Content-based Text Mapping using Multi-dimensional Projections for Exploration of Document Collections

Rosane Minghim, Fernando Vieira Paulovich and Alneu de Andrade Lopes

Instituto de Ciências Matemáticas e de Computação,
CP 668, São Carlos 13560-970, São Paulo, Brazil

ABSTRACT

This paper presents a technique for generation of maps of documents targeted at placing similar documents in the same neighborhood. As a result, besides being able to group (and separate) documents by their contents, it runs at very manageable computational costs. Based on multi-dimensional projection techniques and an algorithm for projection improvement, it results in a surface map that allows the user to identify a number of important relationships between documents and sub-groups of documents via visualization and interaction. Visual attributes such as height, color, isolines and glyphs as well as aural attributes (such as pitch), help add dimensions for integrated visual analysis. Exploration and narrowing of focus can be performed using a set of tools provided. This novel text mapping technique, named IDMAP (Interactive Document Map), is fully described in this paper. Results are compared with dimensionality reduction and cluster techniques for the same purposes. The maps are bound to support a large number of applications that rely on retrieval and examination of document collections and to complement the type of information offered by current knowledge domain visualizations.

Keywords: text visualization, document mapping, domain knowledge visualization, multi-dimensional projection, IDMAP

1. INTRODUCTION

In the quest for relevant information on a particular subject in document data bases, it is a fact that the exploration of the results obtained has become a tiresome task due to the amount of information retrieved by most search engines, and the lack of comprehensive structure in their responses. Most search engines, and other supporting tools for data interpretation, return the information about documents that, in spite of been progressively enhanced, does not allow the identification of higher level structure possibly revealed by the documents, such as new research areas or trends.

In order to support the task of browsing through a recovered sets of documents, there is a number of techniques for document ranking and for visualization and mining. Using such techniques, document maps can be generated to help users to identify the information contained in the documents, and the relationship between them. However, the level of information that users can extract from such maps still needs to be improved. In this context, new visualization support must be sought to help the user pursue useful information.

This paper presents a complete methodology for generation of document maps that improves the existing support to the identification of structure within bodies of documents, as well as the identification of relationships between documents in a document collection. The methodology is based on multi-dimensional projection, resulting on maps that group similar documents effectively. As a result, it allows users to identify similarity, relevance, and sub-areas of activity based on document content. The display is a surface that can be explored using an interactive tool that allows users to examine individual document properties and neighborhoods.

The results are promising and are evaluated in terms of processing speed and effectiveness of representation as compared to other available text visualization techniques.

Further author information: (Send correspondence to Rosane Minghim)
Rosane Minghim: E-mail: rminghim@icmc.usp.br, Telephone: +55 (0)16 3373 9730
Fernando Vieira Paulovich: E-mail: paulovic@icmc.usp.br, Telephone: +55 (0)16 3373 9730

Section 2 presents a background review on text visualization techniques, highlighting the motivation for this work. Section 3 describes the full processing for document map generation as well as its justification. Section 4 discusses the results. Conclusions are drawn in Section 5.

2. PREVIOUS WORK

Due to the complexity and variety of the information and scenarios involved in text examination, alternative means of mappings text sets must be sought. Here we review the works in the literature that deal with this problem that, in our view, cover the main issues relating to visual mapping techniques for documents.

A number of different techniques for visualization of textual results from Web and other searches have being deployed.[1–5] While these techniques are capable of displaying large text bodies, they tend to make location of relevant reading material more troublesome. Our focus in this work is to provide complementary tools to support mapping of documents in a way that helps locate neighboring similarities between texts and groups of texts. So we assume a pre-filtering task that reduces the universe of targeted documents to a few hundreds (maybe thousands) of texts in a few areas of interest (not necessarily pre-determined).

Many techniques for text visualization exist that search for a representation of the content of an individual text,[6,7] of text collections,[2,8,9] or of themes grouping texts[10–12] in order to meet the above mentioned targets.

Usually text processing tasks employ the vector space model[13] whereby texts are represented as points in a vector space. In this representation each text is a vector with dimensions represented by terms (n-grams). The vector coordinates are the weights of the terms based on their frequency. Typically, dimensions reach the thousands even for small to medium databases. Transformation of a text collection into a vector space is preceded by elimination of non-influential words (stopwords), reduction of words to their radicals (stemming), and frequency counting of some sort (various exist). The initial representation is followed by reduction in space dimensions, typically involving cutting off words that are too frequent or too rare in that particular collection.

The most common way to extract structure from a text collection is by applying some sort of dimensional reduction technique over the resulting vector representation. This is the case of systems based on Multidimensional Scaling (MDS), Principal Component Analysis (PCA) or Latent Semantic Analysis (LSA), that work with statistical measures for subspace reduction, and Self-Organizing Maps (SOM), that employ neural computation.[8,9,11,12,14] Those techniques can be used to plot the original data in bidimensional (2D) space, when dimension is reduced to 2.[15,16]

Although dimensionality reduction is a natural processing trend for texts, these types of techniques have high computational costs and low adaptability to incremental processing. Multidimensional reduction techniques also cause other difficulties, such as[17]: high information loss when applied directly to two dimensions (for display); reduction in input dimensions do not seem to affect greatly the outcome; and there is an inherent discretization problem associated with techniques such as SOM, by which individual documents in groups are not distinguishable. For the target of this work, dimension reduction poses and additional problem: when used to display the results in 2D, the mappings to subspaces may define groups of 'similar documents', but locally it is not possible to relate neighboring texts. In a previous work[16] LSA has been successfully applied for the generation of document maps with high local content relationships, but the high computational cost remains a problem.

