UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação
ISSN 0103-2569

Creating Interactive Document Maps Through Dimensionality Reduction and Visualization Techniques

Alneu de Andrade Lopes
Rosane Minghim
Vinícius Melo

Nº 259

RELATÓRIOS TÉCNICOS

São Carlos – SP
Jun./2005

Creating Interactive Document Maps Through Dimensionality Reduction and Visualization Techniques

Alneu de Andrade Lopes Rosane Minghim Vinícius Melo

Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo, São Carlos – Brazil
{alneu, rminghim,vmelo}@icmc.usp.br

Abstract. The current availability of information many times impair the tasks of searching, browsing and analysing information pertinent to a topic of interest. This paper presents a methodology to create a meaningful graphical representation of corpora of documents targeted at supporting exploration of correlated information. The purpose of such an approach is to produce a map from a document body on a research topic or field based on the analysis of their contents, and similarities amongst articles. The document map is generated, after text pre-processing, by projecting the data in two dimensions using Latent Semantic Indexing. The projection is followed by hierarchical clustering to support sub-area identification. The map can be interactively explored, helping to narrow down the search for relevant articles. Tests were performed using a collection of documents pre-classified in three research subject classes: Case-Based Reasoning, Information Retrieval, and Inductive Logic Programming. The map produced was capable of separating the main areas and approaching documents by their similarity, revealing possible topics, and identifying boundaries between them. The tool can deal with the exploration of inter-topics and intra-topic relationship and is useful in many contexts that need deciding on relevant articles to read, such as scientific research, education, and training.

Introduction

One of the challenges for today's researchers and students is to deal with the huge amount of information available in databases and on the Internet. Even when the information is pre-filtered, i.e., when the collection of possibly relevant documents is narrowed down to a few hundred, the final mining process for relevant documents to examine relies very much on the user's expertise. There are very few alternatives to help make sense of the available materials for less experienced users. In general, it is not the purpose of commercial information systems either to analyse or to establish relationships between contents of these retrieved documents. Nor do they produce a representation that helps to locate documents of actual relevance.

The semantic web project has been proposed to ease information retrieval from the web, [2], and techniques for corpus visualization have been developed to help the user understand document relevance regarding a particular topic and its related sub-topics. Information Visualization and Visual Data Mining [25] techniques can help handle the huge amount of information available today [11]. Techniques from these fields

can support the user to get a broader view of the field and its frontiers with related areas, and also to get information on a set or sample of papers. WEBSOM [8], for instance, is a method based on the Self-Organizing Map (SOM) for organizing documents in meaningful maps for exploration and search [13]; other clustering approaches can also be employed for this task [10]. These are fine techniques in support of grouping similar information. However, their results do not aggregate much more understanding and exploration facilities than a list of ranked documents provides. Also, they do not allow analysis of relationships amongst individual elements; only groups, themes or subjects are distinguishable. To explore in more detail a set of focus documents, other techniques must be made available.

In that context this work presents a methodology to build a topographical map of a document collection that reflects various levels of similarity relationships between document contents, and a possible organisation of areas and sub-areas related to the user's query. It is meant to support researchers and general users to organize and explore documents related and relevant to some user quest, helping select what texts to examine and helping create a mental model from the retrieved data. The map is content-based, taking as input information extracted from the texts (titles, terms from abstracts, year of publication, authors, and references). The resulting map can be interacted with to allow the user to explore the relationships and properties inter-topics (between clusters) and intra-topic (between document pairs in a found cluster).

The map is created by employing Latent Semantic Indexing to generate a projection of the documents. Following that, a number of visual (and aural) attributes are employed to add further levels of information to the map. Various attributes can be mapped, such as relevance, number of citations, year of publication and labels. The aim here is not to deal with extremely large corpora; other techniques exist for that. Instead, we look at in-depth exploration of a collection of pre-filtered documents by dimensionality reduction and visualization techniques, complementing other existing techniques for exploration of text collections.

