list of common associations suggests a strong relationship between these two articles.