Another recurring strategy for dealing with the organization of information from a text collection is document clustering,[18,19] many times employed in combination with dimensional reduction and SOM.[2,9,20] They provide a way of relating documents with varying success rates. When clustering techniques are applied, here too the intra-cluster relations are not given as a result. However, they are very useful to provide general overviews of large collections, although they usually have to be interpreted by users with certain level of expertise.

Point placement strategies and force-based point placement improvement have been used before to generate document displays.[5,21] Also based on vector representation, they can avoid partly or completely the extensive calculations needed in dimensional reduction techniques by starting with a semi-random point placement and re-adjusting their position based on attraction by similarity.

There are approaches that completely avoid the problem of high dimensionality by simply ordering the most used terms in the text and employing the first N terms.[7] These strategies work well for single text representation and for association of a limited number of texts, and even for some degree of clustering. However, it also lacks a way of clearly relating different documents and displaying levels of similarity. Other approaches (such as the one by Carey and others[18]) combine a number of different strategies to allow various views of the same document set, potentially improving focusing and analysis tasks.

A few systems are being developed dedicated to viewing maps from multi-dimensional data and some of them are particularly dedicated to text collections. One recently published system[21] adds representational power to the conventional ways of plotting text as points in 2D by separating their contents in thematic areas and handling levels of interaction by hierarchical organization. Final maps resulting from the techniques mentioned above are meant to analyze a number of properties of documents, including similarity, co-citation, term co-occurrence and various others. We refer to the work of Katy Borner and others[22] for a detailed description of the available techniques for text mapping and its applications, systems and challenges.

In general the methods discussed above lack the ability to determine levels of associations between texts contents. Others are computationally expensive. The technique presented here puts forward faster mapping approaches with the ability of approaching texts by similarity, allowing elements with similar content to be placed in the neighborhood of one another. The gain in processing time is attained by using projection techniques, which are faster compared to dimension reduction and less troublesome as far as data pre-processing is concerned. Projections also provide an initial point placement prone to speed up force-based improvement schemes.[23]

The primary visual representation adopted here is the landscape-type of display, which is very useful due to its ability to reveal information without resorting to highly attentive perceptual processes, allowing interpretation even by users with little expertise in the field. It has been the choice of many useful presentations of texts before.[9, 11, 24, 25] Additionally, surfaces are highly interactive and familiar to most users. Surface representations is enriched by mapping further significant information (such as degree of similarity) to visual attributes (such as lines, colors and height) and aural attributes (such as pitch and timbre). The final map can be explored by the users interested in having an overview of a set of texts, locating important texts in the corpora, or finding useful associations between texts through an interaction tool for that purpose. Alternative graph views are also available.

The next section presents the complete mapping process from pre-processing to projection and attribute mapping procedures.

3. PROJECTION TECHNIQUES FOR TEXT VISUALIZATION

A previous work[23] has shown the advantages of projection techniques to obtain useful views of multi-dimensional data sets based on distance metrics. Additionally, those techniques perform well in terms of processing speed and lend themselves to landscape plots. This work set off to find out whether similar techniques could be successfully adapted to representation of text sets.

Different from other techniques that can be used to map data into 2D or 3D, such as dimensional reduction and clustering displays, the goal of distance-based projection techniques - eg. Fastmap[26] and Nearest Neighbor Projection (NNP)[23] - is to place a set of points defined in multi-dimensional space in another space such that the relative distances between points are preserved as much as possible. The degree to which that distance cannot be preserved is called the error of the projection. For projections into a bi-dimensional space (plane), this problem can be stated as:

Let X be a set of points in $\mathbb{R}^n$ and $d : \mathbb{R}^n \to \mathbb{R}$ be a criterion of proximity between points in $\mathbb{R}^n$. We wish to identify a set of points P in $\mathbb{R}^2$ such that if $\alpha : X \to P$ is a bijective relation and $d_2 : \mathbb{R}^2 \to \mathbb{R}$ is a proximity criterion in $\mathbb{R}^2$, then $|d(x_i, x_j) - d_2(\alpha(x_i), \alpha(x_j))|$ is as close to zero as possible $\forall x_i, x_j \in X$. In this paper we call the set P a projection.

In order to create meaningful projections, the definition of a proper proximity criterion between the points of X is very important. Here, each document is represented as a vector, so that it is possible to calculate the distance (proximity) between texts numerically.

In this work, the steps taken to build a document map based on projection are:

1. text corpus pre-processing to build its representation in a vector space;
2. projection to two-dimensional space using a fast algorithm, followed by an improvement strategy (the Force Scheme)[23];
3. hierarchical clustering of the projected data for subgroup identification.

The following sections describe each step of the process in turn.

3.1. Text pre-processing

In order to generate the vector representation of the text set in this work, the original texts (composed by title, authors, abstract and references of articles) were submitted to the following procedure:

1. Stopwords were eliminated from all texts;
2. Stemming was applied to extract word radicals using Porter's algorithm.[27]
3. A frequency count was performed applying Luhn's cut[28] so that terms appearing less then 5 times were ignored.
4. Bi-grams were formed from the remaining words in the texts, that is, we considered as terms the occurrence in sequence of a pair of words.
5. A process to weight terms according to their frequency was carried out; in our case the weight was computed as the *term-frequency inverse document-frequency (tfidf)*.[29]

The result of that process is a matrix $T_{n\times l}$ of documents with n documents and l terms in which each term is weighted according to the tfidf. Each line of the matrix (a document) is a vector, each final bi-gram is a dimension, and the tfidfs are the coordinates.