The tests were performed on a collection of three corpora of scientific texts. Two of them were generated from the Lecture Notes on Artificial Intelligence (Case-Based Reasoning – CBR – 277 articles, and Inductive Logic Programming – ILP – 119 articles), and the other corpus was retrieved from the web by a search on the subject of Information Retrieval – IR (204 articles). To be able to analyse maps, we pre-classified the papers purely according to their sources, assigning labels to the articles as CBR, ILP or IR. The results are very stimulating for the goal of building a comprehensive tool for exploration of document sets.

In the next section, we present some background in Information Retrieval and previous work related to corpus visualization. In Section 3 we describe the LSI (*Latent Semantic Indexing*) and argue for using it as dimensionality reduction method. In Section 4 we present our method for representing text data comparing the results obtained with partitioning clustering and SOM approaches. Finally, in Section 5 we present our conclusions and discussion of future work.

2. Background work

Usually, in Information Retrieval (IR) systems, the user enters a query describing the desired information and the system returns a list of documents that satisfy the query expression. Much of the recent work in the subject has focused on systems that rank documents according to their estimated relevance to a query. Various techniques used in IR rely on Machine Learning algorithms such as documents clustering [23], text categorization [9], and inductive techniques or rule-based approaches for information extraction using a training corpus [14, 23, 3, 7, 18].

Those techniques cannot be applied directly to textual data. Text, represented by the terms deemed useful for the IR purpose – known as *bag-of-words* – must be converted into vectors. Although the conversion process is very efficient in terms of speed, the dimension of the resulting vector space is usually high because each term that was not eliminated or grouped is transformed into an attribute [6]. These attributes are weighted according to the term importance, producing a matrix of *documents x terms*, where the elements weights are the frequency count (*tf*) of a term in a document. The most common weight is referred to as *tfidf – term frequency, inverse document frequency*. Tfidf, given by $tf*idf$, where $idf_{(ti)} = \log(N/n_i)$, assumes that there are N documents in the collection, and that term t_i occurs in n_i of them. [9, 16].

Diverse dimension reduction techniques can be used to reduce the set of terms. The majority of the dimension reduction techniques can be divided in two categories: selection of characteristics (*Document Frequency Thresholding, Information Gain,* and *Chi-Square* [26]) and re-parameterization. Re-parameterization is the process of generating new characteristics by combining or transforming the original characteristics. One re-parameterization method, used in this work, is LSI [21, 5]. In section 3 we present details of this method.

In order to support analysis of the results of such techniques, a number of visualization tools exist that draw visualizations of a document, sets of documents, or web retrieved results. Most techniques are meant to classify large numbers of documents in groups, but fail to show levels of association amongst individual text contents. We believe that showing further relationships is paramount in determining performance improvement for users to analyse new sets of texts or to find new trends within a text collection. We refer to the work of Katy Borner and others [4] for a detailed description of available text visualization techniques.

Two particularly important techniques for the target of this work are SOM and cluster visualizations. SOMs (used in WEBSOM) [12, 13, 8] are competitive Artificial Neuron Networks (ANN) capable of producing maps that try to preserve the distribution between the input and output spaces, mapping high dimensional vectors to a 2D or 3D map. Another approach for cluster visualization is gCLUTO. gCLUTO is a graphical application built on-top of the CLUTO's[1] clustering library, providing tools for visualizing the resulting clustering solutions using mountain (landscape) visualization. The CLUTO framework provides multiple clustering algorithms and similarity/distance functions [10]. The main problem of these text visualization approaches in reaching our goals is that, while they can generate usually effective

[1] http://www-users.cs.umn.edu/~karypis/cluto/

data groupings and meaningful displays, they do not distinguish relationships amongst individual documents. In other words, they are very good for overview, but not for detail exploration. We expect to exemplify that behavior in Section 4 when comparing the results of our approach to results of both SOM and gCLUTO.

In the next section we describe LSI, the underlying technique of our approach, and tests performed to evaluate its application to determine its capability for representing text similarity.

3 Latent Semantic Indexing

Latent Semantic Indexing (LSI) applies singular value decomposition (SVD) to the *term-document* matrix in order to get a projection of the documents and words in a space referred to as the *latent space*, with lower dimension than the original space. In the projected space, co-occurring dimensions (terms) are reduced to a single dimension. After LSI the terms are usually reduced from thousands to hundreds [16].

