Point Placement by Phylogenetic Trees and its Application to Visual Analysis of Document Collections

Ana M. Cuadros* Fernando V. Paulovich† Rosane Minghim‡ Guilherme P. Telles§

Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, Brazil

ABSTRACT

The task of building effective representations to visualize and explore collections with moderate to large number of documents is hard. It depends on the evaluation of some distance measure among texts and also on the representation of such relationships in bi-dimensional spaces. In this paper we introduce an alternative approach for building visual maps of documents based on their content similarity, through reconstruction of phylogenetic trees. The tree is capable of representing relationships that allows the user to quickly recover information detected by the similarity metric. For a variety of text collections of different natures we show that we can achieve improved exploration capability and more clear visualization of relationships amongst documents.

Keywords: Document Visualization, Multidimensional Visualization, Document Analysis, Text Analytics, Phylogenetic Trees.

Index Terms: I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction techniques; I.7.m [Document and text Processing]: Miscellaneous

1 INTRODUCTION

In this paper we contribute to the visual exploration of document collections. We approach the problem with an alternative technique for constructing visual maps of documents. By documents we mean raw text files or files that can be converted to raw text, such as electronic documents, files resulting from Internet searches, emails, or even database dumps.

The problem we investigate can be stated as: given a collection ranging from hundreds to a few thousands text documents, generate a 2D map that reflects the content relationship among those documents, that is, documents with similar content should be clearly related on the map. We also require that the map is amenable to interaction and exploration.

The most common process adopted for generating a content-based document map includes the evaluation of a similarity measure for the texts followed by a point placement strategy, displayed with appropriate visual attributes. Interaction for exploration follows.

In the first step, some distance measure is evaluated for pairs of texts, either directly [27] or by representing every text as a vector and operating on a vector space [24]. Due to the difficult nature of distance evaluation for text, the distances rarely define a metric space. Thus it is not straightforward to place the texts as points in 2D, except for trivially small document sets. To overcome this difficulty, various approaches have been tried for placement of points in 2D. The most widely used is to consider documents as points in a high dimensional space and to project them onto 2D. Projection techniques perform well, but have problems. Techniques designed to obtain scalability make decisions on neighborhoods (clustering for instance) without a complete analysis of the distance relationships. In most high precision techniques some points are always misplaced, at a rate that grows with the size of the collection. Additionally, closely related documents tend to group together to a degree that impairs distinction of the individual points or of the density of groups.

In this work we introduce an alternative approach for the construction of document maps targeted at reflecting well similarity relationships. We create a map directly from a distance matrix by building a "phylogenetic tree of texts". The problem of phylogenetic tree reconstruction is one of inferring ancestors for a group of species, reconstructing its evolutionary history. By replacing species with texts and using a well known heuristic for tree construction, we build an ancestry relationships from higher to lower content correlation. The main advantage of the approach are improved exploration and more clear visualization of similarity relationships. Other data types that lend themselves to similarity calculations can also benefit from this approach.

In the next section we review the main literature related to mapping textual document collections visually based on content. We also present the basic concepts related to our approach, that is, the use of projections for mapping multidimensional data and phylogenetic trees reconstruction. The full description of the process to generate a text map is presented in Section 3. Section 4 presents results of various tests for the phylogenetic tree maps and compare them to displays by high precision projections. That is followed by the the analysis of the contribution of this work, that is, a novel highly meaningful visual technique to reflect content relationships for the exploration of collections of documents and other multidimensional data.

2 RELATED WORK

The strategic task of visual analysis of text collections has various peculiarities that impose a large challenge for the scientific community. There are still many unsolved problems and the effectiveness of solutions depends on the collections, on the questions to be answered, on the individual documents, on the degree of structure that one can count on, amongst many others. Various approaches for visual display of documents exist, and many of them rely on information on the sources of the data and on the texts themselves. For instance, it is common to visualize document collections through networks of citation, co-citation, co-authoring and keywords. A few of the techniques available can build maps of document collections directly from their contents. For a nice classification with description of the various approaches to visual display of document collections we refer to the work of Börner et al. [3].

We focus on content-based displays, that is, layouts that are meant to help locating correlation of content amongst documents, such that the user can gain a good notion of the concepts approached by a text data set and to locate interesting material without actually having to read a lot more than necessary. Many known techniques for data visualization have been investigated to achieve such a goal.

Multidimensional projections have been used to map document

*e-mail: anamaria@icmc.usp.br
†e-mail: paulovic@icmc.usp.br
‡e-mail: rminghim@icmc.usp.br
§e-mail: gpt@icmc.usp.br

IEEE Symposium on Visual Analytics Science and Technology 2007
October 30 - November 1, Sacramento, CA, USA
978-1-4244-1659-2/07/$25.00 ©2007 IEEE

points onto the 2D screen space. Several classical dimensionality reduction techniques can be employed for this purpose – and indeed they are, in a variety of techniques targeted at the displaying general multidimensional data. These include Self-organizing Maps (SOMs), analysis of principal components (e.g., Principal Component Analysis and Latent Semantic Index) as well as Multidimensional Scaling and variations targeted at optimizing performance, such as the Force Directed Placement with stochastic sampling [5], a technique turned much faster as well as multi-scale lately [13, 18]. The basic principles behind this type of layout algorithms are reviewed in Section 2.1.

All these approaches have the ability to map multi-dimensional data of various kinds. Evaluations for text collections have been performed with various degrees of effectiveness. Two typical layout approaches for similarity-based text visual display are galaxies [28, 29], implemented in IN-SPIRE^{TM} [28], now a commercial text visualization system [19]; and Infosky [1]. Both display textual documents from a collection as points in 2D space by attempting to place points close together when they are found to be close by the similarity measure. Clustering and projection are used to place groups of text in regions of the display and to reduce the number of distance calculations necessary to process the data set. Infosky also creates a hierarchical display by embedding structure into the similarity relationships, so the user can focus in and out in an analogy to a telescope.

