Time-Aware Visualization of Document Collections

Aretha B. Alencar
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
Caixa Postal 668
São Carlos, SP, Brazil
aretha@icmc.usp.br

Fernando V. Paulovich
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
Caixa Postal 668
São Carlos, SP, Brazil
paulovic@icmc.usp.br

Katy Börner
School of Library and Information Science
Indiana University
10th Street & Jordan Avenue, Wells Library 021
Bloomington, IN, USA
katy@indiana.edu

Maria Cristina F. de Oliveira
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
Caixa Postal 668
São Carlos, SP, Brazil
cristina@icmc.usp.br

ABSTRACT

Scientific articles are the major mechanism for researchers to report their results, and a collection of papers on a discipline can reveal a lot about its evolution, such as the emergence of new topics. Nonetheless, given a broad collection of papers it is typically very difficult to grasp important information that could help readers to globally interpret, navigate and then focus on the relevant items for their task. Content-based document maps are visual representations created from evaluating the (dis)similarity amongst the documents, and have been shown to support exploratory tasks in this scenario. Documents are represented by visual markers placed in the 2D space so that documents close share similar content. Albeit the maps allow visually identifying groups of related documents and frontiers between groups, they do not explicitly convey the temporal evolution of a collection. We propose a technique for creating content-based similarity maps of document collections that highlight temporal changes along time. Our solution constructs a sequence of maps from time-stamped sub-sets of the data. It adopts a cumulative backwards strategy to preserve user context across successive time-stamps, i.e., maps do not change drastically from one time stamp to the next, favouring user perception of changes.

Categories and Subject Descriptors

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

General Terms

Information Visualization; Multidimensional Projections

Keywords

Time-varying Data; Text and Document Data

SAC 2012 March 26-30, 2012, Riva del Garda (Trento), Italy.

1. INTRODUCTION

The scientific literature available grows at a fast rate, with new results, methods and technologies being reported on a daily basis. This highly dynamic nature of science poses challenges for researchers and research policy makers to track the evolution of a certain discipline or research topic. Some visualization techniques and tools have been proposed specifically to support exploratory analysis of collections of scientific articles [2, 8, 20]. Most of these rely on network analysis – where units like authors, institutions, countries, words, articles and journals are represented as nodes and their relationships as edges of a complex network. In this work, we focus on the evolution of content and content similarity of the articles in a collection, rather than on authorship or citation networks, thus providing a complementary view to existing approaches.

Multidimensional projections and point placement techniques have been employed to generate global views of high-dimensional data sets that can be either embedded in metric space, or for which a matrix of pairwise distances may be computed [3, 16]. They work by mapping high-dimensional data on a low-dimensional visual space, typically 2D, while striving to lay out similar points close to each other. It has been shown that these techniques applied to document collections can generate insightful document maps that are suitable for visualization and intuitive exploration of the topical space addressed by that collection [15, 14, 12, 17]. Because they favor the perception of content similarity and dissimilarity, these maps allow visually identifying groups of highly related documents (addressing similar topics) and frontiers between groups, favouring identification of themes in general, as well as focusing and exploration on themes of interest.

Notably, considering a collection to which new documents are added gradually – as a series of papers from a conference, or the papers authored by a particular person over a certain time period – this topical space is not static along time, as topics rise and decay in the amount of scientific interest they generate. Albeit time plays an important role in many types of data, text inclusive, existing multidimensional projection techniques do not handle it explicitly. In this work we introduce a new temporally-oriented multidimensional projection technique called *Time-based Least Square Projection*, which constructs a sequence of similarity-based maps from time-stamped sub-sets of the data. The goal is to show views of a collection of multidimensional data emphasizing changes in the similarity patterns over time, while preserving global user con-

text. As a proof-of-concept, we illustrate its capabilities analyzing the research trajectory of individual scholars, based on their scientific production reported at ISI Web of Science. In this paper we detail results for a particular researcher.

In the next section we review previous contributions related to visualization of text collections and scientific literature, particularly those concerned with showing temporal evolution. Section 3 presents previous work and concepts required to understand the proposed approach. The full description of the process to generate a sequence of time-stamped document maps is presented in Section 4. Section 5 briefly considers some issues related to application of the proposed backwards cumulative projection strategy to scientific collections and compares it with a conventional multidimensional projection approach. Section 6 illustrates how the technique supports analysis of research trajectories, considering the scientific production of one particular researcher. That is followed by the conclusions and further work we intend to pursue. Results obtained for other scholars are available at `http://lcadfs2.lcad.icmc.usp.br/~aretha/timeaware`, and are not shown here due to space constraints.

2. RELATED WORK

A few visualization tools have been introduced aimed at facilitating user access and interpretation of text collections evolving over time. One simple yet interesting technique is *ThemeRiver* [7], aimed at displaying temporal thematic changes in a document collection by highlighting selected topics (themes). It adopts the visual metaphor of a 'river' that flows through time from the left to the right, where individual topics are represented as colored 'streams' within the river. The width of a flow indicates its strength, and the width of the river at a specific time depicts the collective strength of the selected topics. Analysis allows associating external events with major changes in the river. However, interpreting a large amount of topics is not easy: the visualization becomes cluttered, hampering user ability to discriminate among the areas and colors representing the topics. It is also difficult to relate topics with specific documents. Moreover, a topic is represented by a single term, a summarization that lacks discrimination power in more complex document collections. The choice of topics (terms) to be visualized is manual, based on the most frequent terms.