Using an artificial set of documents, Papadimitrious and others [21] have shown that the LSI space preserves, to the extent possible, the relative distances in the term-document matrix while projecting it to a lower-dimensional space, but a formal proof is not achieved yet. They measured the angles between pairs of texts in the original and in the LSI spaces, considering pairs with texts belonging to the same topic (intra-topic pairs) and belonging to different topics (inter-topic pairs). These measurements have shown the dramatic reduction of the angles of intra-topic pairs, in the LSI space. This reduction suggests the similarity between the intra-topic pairs. In the next section we show, by classification, that the LSI space also preserves the similarities between intra-topic documents. In the case of document clustering, Lerman shows that there is an improvement of quality of clusters when LSI is used prior to clustering, if the optimal dimension for reduction is chosen [15].

In this work LSI is used to reduce the dimensionality of the original document vector in such way that they could be plotted on a two-dimensional plane.

3.1 Dimensionality Reduction by LSI

LSI is actually the application of (SVD) to the matrix *term-document*. Let A be such matrix, its decomposition results in three matrices U, S and V that satisfy:

$$A = U \times S \times V^T \quad (1)$$

The matrices S and V are used to calculate the matrix B that contains the vectors in the reduced space, $B = S \times V^T$. The matrix U can be used to transform a vector b from the reduced space to the original space, $a = U \times b$, or to transform a vector from the original space to the reduced space, $b = U^T \times a$, the matrix U is orthonormal, so $U^{-1} = U^T$.

In the classification task, to obtain the test vectors in the reduced space we have used the procedure described in the previous paragraph, $b = U^T \times a$, where a is a test vector created after the pre-processing and b is the corresponding test vector in the reduced space, obtained with the training set.

It is important to notice that the dimensions found with the LSI are "ordered" by relevance – the best representation of the data using n attributes is obtained using the first n dimensions found. The concept of better representation used here is purely mathematical (it is the distance between the matrices calculated using the 2-norm).

3.2 Similarity Analysis between Documents by LSI

In order to evaluate the precision of the articles representation as vectors by theirs latent-space dimensions, we compared the classification results obtained by the KNN learning method with 10-*fold cross validation*, using both representations (the document as an original vector in a n-dimensional space and in the reduced latent-space). The division between training and testing was made in two ways: pre-processing and LSI were applied before (P1-the optimistic accuracy) and after (P2-the realistic accuracy) the corpus division. In Figure 1 we show the result comparing the precisions (P1 and P2). The standard deviations (*SD1* and *SD2*) were also compared. The tests were carried out with dimensions ranging from 1 to 600.

As shown in Figure 1, in which x axis represents the dimension of the vectors and the y the precision, the articles represented by the LSI applied after division of corpus were satisfactorily classified, particularly when represented with a number of dimensions between 2 and 5.

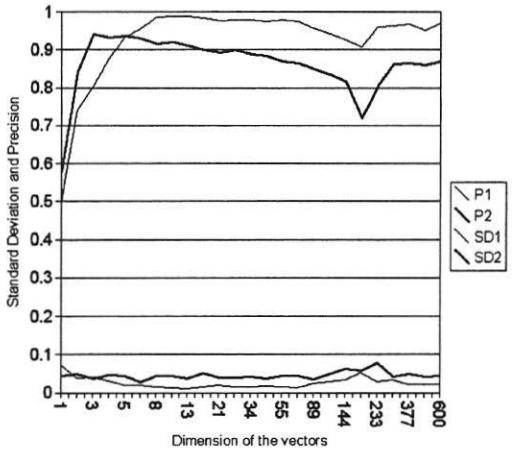


Fig. 1. Classification accuracies (P1, P2) and standard deviations (SD1, SD2) using LSI.

Thus, we can presume that LSI projection preserves the similarity between articles represented by its LSI dimensions.

Finally, if each article is represented by a vector with two dimensions, it is possible to obtain a map of the articles by drawing them on a plane, each article represented by its coordinates using the two final LSI dimensions. This test was performed using the original corpora (CBR, ILP, and IR), creating the vectors from the bigrams. Figure 4a shows that LSI projection to 2D actually distinguishes the different article groups, with few misrepresentations.