The above mentioned systems for document mapping have been devised (and improved) to handle massive amounts of documents and to produce an initial global display that can be explored for focus on particular subgroups. However, a consequence of the approach usually caused by pre-clustering or selection heuristics, a considerable number of documents that should, from the user's perspective, be placed together, get placed in different groups. Additionally, dimensionality reduction in heterogeneous text sets gather different groups in overlapping regions during placement. For performance reasons, a decision is made too early in the process as to what the neighborhood of a particular document is, that is, identification of relationships intra- and inter-groups is impaired and information that may have been reflected by the similarity measure is lost during preprocessing steps. These are good tools to organize massive data sets, but under the requirements of some investigations, they could use a counterpart capable of identifying content relationship with higher precision. This is the type of requirement that can be provided by techniques such as the one put forward in this article.

In this and in our previous work we have taken as goal to find a point placement in 2D dictated by similarity in a global level, so that applications that need to examine considerable amounts of documents can be benefited by finding out general themes as well as local content relationships using the same display. Typical applications are those of research, investigation and examination of topics in an exploratory environment.

2.1 Point placement via multidimensional projections

Data sources have increased substantially both in size and complexity, and extracting useful information from them is still a challenge. One measure of data complexity is the number of attributes associated to each instance of data. Consider, for example, data from a demographic census: a data instance records attributes such as age, sex, education, occupation, income, and so forth. Considering each data attribute as a data dimension, if we have m such attributes then each data instance can be interpreted as a m-dimensional vector placed in a m-dimensional space.

A common way to handle dimensionality is to reduce the number of dimensions, so that strategies that are known to work well with low-dimensional data can be applied. *Multidimensional Projection* techniques are one example of such a strategy. A multidimensional projection technique typically maps the data into a p-dimensional space with $p = \{1,2,3\}$, whilst retaining, on the projected space, some information about distance relationships among the data items in their original definition space. In this way, a graphical representation can be created to take advantage of the human visual ability to recognize structures or patterns based on similarity, such as clusters of elements.

Formally, let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of m-dimensional data, with a dissimilarity distance measure $\delta(x_i, x_j)$ between two m-dimensional data instances, and let $Y = \{y_1, y_2, \ldots, y_n\}$ be a set of points into a p-dimensional space, with $p = \{1,2,3\}$ and with Euclidean distance $d(y_i, y_j)$ between two points of the projected space. A multidimensional projection technique can be described as a injective function $f : X \rightarrow Y$ that seeks to make $|\delta(x_i, x_j) - d(f(x_i), f(x_j))|$ as close to zero as possible, $\forall x_i, x_j \in X$ [26].

Multidimensional projection techniques can be split into two major groups, according to the functions f employed: *linear projection techniques* and *nonlinear projection techniques*.

Linear projection techniques create linear combinations of the data attributes, defining them in a new orthogonal basis of small dimension. A widely known linear technique is *Principal Component Analysis* (PCA) or *Karhunen-Loève Expansion* [12]. PCA is a second-order technique, that is, it employs information embedded in the covariance matrix of the data[1]. Second-order techniques are particularly suitable for data presenting Gaussian distributions, since in this case it captures almost all data distribution [16].

Although linear techniques perform well on Gaussian data, in handling data with nonlinear relationships such techniques typically fail to capture the relevant patterns. In such cases, nonlinear techniques are better candidates. Rather than relying on linear combinations of the attributes, nonlinear techniques attempt to minimize a function of the information loss incurred in the projection. Normally, this function is based on the dissimilarities amongst the m-dimensional instances and on distances among the p-dimensional points. Hence, it does not require representing the original data as vectors, it is sufficient to have a mechanism to measure instance dissimilarity in the original space.

One example of nonlinear projection technique is *Multidimensional Scaling* (MDS) [7]. Originating from the psychophysics domain, MDS actually comprises a class of techniques aimed at mapping instances belonging to a m-dimensional space into instances on a p-dimensional space ($p \leq m$), striving to maintain some distance relations.

Amongst the various MDS techniques, the simplest ones are those based on *Force-Directed Placement* (FDP) [10]. Originally proposed as a graph drawing heuristic, the FDP model aims at bringing a system composed by instances connected by imaginary springs into an equilibrium state. Instances are initially placed randomly and the spring forces iteratively push and pull them until reaching equilibrium.

In the general case, where each instance is connected to any other instance, the iteration of the FDP model takes time $O(n^2)$ for n points. Once it is necessary at least n iterations in order to reach the equilibrium state, the FDP model takes $O(n^3)$. Aiming at reducing this complexity, Chalmers [5] presented a technique, based on data samples, with linear iterations. Even though it reduces the complexity, once n iterations are necessary to create a stable layout the final complexity is still expensive ($O(n^2)$). Approximating the results obtained using this technique, Morrison et al. [18] presented an approach which defines an $O(n^{\frac{3}{2}})$ FDP model. On this approach a random sample of instances is first projected using Chalmers' technique. After that, the remaining instances are interpolated on

[1] For m attributes, a covariance matrix is a $C_{m \times m}$ matrix whose elements c_{ij} denote the covariance between data attributes i and j, which indicate the degree of linear relation between those two attributes.

the final layout. Improving the interpolation process, Morrison et al. [17] created an O($n^{\frac{5}{4}}$) model, and Jourdan and Melancon [13] suggested another approach to reduce this complexity to O($n \log n$), obtaining layouts of comparable quality.

We also have addressed the problem of creating useful multidimensional projection techniques. In our previous work we developed and tested two different techniques, namely *Projection by Clustering* (ProjClus) [20] and *Least-Square Projection* (LSP) [21]. Both were successfully employed for creating document maps, being valuable tools to help users to extract relevant information from document collections.