On a similar line, but with enhanced analytical strategies, TIARA (***T**ext **I**nsight via **A**utomated **R**esponsive **A**nalytics*) [22, 11] is visual text analytics tool also aimed at highlighting the temporal evolution of topics in a collection of documents. It employs automatic topic extraction techniques such as LDA (Latent Dirichlet Allocation [1]) to summarize texts into a set of topics and derive time-sensitive keywords to depict how topics evolve on collections of news or emails. The visual metaphor resembles that of *ThemeRiver*. In both tools relevant topics are identified first, and then a visual representation is created to convey a time-based visual summary. Again, resulting views do not immediately relate documents and topics. Both tools were conceived for visualizing text collections in general, therefore they do not consider specific properties of scientific articles, e.g., citation patterns or keywords.

The state of the art of a research field defines its research front [2]: a set of highly cited articles and the articles heavily cited by them that form the intellectual basis of the field. Scientists tend to cite the most recently published articles, and therefore the research front and intellectual base are characterized by a transient nature. The *CiteSpace II* tool [2] builds a visual representation that aims to show how research fronts and intellectual bases of a research area change over time and their transient patterns. Research front terms are identified using the Kleinberg's burst detection algorithm [9], which returns a ranked list of the most significant word bursts and the time interval when they occurred. The intellectual base is formed by groups of articles cited by articles in which research-front terms were found. The final visual representation is a hybrid network with three types of links: co-occurring research front terms; co-cited intellectual base articles; and research-front term citing an intellectual base article. The betweenness centrality metric over networks is employed to identify and highlight potential points of paradigm shift over time. Despite using Pathfinder network scaling to reduce the number of links shown at a time, the network remains very dense in most cases. Similarly to *ThemeRiver* and *TIARA* for general collections, *Citespace II* seeks to represent the global dynamics a science area through a single static visual representation.

The *Visone* tool [10] employs an MDS-based algorithm to layout time series of social network data dynamically by optimizing the stress both within the current network and over the previous and consecutive networks. This modified stress function penalizes drastic movements of a node from a network to the next. In this manner, the authors expect to promote stability along the networks and preserve the mental map between consecutive layouts. However, stability is achieved through a parameter in the stress function rather than being dictated by the data. The algorithm has been applied to different networks, including journal citation maps and heterogeneous maps composed by title words, authors, and journals. Despite using title words in their heterogeneous visualization, representation is not focused on topical events. This algorithm is $O(n^2|T|)$, where T is the number of networks (i.e., the number of time intervals) and n is the number of nodes in each network.

In this paper we propose generating views of a collection that highlight groups of similar, i.e. content-related articles, on a sequence of document maps capable of showing the temporal evolution of these groups. It differs from the above solutions in that it is aimed at gradually identifying the major topics addressed by a collection of scientific papers and highlighting their evolution, relating topics and articles in a straightforward manner. The maps focus on content, rather than on citation or co-authorship patterns, and has such provides a complementary approach to current visualizations.

3. BACKGROUND

Multidimensional projections provide a general framework for creating interactive visual representations of high dimensional data, which take advantage of the human visual ability to recognize structures or patterns based on similarity, such as clusters of elements. Many types of projection techniques exist, but all of them share the same underlying concept: data is projected from an m-dimensional space into a d-dimensional space with $d \ll m$ (typically, $d = 2$) while retaining, on the projected space, some information about distance relationships among the m-dimensional data items. Formally, let $X = \{x_1, \ldots, x_N\}$ be an m-dimensional dataset, with $\delta(x_i, x_j)$ being a dissimilarity measure between two m-dimensional data instances; and let $Y = \{y_1, \ldots, y_N\}$ be a set of points at the d-dimensional space, and $d(y_i, y_j)$ a distance (usually Euclidean) between two points on the projected space. A multidimensional projection technique can be defined as an injective function $f : X \to Y$ that seeks to make $\left|\delta(x_i, x_j) - d(f(x_i), f(x_j))\right|$ as close to zero as possible $\forall x_i, x_j \in X$ [21].

3.1 Least Square Projection

The Least Square Projection (LSP) [15] adopts a strategy different from most conventional projection techniques, in that it seeks to preserve local data neighborhoods identified in the original m-dimensional space, rather than global neighborhoods rela-

tions. Two major steps are involved in the LSP projection process. In a first step a subset of the dataset, called "control points", are projected onto $\mathbb{R}^d$ employing a highly precise projection technique known as Force Scheme [21]. In the second step, departing from the neighborhood relationships amongst the data instances in $\mathbb{R}^m$ and the Cartesian coordinates of the control points in $\mathbb{R}^d$ obtained in the first step, a linear system is constructed and solved to obtain the projected coordinates of the remaining points in $\mathbb{R}^d$.