4 Our Corpus Visualization Approach

The output of dimensionality reduction by LSI with dimension reduced to 2 is a set of points on a plane so that their distances reflect some structure based on similarity between document contents. Our approach is to use that point placement (see Figure 2a) as basis for a landscape view of the corpus.

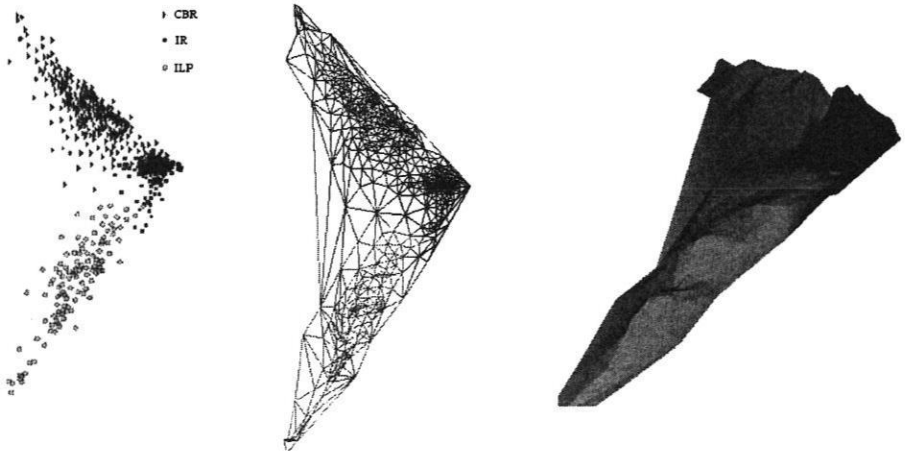
The projected points can be mapped to a surface where their relationships can be observed and explored particularly regarding neighbourhoods. In order to achieve a surface representation, a Delaunay triangulation [17] of the points was performed. The technique is nearly linear in time. Figure 4 b) shows the triangulation of the articles.

From the surface approximation given by the triangulation, it is possible to construct various types of graphical outputs taking the surface topology as a basis and assigning to each point scalar values. Those values can be mapped to visual properties, such as height and color. For instance, if the value represented at the points reflects the relevance of the paper, the most important articles could be shown at the top of a "hill" (or with a particular color).

The corpus projection aims at reflecting similarity by the proximity between documents in the projection plane. By doing so, it is bound to define regions of similar articles. Sub-groups within those groups should also form. In order help locate such groups and sub-groups, a hierarchical clustering (HC) [20] was performed and the resulting cluster values (their depth in the dendogram) were mapped to visual attributes, revealing possible sub-areas. These cluster levels take into account the fact that in the dendogram generated by the HC the depth of each node is proportional to the value of the inter-clusters distance ($dist(a,b)$) between its two daughters a and b. Given max_dist the distance between the two most different clusters, we can use, for instance, the measure ($d = max\_dist - dist(a,b)$) as a third coordinate, plotting it as height or colour in the map. That enables exploring groups on top of the surface. In this case, on the peaks of such a "topographical" representation one would find the most similar articles, while at the valleys one would find the nodes (articles) more likely to be dealing with hybrid approaches, applications of a technique in another field, or that had less similarity with their neighbours. That height map is presented in Figure 4c.

The same information (HC depth) can be redundantly mapped to colour; in that case the hill tops will have the same colours, the next similarity level another colour and so forth (Figure 4d). Note that the hill tops (most similar texts) are highlighted. With that scalar field assigned to vertices on the surface, isolines can help distinguish regions of similar HC depth (Figure 4e). That helps defining boundaries between sub-groups of articles. The combination of these visual mappings provides the user with a more complete map (see Figure 4f).

Naturally, cluster depth is not the only interesting scalar value do draw the visualizations. To each point in the triangulation it is possible to associate a number of different attributes related to the text that can help the user gain insight into the data sets.