The LSP technique, employed here for comparison with the new layout proposed, is a generalization of an approach for mesh-recovering and mesh-editing in order to deal with high dimensional spaces. In this technique, a subset of high dimensional points is projected onto the plane, and the remaining points are projected using an interpolation strategy that considers only the neighborhood from the high dimensional points in the original domain. We refer to [21] for further details of LSP.

2.2 Phylogenetic reconstruction

Phylogenetic reconstruction is the biological problem of building a tree that reflects evolutionary relationships. Every leaf in a tree represents a species (or family or individual or other taxonomic unit). Internal nodes represent hypothetical ancestors. A phylogenetic tree may be rooted or unrooted, and it may represent only ancestry relationships or also represent evolutionary distance, coded as edges' weights or lengths.

There are two types of input for the problem: a character matrix that stores m character values for n species or a $n \times n$ distance matrix for n species. For characters, the problem is solvable in polynomial-time if the number of characters is limited or in the absence of reversals and convergence, which is rarely the case. For distances the problem is polynomial time solvable if the distances form a metric space. This is also rare. Other situations lead to NP-hard problems [25].

While working with documents we are interested in the problem for distance matrices. With texts, too, we rarely have a metric space. The leaves in a tree for a text collection are going to represent the documents. The internal nodes represent "ancestor" texts. We want a tree where edges' lengths represent distance between texts.

As mentioned before, constructing phylogenetic trees when distances are not a metric is hard. We rely on the widely known neighbor-joining (NJ) [23] heuristic algorithm for tree construction. NJ builds an unrooted tree, greedily selecting the closest pair of species and joining them by a hypothetical ancestor in the tree.

Suppose there are n objects that should be related through a phylogenetic tree, and that the distances D_{ij} for every pair (i, j) are known. Objects may be biological entities but are texts in the application we envision here. NJ starts with a star-like tree, with n leaves connected to a single internal node. At every step, the algorithm: (1) selects a pair of nodes (i, j) with the smallest sum of branch lengths S_{ij}; (2) adds a node x to the tree, with i and j as children and connected to the the common ancestor of i and j; (3) evaluates the branch lengths L_{ix} and L_{jx}; and (4) replaces i and j by x in the distance matrix, evaluating D_{xy} for every y in the matrix. This step is repeated while there are more than two nodes in the matrix. As a final step the branch length for the two last nodes is evaluated. The equations are listed below. The overall running time is $O(n^3)$.

$$S_{ij} = \frac{1}{2(n-2)} \sum_{k \neq i,j} (D_{ik} + D_{jk}) + \frac{D_{ij}}{2} + \frac{1}{N-2} \sum_{\substack{k<l \\ k,l \neq i,j}} D_{kl}$$

$$L_{ix} = \frac{D_{ij} + \frac{\sum_{k \neq j} D_{ik}}{n-2} - \frac{\sum_{k \neq i} D_{jk}}{n-2}}{2} \qquad L_{jx} = \frac{D_{ij} + \frac{\sum_{k \neq i} D_{jk}}{n-2} - \frac{\sum_{k \neq j} D_{ik}}{n-2}}{2}$$

$$D_{xy} = \frac{D_{iy} + D_{jy}}{2}$$

3 Process to create the document map

The process to create and explore visual representations of document collections has four steps and can be summarized as follows.

1. build a triangular matrix with text distances.
2. build a tree using NJ.
3. plot the tree.
4. explore the tree.

In the first step a distance matrix with the distances among the documents is computed. This matrix is the basis for the execution of the NJ algorithm that creates a phylogenetic tree (Section 2.2).

In our experiments we used two different approaches to evaluate the distances: the document vector representation and the NCD. In the former, documents are represented as vectors on a multidimensional space, and the distances are defined as the distances between vectors. Usually the steps to create the vector representation are: (1) the stopwords are eliminated from the documents; (2) stemming is applied to extract word radicals, for instance, using the *Porter's* stemming algorithm [22]; (3) a frequency count is performed, and the Luhn's cut-off [14] applied, eliminating terms too frequent or too rare; and (4) terms are weighed according to the *term frequency inverse document frequency (tf-idf)* [24]. The result of this process is a matrix where each line (vector) represents a document, and the columns (dimensions) represent the terms. The measure used to define distance amongst the documents is a cosine-based metric [8].

Normalized Compression Distances (NCD) [6], approximates the Kolmogorov complexity for character strings, slightly modified to accommodate problems related with compressor accuracy [27]. To evaluate such measure we need to compress the texts individually and pairwise, and perform simple operations with the sizes of the compressed files. This approach does not requires preprocessing the document collection. In the experiments we have used bzip2 for compression.

After distance matrix evaluation, the tree is constructed using NJ and plotted in the plane using a linear algorithm designed to create radial layouts of trees [2]. Starting with the root v of the tree, the algorithm assigns to each subtree descending from v a wedge of angular width proportional to the number of leaves in T. The same partition is applied recursively. Tree edges are drawn along wedge angle bisectors and subtrees are kept disjoint on the plane. This algorithm results in a layout where edges' lengths are proportional to those generated by the NJ algorithm. Our trees are unrooted so we use one of its centers as the root. The initial layout provides a high level view, useful to inspect the tree's overall structure. It does, however, present a high degree of overlapping. To obtain the final layout, some iterations of a simplified FDP model is applied to the nodes constrained by the tree connectivity.

The full process, including NJ and the layout algorithms was implemented as an extension of a multidimensional visualization tool named *Projection Explorer* (PEx)[2] and therefore can be explored using the set of resources implemented in the tool. Among other things, PEx allows different neighborhood relationships to be visualized as edges on the map (for example, on the multidimensional space and on the bi-dimensional space), easy content display of documents and its neighbors to support in-depth analysis, coloring nodes that represent the documents according to the frequency of occurrence of a word or group of words in the documents, and creation of labels identifying the main topics in a selected group of documents.

In the next section we present the results of applying this process to data sets of various kinds and compare that with the results of projections of the same data sets.