For many text processing activities, the number of dimensions in the resulting table is still too large (usually it reaches few thousands). In many applications, where distance calculations become unstable with high dimensionality,[30] a pre-reduction of attributes (terms) must be performed before the projection takes place. We have experimented with a modified version of k-means feature clustering for attribute reduction for our target text collection but the results were not improved significatilly.

The documents x terms T matrix resulting from the vector space representation is then used to perform the projections of the data on bi-dimensional space, as described in the following section.

3.2. Projection techniques

In order to execute a projection based on the matrix T, a metric is necessary to establish a distance criterion between two different texts (d in the problem statement given previously in this text). A measure that usually performs well for document processing is the cosine metric, given by:

$$scos(t_i, t_j) = \frac{\sum_{k=1}^{l}(t_{ik} \cdot t_{jk})}{\sqrt{\sum_{k=1}^{l}(t_{ik}^2) \cdot \sum_{k=1}^{l}(t_{jk}^2)}} \tag{1}$$

where:

t_i and t_j are two lines of the matrix $T_{n\times l}$ of documents, that is, the vector representations of two different documents.

¿From that, a distance metric can be obtained,[26] given by:

$$d(t_i, t_j) = \sqrt{2 * (1 - scos(t_i, t_j))} \tag{2}$$

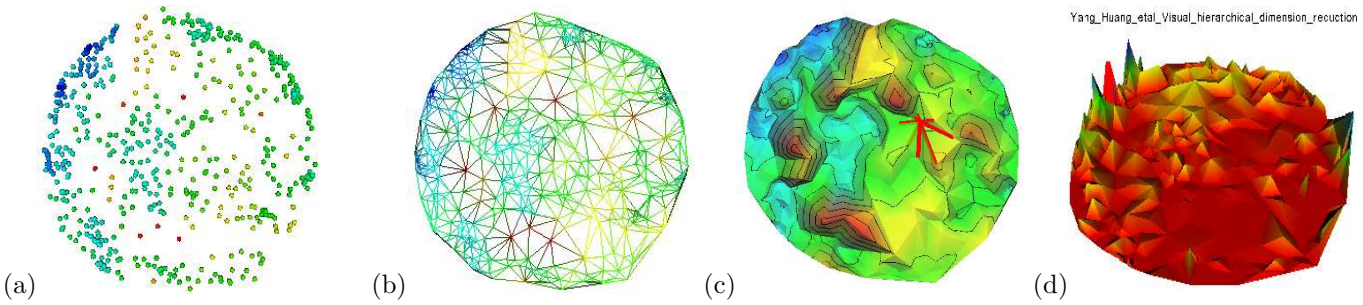

Figure 1. Generation of a surface model from text data. In this picture color and height are determined by the hierarchical clustering of the projected points. (a)'glyphed' view of the projected points via FASTMAP. (b) 2D Delaunay Triangulation of (a). (c) Isolines and height map for subgroup determination. (d) color and height mapping of relevance towards the query 'hierarchical+clustering+dimension'; the highest (blue) is most relevant; the red 'spider' is the exploration cursor.

Two distance based projection techniques were used in this work to generate the projection P of the set T of points. The first was Fastmap and the second was the NNP, with similar results. Fastmap[26] is a well known technique for multi-dimensional projection based on the idea of hyperplane projection. NNP[23] projects points defined in the original space onto 2D plane by trying to find a position that reproduces the original distances between each point to its 2 closest points. Details of both techniques can be found in the references provided.

In the context of multi-dimensional visualization, the original Fastmap tends to produce representations too compact while NNP tends to scramble neighboring points, impairing the effectiveness of the final result. A force-based projection improvement technique (see[23]) was used to enhance point placement, recovering part of the information lost during the projecting process. What this force scheme does is iteratively approach points projected too far and repel points that were projected closer than they should have, in a similar procedure as that adopted before by other researchers, such as Chalmers.[5] The main difference here is that, since the points were already projected, using fast techniques, with the effort to preserve distance, the number of iterations necessary to converge is really small. This step of the algorithm is quite fast when applied in the context presented here.

The final result of the projection is a set of points P in $\mathbb{R}^2$, each representing one document, so they can be plotted using their projected coordinates.

Visualizing the projected documents as points on a plane tends to compromise the interpretation and exploration of similar or related documents. Even when the points are represented by ghyphs such as spheres (see Figure 1(a)), neighborhoods do not get much clearer. In order to produce a better landscape view from the projection, we perform a Delaunay triangulation[31] over the vertices (see Figure 1(b)).

The triangulation is useful for many purposes besides improving the perception of neighborhood. It lends itself to other mappings that can be combined to highlight important information to the user (see Figure 1(c)). The subject of attribute mapping is discussed in the next section.

3.3. Attribute mapping

The projection gives a placement of documents on a plane, where the document positions are based on a similarity measure of documents. The triangulation offers a mesh of these data, whereby the neighbors of a particular point (vertices in incident edges) indicate the closest documents according to the projections.

Classically, visualization techniques can generate a number of graphical displays from this set-up, particularly if scalar (or even vector) data are assigned as attributes to the vertices of the mesh. In the display, colors or heights can help locate highlighted attribute information. Paper relevance, number of citations, year of publication, are all valid attributes in that context.