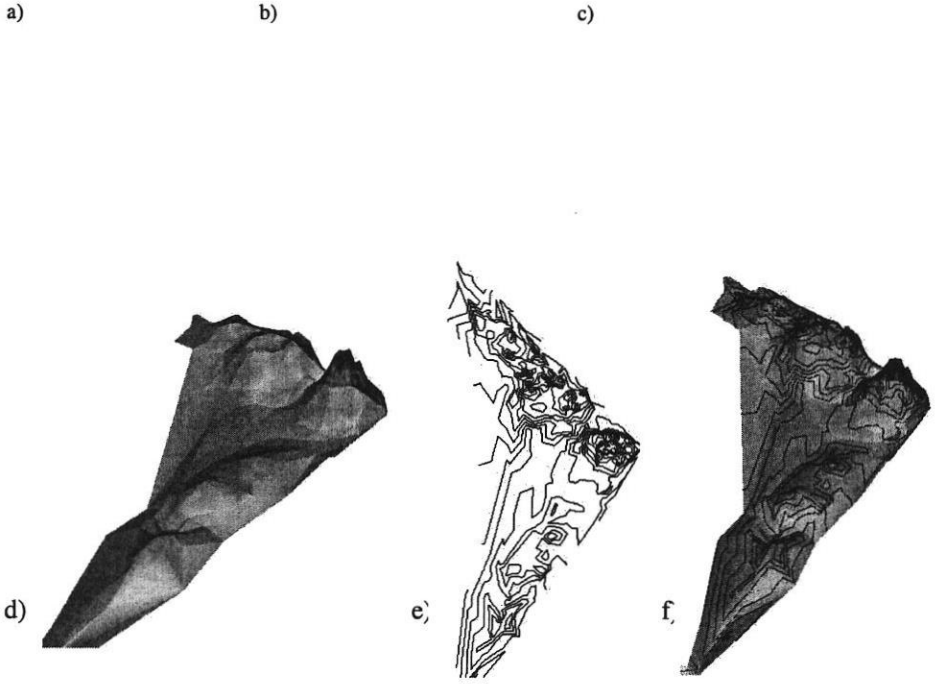


Fig. 2. Document maps. a) Projection of the articles; b) Triangulation of b). c) Colour coding pseudo-classes; height coding HC depth; d) Color and height determined by HC depth. e) Contour lines based on clustering depth; f) Combined view of d) and e).

To gain further insight into the value of the technique, we compared our results with applying the same data set to SOM and gCLUTO.

SOM results were obtained by clustering our corpora using the SOM_PAK tool. SOM_PAK [12], freely available at the developer's site[2], is a set of tools for using SOM networks. The experiment configuration employed an 8x8 map, with the weight vectors started randomly. The training was made in two steps: 1000 epochs for the first stage and 10000 for the final stage. After that, using the Umat tool, provided by the SOM_PAK, the u-matrix visualization was created.

The resulting map is shown in Figure 3. It displays a very compact grouping of the IR class texts (3 neurons correspond to 133 of the 204 documents) and 4 documents were separated from the others (IR 1/1). That characteristic makes it difficult to define sub-areas or other relationships within a cluster. The ILP class is a little better separated and the CBR class is very dispersed on the map, implying that there is more difference amongst those documents.

We also produced a larger map (20x12), in the attempt of getting a better visualization of the situation inside each initial class. In that map, we noticed a large amount of nodes with only 1 document, and blank nodes (without any document) in

[2] http://www.cis.hut.fi/research/som_pak/

the area that should be occupied by the IR class. So, the attempt was not successful in distinguishing trends within groups. WEBSOM's major drawback is the amount of training time and resources required for the fine-tuning of the document collection.

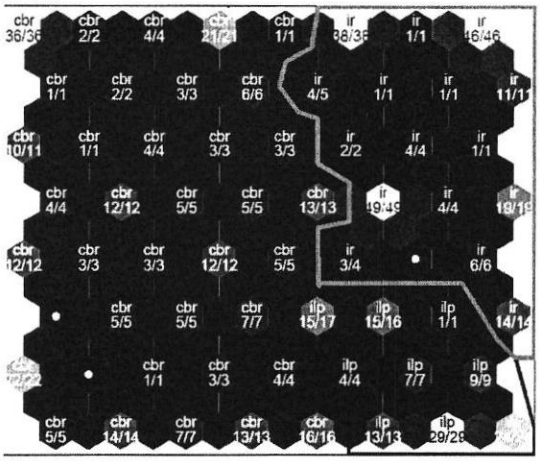


Fig. 3. SOM map 8x8

Compared to LSI maps, SOM: 1) does not offer indication of similarities amongst documents in a cluster. 2) demands extension and retraction of map size to allow investigation of nodes that are too dense. Additionally, the U-matrix tool does not provide additional levels of information.