[2]Freely available at `http://infoserver.lcad.icmc.usp.br`.

4 RESULTS

These results summarize several tests of mappings performed with distinct document collections. Table 1 provides details about such collections. The first collection includes scientific papers with title, authors, abstract and references on four different subjects: Case-based Reasoning (CBR), Inductive Logic Programming (ILP), Sonification (SON), and Information Retrieval (IR). The first three subsets were taken from journals on those subjects, and the other was obtained as a result of Internet searches. Those were all collected by members of our team. The KDVis set was obtained from an Internet repository and includes files in the ISI format on the subjects of Bibliographic Coupling (BC), Co-citation Analysis (SC), Milgrams (MG) and Information Visualization (IV) [4]. The INFOVIS04 data set was made available for the 2004 IEEE Information Visualization Contest [9]. The ALL set is formed by the sets above put together. The set MESSAGES are messages from three different news discussion groups obtained from an Internet repository [11]. And, the NEWS data set is composed by RSS news feed articles, collected from Associated Press (`www.ap.org`), Reuters (`www.reuters.com`), BBC (`www.bbc.com`), and CNN (`www.cnn.com`) Web sites, during two days in April 2006.

Table 1: Datasets used in the tests

Data Set	Type of Source	Number of Files
CBR+ILP+IR+SON	Scientific papers	680
KDVis	Scientific papers	1,624
INFOVIS04	Scientific papers	515
ALL	Scientific papers	2,819
MESSAGES	Discussion messages	300
NEWS	RSS Flash News	2,684

4.1 Mapping scientific data sets

Figure 1 presents a map for the CBR+ILP+IR+SON data set. On this map, the circles representing the documents are colored (pseudo-classified) according to their main subject (red for CBR, yellow for ILP, light blue for IR and dark blue for SON). It is possible to see that the NJ technique keeps documents that are considered to be in the same general in the same subtree, suggesting that the approach presented can separate and group documents based on their content. Also, in that picture, we chose five documents belonging to sonification sub-area which we know to have a high degree of similar content (they are an evolution of the same sonification system) and color the points that represent them in green. These points are identified in region **A** on the picture. This shows that the technique can also repeat the grouping result when the tree is examined in more detail, propagating the hierarchy of the similarity and possibly allowing identification of sub-areas within a more general one.

On the CBR+ILP+IR+SON document map we expected to have some documents seemingly misplaced according to pseudo-classification, since documents in the IR class were classified as such just because they resulted from a search in that subject. For instance, most documents identified in region **B** are classified as IR, and share a branch with documents classified as SON. Closer examination of such placement reveals that they are actually well placed since they discuss 'audio information retrieval', therefore bearing similarity with sonification research issues. Another example is the set of documents identified on that picture as **C**. They were pseudo-classified as IR documents, but in fact they relate to 'learning algorithms', and would arguably be classified better as CBR.

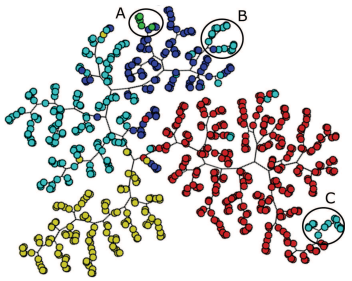
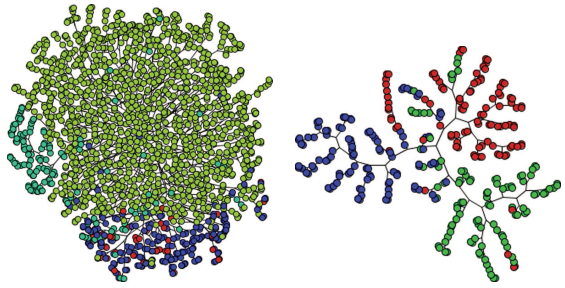


Figure 1: CBR+ILP+IR+SON document map. Points represent the documents, colored according to the area they belong to.

The same test based on pseudo-classification was conducted on various other data sets. In Figure 2(a) the KDVis data set was used with points colored according to the document area they belong (red for BC, blue for SC, light green for MG and yellow for IV). Again, the tree separates these areas well. BC and SC are mixed since they are quite related. On this map, although we have a dominant area with much more documents than the others (IV alone is composed of 1,236 documents) the NJ algorithm performed well. In Figure 2(b) we have the map for the MESSAGES data set, with points colored by message groups (3 groups), reproducing similar results. The tests show that the technique can handle well documents from different types of sources, provided the same is true for the similarity measure.



(a) The KDVIs document map. (b) The NEWS document map.

Figure 2: Document maps for two different collections. Color is predefined class. (a) comprises scientific papers and has a dominant subject area with more than 76% of the documents; (b) is composed by messages of 3 different discussion groups.

The INFOVIS04 data set is composed by documents published in the same conference on information visualization. Thus its content is much more homogeneous, what suggests that it is harder to group documents by content. Figure 3 presents a document map for this data set. Even in this case, documents with similar subjects were grouped and separated well. On this map, four different sub-topics within information visualization field are circled. In such cases, the points representing the documents are colored according to the frequency of occurrence of the words that named these sub-topics.

Using NJ, documents with high degree of similarity will be placed in the same branch. Thus, if it is possible to identify long branches without too many ramifications, they probably represent specific sub-topics inside the collection. In Figure 3, the branches circled and named as 'Treemaps', 'Spreadsheets', and 'Protein Sequences' are examples of long branches. The 'Graph Drawing' branch can be split into different sub-topics, such as aesthetic drawing or drawing large graphs, and this branch actually becomes a sub-tree. It indicates the maturity of the graph drawing subject.