In order to use the scalar values for supporting the location of sub-groups of documents we have performed *Hierarchical Clustering (HC)* of the projected data. HC defines groups of elements progressively, using divisive or agglomerative approaches. By mapping the depth of the HC to each projected point (as a scalar value stored in the triangulation), new visual mappings can be obtained of levels of similarity among documents.

Figure 1 shows the mapping of that information (clustering depth) to color over the document map. In that case, the points colored in dark blue (or green) can be seen as the focuses of the various clusters of documents, that is, the documents grouped initially by the hierarchical clustering process. As the color changes from blue to red (in the rainbow color scale) new levels of grouping are achieved, until clusters merged closer to the root of the tree are viewed in the red regions. In our initial maps, that same information is redundantly mapped to height. Thus, the clusters with the most similar documents are placed closer to the hill peaks, and the roots are the valleys of the landscape. In addition we have also traced level curves (isolines) corresponding to various clustering levels, in order to ease the visualization of the sub-groups formed. From this figure it is possible to see that the curves help identify the borders of groups of documents.

An extensive number of properties can be mapped to visual (or aural) attributes. For instance, Figure 1(d) shows the same mapping as Figure 1(a) to (c), only now colors reflect the relevance of the papers towards a user search based on keywords. That relevance is calculated by cosine distance between the search vector and the papers' vector representations. Other possible mappings are exemplified in the Results section.

Many of the features that a map is capable of representing are more useful under interactive exploration. The possibilities of interaction with the map are presented in the following section.

3.4. Interacting with surface maps of documents

The *spider cursor* is a tool for interaction with visualizations using sound.[32] The tool allows exploration of a data set represented by a triangulation by showing a cursor (called *spider*) on top of the triangulation as the user moves the mouse over the surface. From a central point (located by the cursor) the neighbors in the triangulation are shown with line segments (looking like spider legs). One chosen scalar value stored in the vertex pointed by the cursor is mapped to pitch of a pre-selected instrument. Another document property, represented by a character string, can be shown on a field on top of the presentation window. The visual representation of the spider cursor can be seen in in Figure 1(c). The Spider Cursor is an extension of VTK (The Visualization Toolkit - (http://www.vtk.org)). The maps are therefore represented as VTK files. Other two interaction tools for meshes were developed, one that shows the map as Voronoi Diagrams and the other that is a triangulation display with location of neighbors that prints the names of the neighbors as well as the abstract of the related paper. Illustrations of the use of these tools as well as the color pictures for this paper can be found at: http://www.lcad.icmc.usp.br/∼rosane/textmaps.html.

The sound mapping available in the Spider Cursor is very useful, amongst other things, to resolve ambiguities in the visual mapping. Thus, two different documents visually undistinguishable in terms of their color or height can most times be told apart by the sound they produce when pointed at.

The techniques presented here (jointly called IDMAP - *Interactive Document Map*) resulted in a very useful interactive map for exploration of text collections (and other data). The next section presents some results obtained with the processing of various corpora using this methodology.

4. RESULTS

Using the IDMAP method presented here we have processed and interacted with a number of different data sets. Four of those, composed of research articles, are exemplified here. For the first data set, we processed three corpora (one on Case-Based-Reasoning (CBR), another on Inductive Logic Programming (ILP), and a third on Information Retrieval (IR)) comprehending a collection of 574 documents. The CBR and ILP corpora were manually extracted from Lecture Notes on those subjects and the IR corpus resulted from a web search. All the pre-processing of the text to extract the contents and normalize the references was done by student members of our team.

Reviewing the process, from the pre-processing step we have as result a matrix of documents x terms. That matrix was 574 documents x 5495 attributes for the CBR+ILP+IR corpus using a Luhn's cut off of 5.

These documents were then mapped using NNP and FASTMAP projections with similar results. Figure 2 shows the outputs of this process. In that figure color represents tagging of a particular paper as CBR, ILP, or IR. This is a pseudo-classification, since the source of the paper was the only criterion to determine its class. It is, however, a good basis for preliminary evaluation of the results.

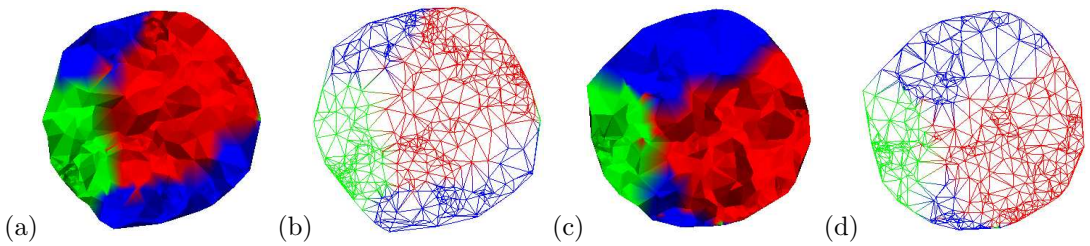

Figure 2. Projections of the CBR+ILP+IR corpus with coloring as pseudo-class. Color is pseudo-class. Red is CBR. Green is ILP. Blue is IR. (a) Projection using NNP with surface view. (b) 2D Delaunay triangulation of the projection in (a). (c) Projection using Fastmap with surface view. (d) 2D Delaunay triangulation of the projection in (d).