Also, comparing these maps with the maps created by the LSI (Figure 2), there is a large difference in the way the two algorithms cluster the data. In our map there is a well defined but slightly dispersed CBR area (red in Figures 4b and 4c), a very compact IR area (in blue), and two adjoining ILP areas (in green). On the maps produced using SOM there is a more compact ILP area, a larger CBR area, occupying about of half of the map, and an IR area very dispersed and little defined.

gCLUTO was run over the data set with its default configuration, which uses a partitioning method called RB. In this method the solution is computed through a sequence of k-1 *repeated bisections* (k is the number of final clusters, given as input). The initial database is partitioned in two clusters. Then, one of these clusters is selected and partitioned. The process continues until the desired number of clusters is reached. In each step, the partitioned cluster results in two new clusters that locally optimise a specific criterion function.

gCLUTO's first drawback in a generic task for document maps is having to inform the k value, not usually known in advance. In our experiments, we chose k=7 to lead RB to divide the documents in sub-areas. After the execution of the gCLUTO with the default configuration (method RB and criterion function $i2$), with k=7, we used the tool to generate the mountain visualization shown in Figure 4.

In that map, it can be seen that the ILP class was kept almost intact in cluster 1 (109 documents). The CBR class was partitioned between clusters 5 and 6 (114 and 154 documents, respectively). The IR class was divided into the clusters 0, 2, 3 e 4

(17, 35, 35, and 82 documents, respectively). In these hills, the height reflects the similarity amongst documents and the surface area reflects the quantity of documents in the cluster. The map produced by gCLUTO provides a good division of the document set. Besides, the terms assigned to the clusters in these corpora are good clues about theirs contents.

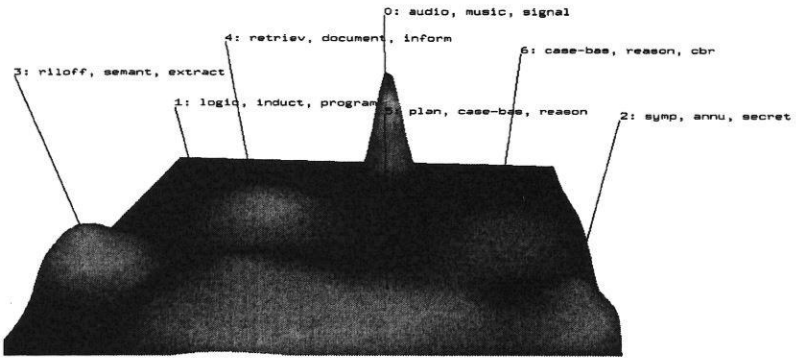


Fig. 4. Map using gCLUTO with RB, I2 and k=7.

Nevertheless, the hills do not give enough information on the structure inside a cluster. Also, the positioning of the "sub-areas" is not properly plotted. For instance, there is a CBR mountain (6) between the IR sub-areas (0 and 2), and part of CBR cluster (5) was merged with ILP (1). Moreover, there is no access to information regarding an individual article (its importance, for instance).

Summarizing, both techniques used for comparison have shown different views of the data set. Compared to LSI maps, they produce a different level of information and the results point towards simultaneous use of all of them to make sense of a particular document collection.

To evaluate the results of the LSI projection further we have employed an interaction tool developed by members of our research teams called spider cursor [24], that allows navigation on the surface and identification of individual elements. A point vertex on the mesh is located using the mouse, and the connection to its neighbours is highlighted. Figure 5 illustrates the spider cursor and the results of the exploration using this tool.

As shown in Figure 5, inter and intra-cluster relationships are represented on the map. The documents in the valley between CBR and ILP mountains, for instance, deal with hybrid case and rule-based approaches. In the border between IR and CBR lie articles dealing with CBR applied to text processing. In the CBR cluster one can see two mountains, the first, larger, in a volcano shape, and the second, smaller. Analysing the articles shown on the peak of the larger mountain all articles refer to CBR applications, and on the top of the smaller mountain the articles deal with methodologies and techniques. We have not reached a conclusion on the division of the ILP group in two mountains from the sample articles analysed.