In order to compare the visual representation by NJ trees with

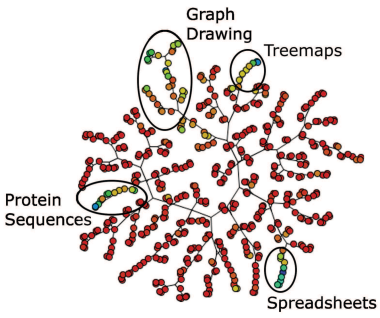


Figure 3: INFOVIS04 document map, a more homogeneous data set. Four different topics on information visualization are identified by coloring points containing words of interest.

projections we performed point placement using LSP for some of them. Although this technique is our own, it has been proved effective before for these same data sets. LSP also shares the aspects we wish to discuss (high density of points cluttering view and groups of similar documents being placed further than the user might expect) with all other projection techniques.

Although the visual representation for projection techniques and for NJ is similar – nodes are documents and edges are relationships between them – their results must be interpreted differently. In the case of NJ, groups of similar documents will compose a branch in the tree. For projection techniques in general, groups of similar documents will be placed in the same neighborhood on the final map. Although this feature of the projections may be very useful to recognize clusters of similar documents, sometimes this can cause an excessive overlap of nodes, impairing interpretation of density. Figure 4 presents a projection of the same data set used in Figure 3. In this map it is possible to identify the same topics of the previous map. However, the nodes on the map's core overlap, which makes it difficult to identify groups of similar documents within that region.

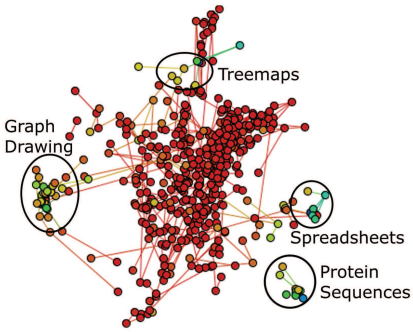


Figure 4: Projection of the same data set as in Figure 3. Edges connect nearest neighbors given by the distance matrix.

In the previous examples the distance measure used to create the maps was cosine distance evaluated on the vector representation of the documents. In Figure 5, the similarity measure used was NCD for the ALL data set. Again, NJ was capable of separating the documents well by their content, showing its consistency across different distance measurements. Actually, NJ phylogenetic trees will be able to join neighbors properly every time the distance measure is able to produce a good distinction of content.

The time complexity of NJ reflects on the times taken to build trees, as shown in Table 2. That aspect is discussed in Section 5.

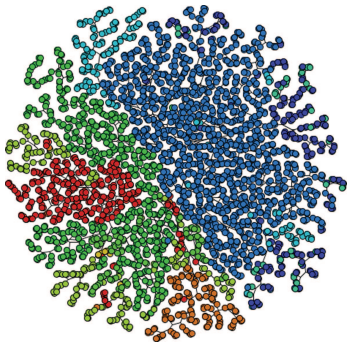


Figure 5: The ALL document map. All previously data sets composed of scientific documents were put together and NCD similarity was employed.

Table 2: Time in seconds to create maps in a 3.2 GHz Pentium 4.

Data Set	NJ	Layout	Total
CBR+ILP+IR+SON	4,55	0,52	5,17
KDVis	66,20	1,26	67,46
INFOVIS04	1,83	0,45	2,28
ALL	454,66	2,17	456,83
MESSAGES	0,35	0,31	0,66
NEWS	359,63	1,70	361,33

In the next section an example of exploration of a text data set using an NJ document map exploration is presented.

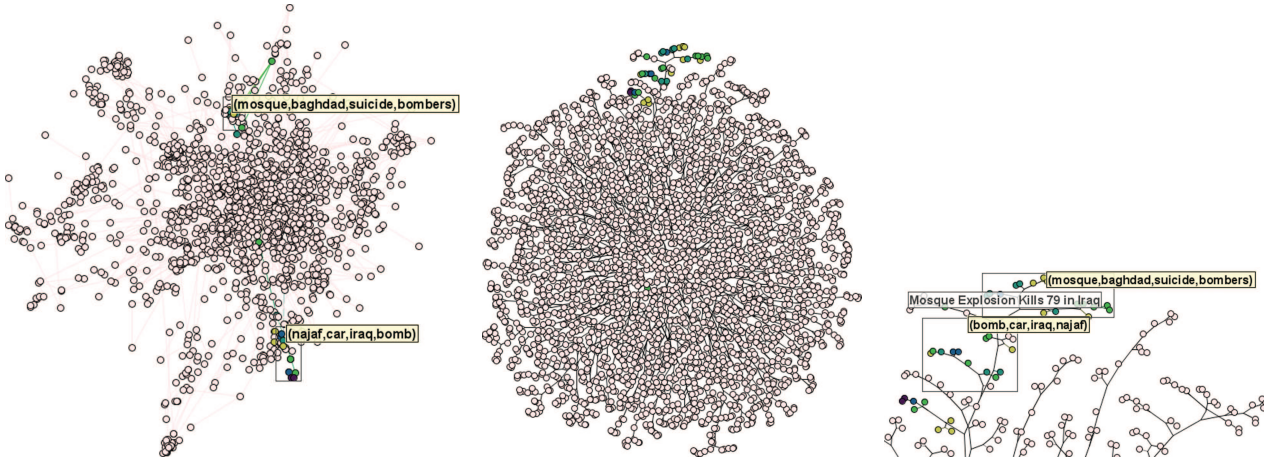
4.2 Exploring RSS feeds of flash news

In this section we present a test case to illustrate the effectiveness of similarity-based NJ phylogenetic trees. For this test we used RSS feeds available in web services of news agencies (NEWS data set in Table 1), a particularly difficult case due to the small size of most news files.