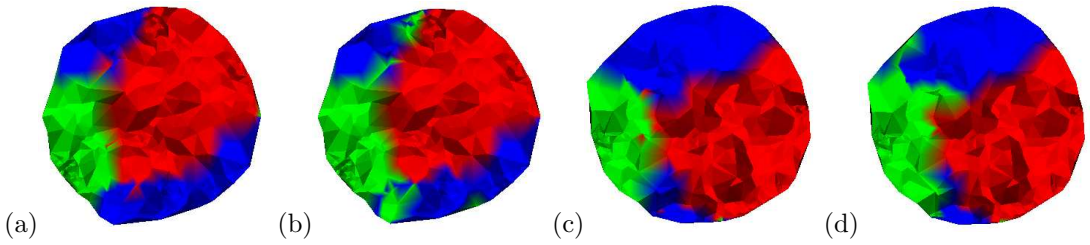

Figure 3. Projections of the CBR+ILP+IR corpus with color mapping. (a) Projection using NNP with color by pseudo-class (b) Projection using NNP with color by k-means classification (c)Projection using Fastmap with color by pseudo-class (b) Projection using Fastmap with color by k-means classification

As it can be observed in those pictures, the projections came a long way towards separating the main classes of documents. They occupy specific 'portions' of the map, apart from a small proportion of 'outliers' clearly distinguishable in the surface views of the projections. The fuzzy boundaries as well as the outliers are expected, particularly because the classes were inferred from their sources, and the three areas of knowledge have many concepts (and certainly expressions and terms) in common. Both NNP and FASTMAP projections have similar results for the maps.

To analyze the grouping by content capability of the projections we have performed two tests:

- grouping of the original set of documents employing a conventional clustering technique (k-means) using the corpus' vector representation.
- addition of a new set of documents to the collection, resulting from another search for a less related subject, sonification *.

The clusters formed by k-means clustering[33] of the original 3-subject text set matched the pseudo-classification very closely. In order to observe the similarities of these groups with the ones obtained by projection, we mapped the resulting k-means classes on top of the projection and visually compared with the original mapping with pseudo-class coloring. The result can be seen in Figure 3. It can be seen that there are very few areas of mismatch between pseudo and calculated clustering are found in the boundaries between the different 'classes'. Numeric values of that matching can be seen in Table 1. It should be noted that clustering per se does not perform point placement. Nor does it offer views of proximity by similarity. Those features are, however, provided by the projections.

The addition of the sonification document subset can be seen in Figure 4(a) and (b). The new corpus (SON) also separated well from the other three and also matched well clustering via k-means. In the same test, to verify the ability to approximate the display of files by content, we 'marked' 6 papers. Their tags started with 'INT-' (for intruders) and the color mapping turned out green. Five of the 'intruders' were papers involving members of our team on the subject of sonification, and represented an evolution of a system called DSVol (*Distributed*

*sonification is defined as the presentation of information through sound

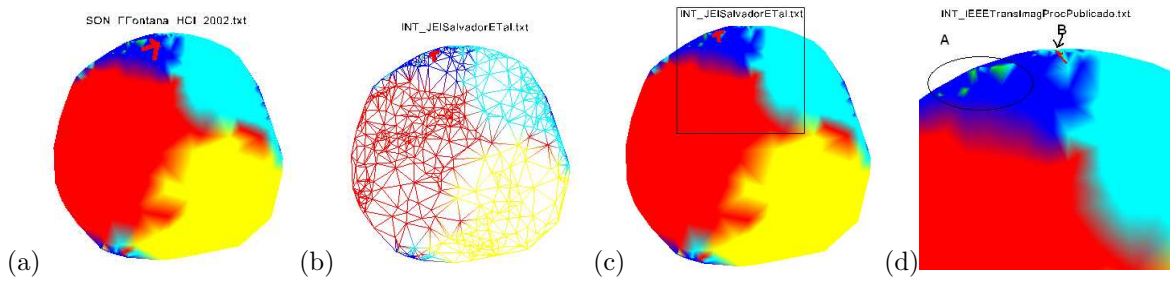

Figure 4. Projection of the CBR+ILP+IR+SON corpus using FASTMAP. Color is pseudo-class. Red is CBR. Light blue is IR. Dark blue is SON. Yellow is ILP. The label is the name of the file located by the pointing interaction tool (the spider cursor) (a) Surface view. (b) 2D Delaunay triangulation. (c) Location of the related papers intentionally added. (d) An amplified part of (c).

Sound for Volumes). The sixth paper was also by members of our team, but it was in a different subject (image segmentation). As Figures 4(c) and (d) illustrate, the papers were mapped quite close (see region marked 'A' in Figure 4(d). The mapping of the sixth, less related paper, was pushed to the border of the sonification region, on the place marked 'B' in Figure 4(d). This results indicate the ability to map related contents nearby in the mesh.

The final document map is obtained with scalar mappings over the projection meshes, as detailed in Section 3.3. Various attributes can be mapped, as described and illustrated before. The scalar field mapping the depth of the hierarchical clustering of the projected points is a numerical value that has the ability to support the location of pockets of articles mapped closer. Figure 5 shows the final maps of another four-subject corpora recovered from an Internet repository[†]. It comprehends 1624 files in the ISI format on the subjects of Bibliographic Coupling (BC - in red), Co-citation Analysis (SC - in blue), Milgrams (MB - in green) and Information Visualization (IV - in orange). It can be seen in those pictures that the IV group, which dominates the plot due to the number of files it comprises (1236 files), possesses five 'hill (blue) peaks', while other 'hill peaks' on the map also appear, one for each of the other classes (once BC and SC are very correlated areas, there is only one 'peak' for both of them). Those peaks indicate the areas of more concentration of closely related articles.