Interaction with the spider cursor is improved by sound mappings, that can reflect different quantities in terms of different pitches as the cursor moves over the points. During interaction that is useful to dissipate interpretation ambiguities, particularly

due to perspective misconceptions. Sound mappings can help guide the user during analysis.

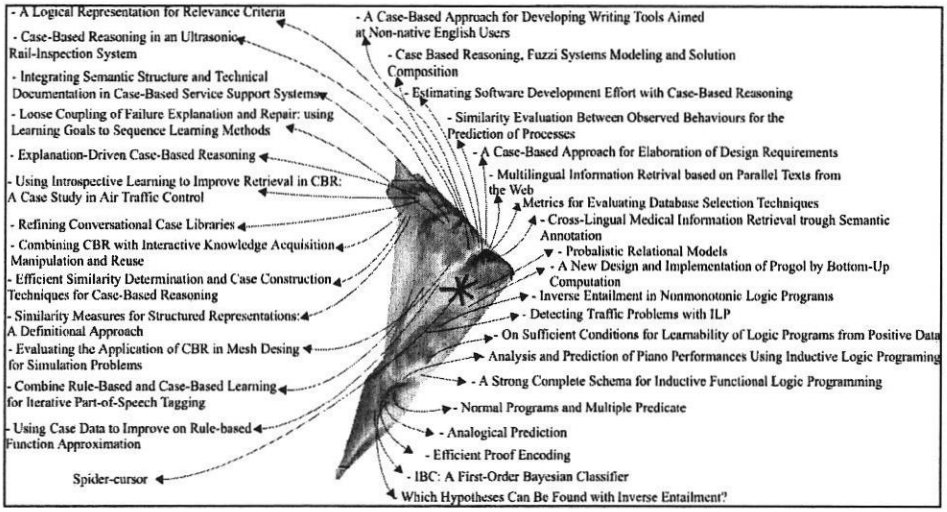


Fig. 5. Result of analysis after interaction with the resulting map using the spider cursor.

6 Conclusion and Future Work

In this work we have presented a new methodology for text information visualization by dimensionality reduction and hierarchical clustering that overcomes typical limitations of the well known text visualization techniques in the exploration of the inter and intra-clusters relations.

gCLUTO clustering approach shows a map that provides clues on the fields or topics related to the documents, as well as the documents belonging to it, but the exploration of the intra-cluster relations are not provided. Even the inter-cluster relations are not properly reflected in the map. Its purpose, as well as SOM's, is to gather similar documents in the same cluster. The possibility of a hierarchical presentation of the SOM map may be useful for further exploration, but the available graphical representations are poor in added attributes. Empirical choices that reflect highly on the results impair the use of both techniques for generic data sets (e.g., pre-defined number of clusters in gCLUTO, and map size, number of epochs, learning rates and others in SOM).

In this work we have proposed an approach that overcomes those limitations, by showing intra- and inter-clusters relationships as well as overview of the document collection. Our tests of formation of LSI maps using every two of the three mentioned areas of research, as well as two of them plus a different area (sonification), all resulted in maps that maintained the same properties of the map presented here, indicating a consistency not subject to empirical parameter settings.

The map obtained reflects similarity relationships between documents, and a possible organisation of areas and sub-areas within groups. Frontiers are also satisfactorily defined.

The spider cursor tool is being extended to allow selections of attributes to be mapped to the various visual and aural attributes of the maps and to allow hierarchical views. The pre-processing steps, although manually realized in the context of this work, can be clearly automated.

In order to deal with the computational complexity of the dimension reduction by LSI, which makes it difficult to treat scalability of the problem (to tens of thousand or millions of documents), we are approaching the use of other faster mapping strategies to produce maps with similar quality as the one found here. Since the complexity of LSI is due to the calculations of SVD, another alternative we are seeking is to implement a fast approximation of SVD [5].

We are also looking into ways of defining similarity between texts without the need for the vector representation, in search for a map that can be produced incrementally, a feature that LSI does not possess.