Some of the main news developing during those two days (April 6th and 7th 2006) were:

- A swan was found dead in Scotland, then there were fears of bird flu, then the bird was tested, then the virus was confirmed and then authorities put the population at ease concerning the case.
- New Immigration bills in USA had one of their first rounds at the Senate.
- The case of intelligence leak at the White House concerning Iraq was being investigated.
- The case of authorship for Da Vinci Code made headlines.
- Tornadoes pounded Tennessee.
- Two separate bomb attacks in Iraq took place, one in Najaf, another in Baghdad, killing different numbers of people.
- In the Middle East, recognition of Hamas Palestinian government and of Israel by the Hamas government were under scrutiny; there were talks of cutting aid to Palestinians; Israeli air raids happened in the Gaza strip.
- Augusta Golf Masters tournament was taking place.

In the pictures that follow we use one of our labeling systems, available in PEx, to display the location of the main subjects on the news map. Labels are key elements for exploring document maps. In order to enable visual identification of the main topics

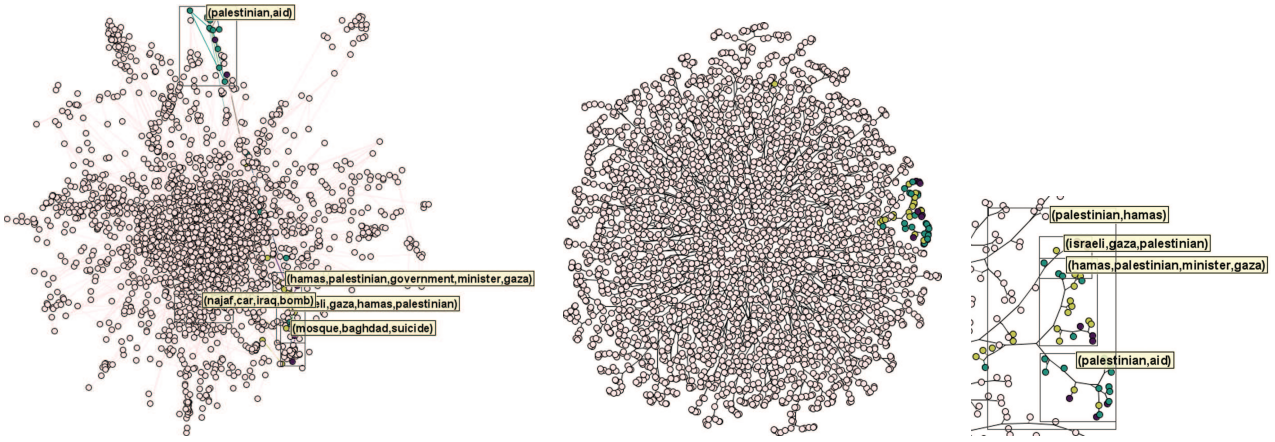


(a) Projection of the NEWS data set. Highlight of the word search 'bomb and iraq'.

(b) The 'bomb and iraq' search in the phylogenetic document map.

(c) Detail of the region of news on bombings in Iraq.

Figure 6: Projection and phylogenetic document maps for the 'bomb and iraq' search.



(a) Second projection of the news data set. Highlight of 'palestinian' word search.

(b) The 'palestinian' search in the phylogenetic document map.

(c) Detail of the region of news on Palestinian affairs.

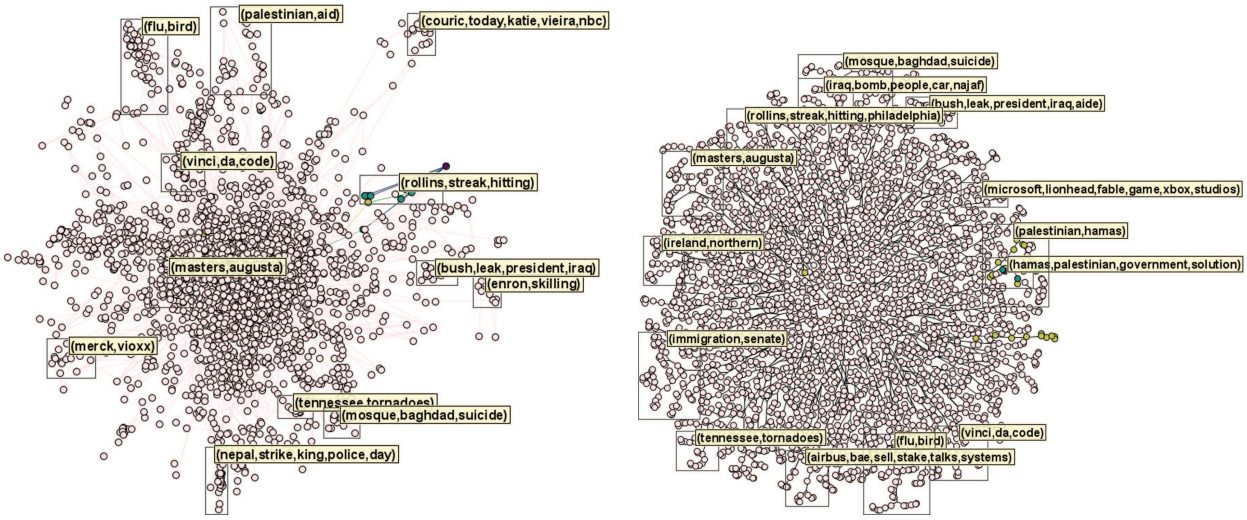
Figure 7: Maps of the NEWS document set by projection and phylogenetic tree.

discussed in groups of texts, an area of the layout can be selected using the mouse; a label is then generated that is a representative of the documents within this area. That label is created by first selecting the two words with the highest covariance in the vector representation of the selected documents. For each remaining word the mean of the covariances taking the two words selected in the first step is evaluated, and if the mean is higher than a threshold, the word is also included in the label.