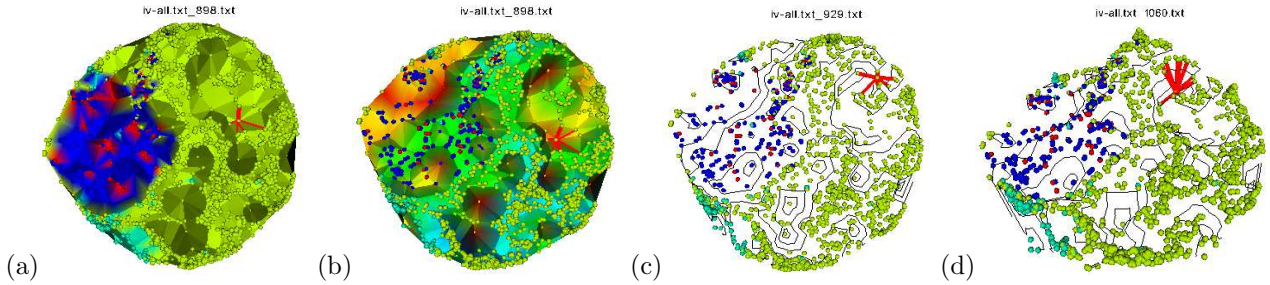

Figure 5. Final document collection maps of BC+IV+MB+SC corpus after projection by NNP. (a) Surface and glyph color is class (b) Surface color is HC, glyph color is class (c) Adding level curves and eliminating the surface to help locate sub-groups. (d) Perspective view of (c).

Another text collection mapped was the corpus from IEEE Information Visualization 2004 contest. They represent papers from 10 years of InfoVis plus papers most referenced by them. 574 of the original 614 papers were used in our maps (the others did not posses abstracts in the corpus).

In that subject area, classifying papers in big groups is not trivial, once most papers in information visualization do deal with a number of basic common techniques and concepts. Clustering and classification algorithms, therefore, have difficulty making sense of that corpus. All the more interesting to employ user centered mapping strategies in that context. Figure 1 presented the basic maps from that corpus. Additional pictures using the

[†]ella.slis.indiana.edu/∼katy/outgoing/hitcite/{bc,sc,mb,iv}.txt

spider cursor are shown in Figure 6. On top of the original map we have mapped to colors and heights to frequency of reference to certain terms. Number of references and relevance towards term query were mapped. Those maps allow to distinguish the region or regions of the map that do mention particular subjects of interest.

Combination views (such as that presented in Figure 6(d)) can help the user determine what articles are bound to fit more than one query at once. In that particular map, points on top of hills that are also colored different from red (which means low) will determine the papers that deal with visualization of text or documents and *also* with user interface. The map does not exclude the presentation of either subject group, so papers with either one of them can still be explored in the same display. Additionally, it can be seen in Figure 6(d) that those papers matching both searches (document visualization and user interface) appear in the border between the first and the second group of papers.

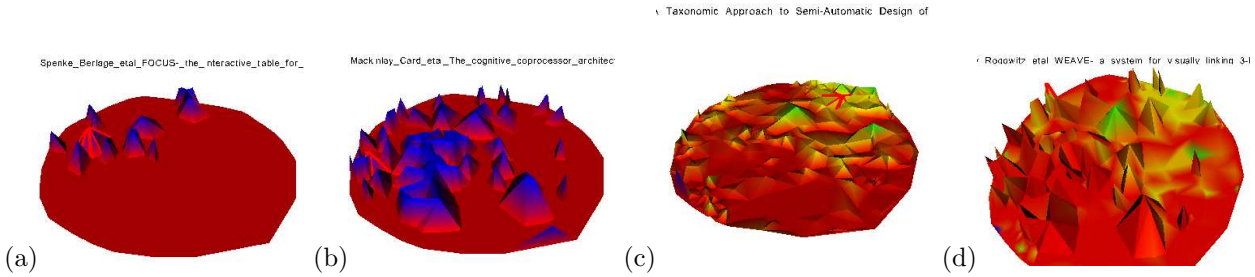

Figure 6. Final document collection maps of InfoVis04contest corpus. (a) Height (an blue) shows papers with above average mention of focus+context. (b) Height (an blue) shows papers with above average mention of user interface. (c) Height (or colors - blue higher - red lower) shows relevance towards the query 'document+text+visual'. (d) Combination view of (c) and (d). Higher AND colored are relevant towards the query and also mention user interface. Points just colored (other than red) have low or no mention of user interface. High red points are papers with no relevance towards 'document+text+visual'

These and other results of exploration of the maps indicate that the subgroups formed in the visualizations match the general idea of close association between papers, with few exceptions. Closer examination (both theoretical and practical) of that issue is being pursued.

Besides the spider cursor (mentioned in section 3.4), a specific tool was developed to explore neighborhood of plotted texts. Figure 7 shows an example of that tool.