References

1. Banerjee, S.; and Pedersen, T. (2003). The design, implementation and use of the ngram statistics package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City.
2. Berners-Lee, T.; Hendler, J.; and Lassila, O. (2001). The semantic web. *Scientific American*, 284(5):34–43.
3. Borkar, V.; Deshmukh, K.; and Sarawagi, S. (2001). Automatic segmentation of text into structured records. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, pages 175–186, Santa Barbara, California.
4. Borner, K.; Chen, C.; and K. Boyack (2003). Visualizing knowledge domains. *Annual Review of Informtion Science & Technology*, 37:1–51.
5. Drineas, P.; Frieze, A.; Kannan, R.; Vempala, S.; Vinay, V. (2004). Clustering Large Graphs via the Singular Value Decomposition. Machine Learning, 56, 9-33.
6. Ebecken, N.; Lopes, M.; Costa, M. (2003). Mineração de Textos, Sistemas Inteligentes: Fundamentos e Aplicações, Manole, p. 337-370.
7. Geng, J. and Yang, J. (2003). Automatic extraction and integration of bibliographic information on the web. Technical report, Department of Computer Science, Duke University.
8. Honkela, T.; Kaski, S.; Lagus, K.; and Kohonen, T. (1996). Newsgroup exploration with WEBSOM method and browsing interface. Technical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland.
9. Joachims, T. (1997). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, Machine Learning: Proceedings of the 14th International Conference(ICML 97), p. 143-151.
10. Karypis, G. (2002). CLUTO a clustering toolkit. Technical Report 02-017, Department of Computer Science, Un. of Minnesota. Available at http://www.cs.umn.edu/~cluto.
11. Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1).
12. Kohonen, T.; Kaski, S.; Lagus, K.; and Honkela, T. (1996). Very large two-level SOM for the browsing of newsgroups. In von der Malsburg, C., von Seelen, W., Vorbrüggen, J. C.,

and Sendhoff, B., editors, *Proceedings of ICANN96, Bochum, Germany,* Lecture Notes in Computer Science, vol. 1112, p. 269-274.
13. Kohonem, T.; Kaski, S.; Lagus, K.; Salojärvi, J.; Paatero, V.; Saarela, A. (2000) Organization of a Massive Document Collection. IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Minining and Knowledge Discovery. 11(3), pp.574-585.
14. Lawrence, S.; Giles, C. L.; and Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71.
15. Lerman, K. (1999). Document clustering in reduced dimension vector space. Technical report, USC Information Sciences Institute.
16. Manning, C. D. and Schütze, H. (2001). *Foundations of statistical natural language processing*. MIT Press.
17. Maus, A. (1984). Delaunay triangulation and convex hull of n points in expected linear time. *BIT*, 24(2):151–163.
18. Melo, V.; and Lopes, A. A. (2004). Efficient identification of duplicate bibliographical references. In Logic Applied to Technology, Japan, pp. 1- 8.
19. Mitchell, T. (1997). *Instance-based learning*, chapter 8. McGraw-Hill.
20. Murtagh, F. (1984). A survey of recent advances in hierarchical clustering algorithms which uses cluster centers. Computer Journal (26), pp.354-359.
21. Papadimitriou, C. H.; Raghavan, P.; Tarnaki, H.; and Vempala, S. (1998). Latent semantic indexing: A probabilistic analysis. In *Proc. 17th ACM Symp. on the Principles of Database Systems*, pp. 159–168.
22. Peñas, A.; Verdejo, F.; and Gonzalo, J. (2001). Corpus-based terminology extraction applied to information access. In *Proceedings of the Corpus Linguistics*, volume 13, pages 458–465.
23. van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworths, London, 2 edition.
24. Salvador, V. C. L.; Minghim, R.; and Levkowitz, H. DSVOL II - A Distributed Visualization and Sonification Application. In*: XV Brazilian Symposium On Computer Graphics and Image Processing*, 2002, Fortaleza – CE, IEEE CS Press, 2002. v. 1, pp. 35-42.
25. Wong, P. C. (1999). Visual data mining. *IEEE Computer Graphics and Applications*.
26. Yang, Y.; Pedersen, J. (1997). A Comparative Study on Feature Selection in Text Categorization. In International Conference on Machine Learning.