The first aspect we would like to illustrate regarding the organization of NJ phylogenetic trees is in regards to the influence that certain documents can exert in more sensitive layouts such as clustering or projection. Figure 6(a) shows a first attempt projection for the NEWS data set with points matching to the word search 'bomb and iraq' colored. In the picture it can be seen that the news on each bombing (Najaf and Baghdad) were placed close together but the two groups of news were laid-out far apart on the map; generally they refer to 'suicide bombs in Iraq' and one might expect them to be placed nearby. Closer examination reveals news regarding bombings in other places and their relationships to other issues. However, since the higher similarities between Najaf's and Baghdad's bombings are identified by the cosine metric, the phylogenetic tree places all news regarding those bombings in Iraq closer together in neighboring branches (see Figures 6(b) and 6(c)). Other neighboring news in the same branches are either the same news using other words (such as explosions instead of bomb) or on the subject of bombings some place else.

A second projection attempt of the same data placed those two groups close together (see Figure 7(a)), but well mixed with other groups. On that same map, three different subjects relating to Palestinian affairs were placed apart from one another. The placement of these news regarding 'palestinian affairs' in a phylogenetic tree can be examined in Figures 7(b) and 7(c). The example above illustrates the relative low sensitivity of the phylogenetic trees for textual documents that are 'connected' with two subjects yet belonging to a third.

Regardless of possible difficulties with placing small text samples (such as RSS news feeds) in consistent levels of abstraction (such as illustrated in the previous figures), projections tend to group highly related text close together in these cases. Figure 8(a) shows the location of topics on the map of the NEWS data set, consistent with grouping of news about the same event in the same neighborhood. For NJ phylogenetic trees that is also the case. The



(a) Grouping by topic on the map by LSP projection.

(b) Grouping by topic on the map by phylogenetic trees.

Figure 8: Groups of news occurring in the NEWS maps.

groups are located in the same branches, as shown in Figure 8(b). Notice that documents with higher degree of correlation are placed mostly in the outer branches.

One type of textual document that usually makes point placement difficult is that for which the similarity measure returns a high degree of similarity to many other documents (i.e., text that is 'nearest neighbor' to too many others). These tend to unbalance local point placement and unite in the same groups too many topics or subjects. Figure 9 shows the projection of 18 artificially generated variations of the same C program, with one document (in red) being the nearest neighbor to 11 others (in yellow). Grouping occurs around it, disturbing the location of sub-grouping amongst the others. Figure 10 shows the phylogenetic tree for the same set. Once the common nearest neighbor is placed, naturally in the first steps of the algorithm, the remaining ones will be placed next, but it will still be possible to locate degrees of proximity between the others on the display by the type of branching because relative sizes of branches reflect as best as possible the actual distance.

Although text visualization was the core application to which we employed the technique, neighbor-joining phylogenetic trees build in this way can, of course, be used for many different multidimensional data, provided users count on a similarity measure between them. As a final example, we present a phylogenetic map of time series representing stream-flow measurements for the year 2005 of 76 hydroelectric reservoirs on the basin of Paraná river in Brazil. In Figure 11, color represents the sub-basin to which the reservoir belongs. The display shows that similar behavior, according to that particular metric, occurs within the same sub-basing, with few exceptions.

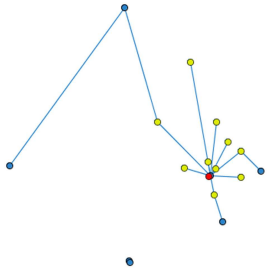


Figure 9: Placement by projection with a point (in red) with a high number of nearest neighbor connections.

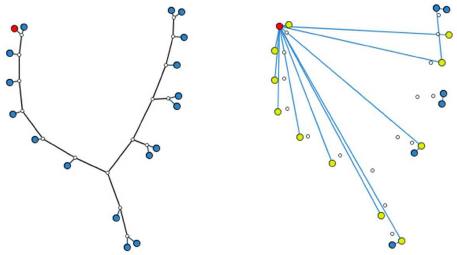


Figure 10: Placement by phylogenetic tree (left) and nearest neighbor connections (right).

5 Concluding Remarks

We have proposed a novel approach to reflect content relationship in a visual representation of multidimensional information, particularly collections of textual documents.

For such application, as well as others, the technique can reflect similarity relationships more precisely than the available point placement strategies. Our strategy, being capable of constructing a hierarchy from that similarity relationship contributes largely to decrease the loss of information caused by the 'attraction by distance' nature of conventional FDP.

The approach borrows from phylogeny the concept of reconstructing ancestry from similarity. Since in this case ancestry is also similarity (first ascendant more similar than 'older' ascendants), interpretation of this particular display is naturally obtaining by following the branching of the trees.

The interpretation is complementary to that of the projections and can be used in joint and multiple views of the same data set (since the preprocessing is basically the same for both). While groups take the form of larger densities of points in projections and other point placement strategies, our method places more consistent groups in the outer branches. The inner branches comprise those documents with lesser correlation to the more dense ones and also the documents that bridge different subjects.

In phylogenetic trees, points placed in the inner branches signify larger distances to the others than the ones at the outer perimeter

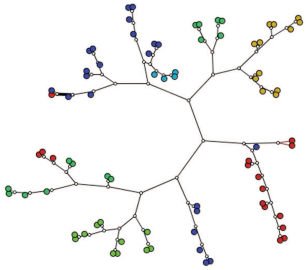


Figure 11: Placement by neighbor joining of stream-flow in hydroelectric plants of Parana River (Brazil). Color is sub-basin of the river.

of the tree. That should be evaluated further in terms of grouping of documents with less correlation but yet belonging to the same subject.

This paper also presented a form for automatically detecting subjects under discussion in groups of points in any display by using a term co-variance algorithm. This, coupled with the exploration of the display, present valuable tools for identifying subjects approached by a text collection as well as to identify regions of interest for further examination.

An additional advantage of this type of display is that a particular distance matrix always generate the same tree, which helps interpretation by fixing a mental model of the map faster than most point placement strategies, that have to have a decision as to where it stabilizes or as to how to combine the points into neighborhoods, producing a number (many times a large one) of possible solutions for the same distance matrix.