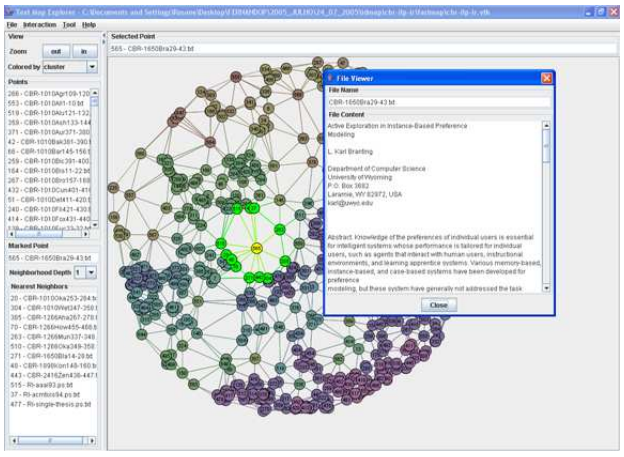

Figure 7. TextMapExplorer: An exploration tools for text maps.

Table 1 shows a summary of some of the aspects discussed on the mapping presented here, including times consumed, matching between k-means and the pseudo-classification used, and number of mismatches between

k-means clustering and projection. Times were measured using a AMD Athlon 2.16 GHz processor, 512 Mbytes RAM.

Corpora	Entries	Dimension	Time(s)		K-means matches	Number of mismatches	
			Fastmap	NNP		Fastmap	NNP
CBR+ILP+IR	574	5495	11.609	11.484	98.61%	5	6
CBR+ILP+IR+SON	682	6371	18.703	18.781	91.94%	20	26
KDViz	1624	9398	165.844	159.546	90.83%	45	52
InfoVis04	534	4287	8.266	-	-	-	-

Table 1. Summary of details on data sets and their mappings

As far as algorithm complexity goes, FASTMAP is $O(log(n))$, NNP is $O(n^2)$. Force is $O(n^2)$, but, due to the initial point placement provided by the projections, the number of iterations is actually quite small compared to conventional force based placement strategies, rendering the whole process in fact quite reasonable in terms of time.

The above maps generated by IDMAP were compared with plots produced by projections using dimensional reduction via LSI .[15, 16] In terms of quality, LSI maps with reduction to 2 tend to separate well up to three classes of texts, with slightly less outliers. However, in the CBR+ILP+IR data set, when the class SON was added to the previous 3-subject corpus, there was a mix between SON and IR documents that forced a reduction to 5 followed by projection. Comparatively, IDMAP has separated well the 4-subjects corpora. Also, computational times for LSI based maps ran in the order of minutes for this data set (30 minutes for the KDVis data set).

To compare our results with direct clustering of documents, we have used, as illustrated here, k-means. In this case (as in clustering in general), the result is pockets of documents with no distinction within the pocket or in their neighborhoods. Also, frontiers of similar areas are not found by clustering. Plots by similarity or individual exploration are not naturally available by processing these data by clustering only. Rather, clustering and IDMAP are complementary techniques to each other.

Compared to other knowledge domain plots,[22] the type of information presented by IDMAPs is quite complementary to those presented by co-citation and bursts graphs, presenting an alternative view. Also, most of the pre-processing here can be made automatic from text files, allowing faster production of the final mapping.

5. CONCLUSIONS AND FURTHER WORK

In terms of usability, generation and separation of groups of documents and speed for exploration of texts, IDMAP compares well with other techniques. It complements those in the sense that it can generate maps that help explore structure and relationships in document collections inter and intra-groups and allow exploration of those collections in various levels (from overview to individual browsing).

In itself it presents a novel, unique form of document map generation, to support information gathering from groups of texts without the need for extensive individual document examination.

A re-engineering of the projection improvement part of the algorithm is currently under way to speed up the force calculations in a similar form as that employed by Chalmers[5] in order to test the approach for really large data sets (not the original goal here, but it will still be tested in that set up).

We also plan to add semantic levels to the display and multiple-view setups employing other available document visualization techniques.

A software system itself is to be developed and made available as a Web system for general use. All the pre-processing tasks that were performed are possible to be made automatic, and this is also in our plans.

The mapping process can be applied, from the distance matrix, to any data set that can be expressed that way, so the system will encompass data visualizations from other sources other then text.

We are working on other types of similarity metrics between texts to extend the flexibility and usability of IDMAP.

Acknowledgments

This work is funded by FAPESP research financial agency, São Paulo, Brazil (proc. no. 04/09888-5,04/07866-4). We wish to acknowledge the work of our undergraduate and research students as well as research colleagues in processing some data and discussing various issues of the work.