In our path of development there is a plan for constructing a set of tools for proper exploration of phylogenetic trees, allowing the same power of exploration that some tools, specially developed to explore projections, already posses (e.g. PEx).

Other algorithms exist for phylogenetic reconstruction. Although Saitou and Nei have already compared some of them to NJ [23], the question of whether there is a better algorithm to be used with texts is still to be answered. Asymptotically faster versions of NJ exist (e.g. [15]), but they were not tested yet for sets of documents. They could help to reduce processing time, strengthening the technique introduced here.

ACKNOWLEDGEMENTS

This work has been supported by Brazilian Agencies FAPESP and CNPq. We wish to acknowledge the support of Roberto Pinho for the original search code to recover RSS news feeds, Isaura Cronemberger for the similarity data in program plagiarism, and Aretha Alencar for the similarity data on stream-flows

REFERENCES

[1] K. Andrews, W. Kienreich, V. Sabol, J. Becker, G. Droschl, F. Kappe, M. Granitzer, P. Auer, and K. Tochtermann. The infosky visual explorer: exploiting hierarchical structure and document similarities. *Information Visualization*, 1(3/4):166–181, 2002.

[2] C. Bachmaier, U. Brandes, and B. Schlieper. Drawing phylogenetic trees. In X. Deng and D. Du, editors, *Proc. Intl. Symp. on Alg. and Comp. , ISAAC 2005*, volume 3827, pages 1110–1121, 2005.

[3] K. Borner, C. Chen, and K. Boyack. Visualizing knowledge domains. *Annual Review of Info. Sci. & Tech.*, 37(1):179–255, 2003.

[4] K. Börner. *KDVis*. `http://ella.slis.indiana.edu/~katy/outgoing/hitcite/{bc,sc,mb,iv}.txt`, 2005.

[5] M. Chalmers. A linear iteration time layout algorithm for visualising high-dimensional data. In *Proceedings of the 7th IEEE Visualization, (VIS)'96*, pages 127–132, Los Alamitos, CA, USA, 1996. IEEE Computer Society Press.

[6] R. Cilibrasi and P. Vitányi. Clustering by compression. *IEEE Trans. Information Theory*, 51(4):1546–1555, 2005.

[7] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman & Hall/CRC, second edition, 2000.

[8] C. Faloutsos and K. Lin. Fastmap: A fast algorithm for indexing, datamining and visualization of traditional and multimedia databases. In *ACM SIGMOD Intl. Conf. on Management of Data*, pages 163–174, San Jose-CA, USA, 1995. ACM Press: New York.

[9] J.-D. Fekete, G. Grinstein, and C. Plaisant. IEEE InfoVis 2004 Contest, the history of InfoVis. `http://www.cs.umd.edu/hcil/iv04contest`, 2004.

[10] T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software - Practice and Experience*, 21(11):1129–1164, 1991.

[11] S. Hettich and S. Bay. *The UCI KDD Archive*. `http://kdd.ics.uci.edu`, Irvine, CA: University of California, Department of Information and Computer Science, 1999.

[12] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 2 edition, 2002.

[13] F. Jourdan and G. Melancon. Multiscale hybrid mds. In *IV '04: Proc. of the Conf. Information Visualisation*, pages 388–393, Washington, DC, USA, 2004. IEEE Computer Society.

[14] H. Luhn. The automatic creation of literature abstracts. *IBM J. of Research and Development*, 2(2):159–165, 1968.

[15] T. Mailund, G. Brodal, R. Fagerberg, C. Pedersen, and D. Phillips. Recrafting the neighbor-joining method. *BMC Bioinformatics*, 7:29, 2006.

[16] K. V. Mardia, J. T. Kent, and J. M. Bibby. *Prob. and Math. Statistics*, chapter Multivariate Analysis. Academic Press, 1995.

[17] A. Morrison and M. Chalmers. A pivot-based routine for improved parent-finding in hybrid mds. *Information Visualization*, 3(2):109–122, 2004.

[18] A. Morrison, G. Ross, and M. Chalmers. Fast multidimensional scaling through sampling, springs and interpolation. *Information Visualization*, 2(1):68–77, 2003.

[19] Pacific Northwest National Laboratory (PNL). IN-SPIRE^{TM} Visual Document Analysis. `http://in-spire.pnl.gov/`, 2007.

[20] F. V. Paulovich and R. Minghim. Text map explorer: a tool to create and explore document maps. In *IV '06: Proc. of the conf. on Information Visualization*, pages 245–251, Washington, DC, USA, 2006. IEEE Computer Society Press.

[21] F. V. Paulovich, L. G. Nonato, R. Minghim, and H. Levkowitz. Visual mapping of text collections through a fast high precision projection technique. In *IV '06: Proc. of the conf. on Information Visualization*, pages 282–290, Washington, DC, USA, 2006. IEEE Computer Society Press.

[22] M. F. Porter. An algorithm for suffix striping. *Program*, 14(3):130–137, 1980.

[23] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4(4):406–425, 1987.

[24] G. Salton. Developments in automatic text retrieval. *Science*, 253:974–980, 1991.

[25] J. C. Setubal and J. Meidanis. *Introduction to computational molecular biology*. PWS Publishing Co., 1997.

[26] E. Tejada, R. Minghim, and L. Nonato. On improved projection techniques to support visual exploration of multidimensional data sets. *Information Visualization*, 2(4):218–231, 2003.

[27] G. Telles, R. Minghim, and F. Paulovich. Normalized compression distances for visual analysis of document collections. *Computer & Graphics, Special Issue on Visual Analytics (to appear)*, 2007.

[28] J. Wise, J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information for text documents. In *Readings in information visualization: using vision to think*, pages 442–450, San Francisco, CA - USA, 1995. Morgan Kaufmann Publishers Inc.

[29] J. A. Wise. The ecological approach to text visualization. *J. of the American Soc. for Inf. Sci.*, 50(13):1224–1233, 1999.