REFERENCES

1. O. Alonso and R. Baeza-Yates, "Alternative implementation techniques for web text visualization," in *Proc. of the First Latin American Web Congress*, pp. 202–203, IEEE Computer Society, IEEE Press, (Santiago, Chile), November 2003.
2. A. Leuski and J. Allan, "Lighthouse: Showing the way to relevant information.," in *InfoVIS*, pp. 125–130, IEEE Computer Society Press, 2000.
3. M. Sebrechts, J. Cugini, S. Laskowski, J. Vasilakis, and M. Miller, "Visualization of search results: A comparative evaluation of text, 2d, and 3d interfaces," in *22nd ACM-SIGIR Conf. Research and Development in Information Retrieval*, pp. 3–10, ACM Press, 1999.
4. R. Baeza-Yates, "Visualizing large answers in text databases," in *Int. Workshop on Advanced User Interfaces (AVI'96)*, pp. 101–107, ACM Press, (Ubbio, Italy), 1996.
5. M. Chalmers, "A linear iteration time layout algorithm for visualising high-dimensional data," in *Information Visualization 1996*, pp. 127–132, IEEE CS Press, (San Francisco - CA, USA), 1996.
6. N. Miller, P. Wong, M. Brewster, and H. Foote, "Topic islands - a wavelet-based text visualization system," in *Proc. of the conference on Visualization '98*, pp. 189–196, IEEE Computer Society, IEEE Computer Society Press, (Research Triangle Park, North Carolina, United States), 1998.
7. R. Rohrer, D. Ebert, and J. Sibert, "The shape of shakespeare: Visualizing text using implicit surfaces," in *Proc. the IEEE Symposium on Information Visualization*, pp. 121–129, IEEE Press, 1998.
8. A. Booker, M. Condliff, M. Greaves, F. Holt, A.Kao, D. Pierce, S. Poteet, and Y.-J. Wu, "Visualizing text data sets," *Computing in Science and Eng.* **1**(4), pp. 26–35, 1999.
9. E. Weippl, "Visualizing content based relations in texts," in *Proc. of the 2nd Australian conference on User interface*, pp. 34–41, IEEE Computer Society, IEEE Computer Society, (Queensland, Australia), 2001.
10. S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "ThemeRiver: Visualizing thematic changes in large document collections," *IEEE Transactions On Visualization And Computer Graphics,* **8**, pp. 9–20, Jan-Mar 2002.
11. J. Wise, "The ecological approach to text visualization," *Journal of the American Society for Information Science* **50**, pp. 1224–1233, November 1999.
12. 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*, pp. 442–450, Morgan Kaufmann Publishers Inc., (San Francisco, CA - USA), 1995.
13. G. Salton, "Developments in automatic text retrieval," *Science* **253**, pp. 974–980, 1991.
14. S. Kaski, T. Honkela, K. Lagus, and T. Kohonen, "WEBSOM - self-organizing maps of document collections," *Neurocomputing* **1**(1-3), pp. 110–117, 1998.
15. T. K. Landauer, D. Laham, and M. Derr, "From paragraph to graph: Latent semantic analysis for information visualization," *Procedings of the National Academy of Sciences* **101**(suppl. 1), pp. 5214–5219, 2004.
16. A. Lopes, R. Minghim, V. Melo, and F. Paulovich, "Mapping texts through dimensionality reduction and visualization techniques for interactive exploration of document collections," in *IST/SPIE Symposium on Electronic Imaging - Workshop on Visualization and Data Analysis*, (San Jose, California), 2006.
17. S. Huang, M. Ward, and E. Rundensteiner, "Exploration of dimensionality reduction for text visualization," Tech. Rep. TR - 03–14, Worcester Polytechnic Institute, Computer Science Department, 2003.
18. M. Carey, D. Heesch, and S. Ruger, "A visualization tool for document searching and browsing," in *Proc. of Intl. Conf on Distributed Multimedia Systems*, 2003.

19. M. Rasmussen and G. Karypis, "gCLUTO - an interactive clustering, visualization, and analysis system," Tech. Rep. CSE/UMN TR 04-021, Univ. of Minnesota, Dep. of Computer Science and Engineering, 2004.
20. S. Iritano and M. Ruffolo, "Managing the knowledge contained in electronic documents: a clustering method for text mining.," in *12th International Workshop on Database and Expert Systems Applications (DEXA) Workshop*, pp. 454–458, IEEE Computer Society Press, 2001.
21. M. Granitzer, W. K. V. Sabol, K. Andrews, and W. Klieber, "Evaluating a system for interactive exploration of large, hierarchically structured document repositories," in *Information Visualization 2004*, pp. 127–132, IEEE CS Press, (Austing- TX, USA), 2004.
22. K. Borner, C. Chen, and K. Boyack, "Visualizing knowledge domains," *Annual Review of Informtion Science & Technology* **37**, pp. 1–51, 2003.
23. E. Tejada, R. Minghim, and L. Nonato, "On improved projection techniques to support visual exploration of multi-dimensional data sets," *Information Visualization Journal* **2**(4), pp. 218–231, 2003.
24. M. Chalmers, "Using a landscape methaphor to represent a corpus of documents," in *Proc. of COSIT '93*, A. Frank and I. Campari, eds., *Lecture Notes in Computer Science* **716**, pp. 377–390, Springer, 1993.
25. M. Chalmers and P. Chitson, "Bead: Explorations in information visualization," in *Proc. of ACM SIGIR*, pp. 330–337, ACM Press, 1992.
26. C. Faloutsos and K. Lin, "Fastmap: A fast algorithm for indexing, datamining and visualization of traditional and multimedia databases," in *Proc. of International Conference on Management of Data*, pp. 163–174, ACM Press: New York, (San Jose-CA, USA), 1995.
27. M. Porter, "An algorithm for suffix striping," *Program* **14**(3), pp. 130–137, 1980.
28. H. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development* **2**, pp. 159–165, 1958.
29. G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Management* **24**(5), pp. 513–523, 1988.
30. K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is nearest neighbor meaningful?," in *Lecture Notes in Computer Science*, **1540**, pp. 217–235, 1999.
31. H. Edelsbrunner, *Geometry and Topology for Mesh Generation*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge, 2001.
32. R. Minghim, H. Levkowitz, L. G. Nonato, L. Watanabe, V. Salvador, H. Lopes, S. Pesco, and G. Tavares, "Spider cursor: A simple verstile interaction tool for data visualization and exploration (to appear)," in *Proceedings of GRAPHITE05*, ACM Press, (Dunedin, New Zeland), 2005.
33. J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceeedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, University of California Press, 1967.