These remaining instances are projected in the convex hull of their neighbors, while taking the control points as anchors to add geometrical information to the system. Further details on how this linear system is built and solved may be found elsewhere [15]. Because it seeks to preserve local neighborhoods LSP generates good similarity-based maps of data defined on sparse high-dimensional spaces, which is the case of text collections. It achieves excellent compromise between precision, measured in terms of neighborhood preservation, and computational cost, as discussed in [15].

3.2 Evaluating Similarity

Generating content-based maps of documents using multidimensional projection techniques requires an approach to assess how (dis)similar documents are. The input to such techniques is a matrix of pairwise dissimilarities amongst all documents. In generating this matrix some data preprocessing is typically required to improve projection accuracy. A common approach is to employ the vector space model [19], which represents each document by a vector of term frequencies. Given a set of N documents $D = \{d_1, \ldots, d_N\}$ represented by M terms $T = \{t_1, \ldots, t_M\}$, each document d_i is represented by a vector $d_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{iM})$, where the value of α_{ij} must represent the relative influence of a term t_j on a document d_i. One may adopt some weighting schemes to compute the value of α_{ij}: the *term frequency* (tf), the *term frequency-inverse document frequency* (tf-idf) and their normalized versions. In principle, the set of terms T consists of all the terms occurring in the collection, but it is usually reduced through some pre-processing steps, e.g., removing terms known as stopwords (adjectives, verbs, etc.) that add no discriminative power; applying a stemming algorithm such as *Porter's* [18] to reduce words to their radicals; and removing terms that occur too sparsely or too often and hence have limited differentiating capability through *Luhn's cut-off* values [13].

Once the vector space model of the collection is computed, the cosine between two vector representations may be taken as a measure of semantic dissimilarity between their two corresponding documents. If the documents are scientific papers for which the cited references are known, one may extend the vector space model to consider their shared cited references, as they usually reflect topic (content) similarity. In this extended version, each unique reference in the collection is added as an additional column in the vector space model and the corresponding value for each article is set to 1 if that reference appears in its cited references, and to 0 otherwise.

4. TIME-BASED LEAST SQUARE PROJECTION

In the following we introduce a multidimensional projection technique that handles the time attribute explicitly so as to convey the changes in the similarity relationships of a multidimensional dataset, along time. It outputs a temporal sequence of similarity-based maps, given a time-stamped multidimensional dataset.

The following steps explicit the time-based multidimensional projection:

1. Given a set of (possibly pre-processed) time-stamped data instances, split the data into a list of batches $X = \{X_1, X_2, \ldots, X_T\}$, according to some temporal property. This list must be organized in ascending order regarding the temporal property chosen.
2. Project the entire dataset using LSP, creating a data map P_T. Create an index *current_map* with value T.
3. Remove the latest batch from list X. Decrease the value of index *current_map* by one, i.e., $current\_map = current\_map - 1$.
4. The next data map, $P_{current\_map}$, is created using a backward scheme as follows. Identify in the previous data map, $P_{current\_map-1}$, the following data instances:
   (a) Data instances that belonged to the batch removed in Step 3, which must be removed from the current map.
   (b) Data instances which would have their high-dimensional neighborhoods changed as a result of the previous step, i.e. that had in their neighborhood instances identified in Step 4(a). The data neighborhoods computed when applying LSP to generate the previous map are inspected to find the data points with modified neighborhoods.
   (c) Data instances outside the neighborhood of the instances identified in Step 4(a).
5. Use LSP to reproject the data instances identified in step 4(b), taking the instances identified in step 4(c) as control points. This actions will create map $P_{current\_map}$. Add this new map to the list of maps P.
6. Repeat steps 3 through 5, until there are only two batches on the list of batches X.
7. Reverse the order of the list of maps to $\{P_1, \ldots, P_T\}$ for display.

A problem may occur when there is an insufficient number of control points. This happens when the projected positions of most data instances need to be updated. This is solved by choosing new control points among the data instances that need to be updated. With this purpose in mind, each data instance to be updated is assigned a weight ω, which infers how close to it the removed instances were from its neighborhood:

$$\omega(x_i) = \sum_{i=1}^{k} \left\| f(x_i), f(x_j) \right\|, \ \forall x_j \in r(x_i) = \{x_1, \ldots, x_k\} \quad (1)$$

where $r(x_i)$ is a function that returns a list of the removed data instances in the neighborhood of data instance x_i. These documents are sorted in ascending order of this weight ω. Supposing that the current map has l control points and it is necessary to achieve at least *min_cp* control points, then the first ranked $(min\_cp - l)$ data instances are included as control points.

By using information from the previous map to build the current one we seek to maintain a global spatial coherence throughout the sequence of maps. It is expected that data instances that have similar content and are positioned at a certain region in a map P_i stay roughly in the same region in the subsequent map P_{i+1}. This behavior is desirable in order to preserve the user's mental map: despite modifications, layouts should remain consistent throughout the time-based sequence in order to avoid user confusion. The term "mental map" refers to the structural cognitive information a user creates internally by observing the layout of a visual representation [4].

5. APPLICATION TO SCIENTIFIC COLLECTIONS

Figure 1(a) illustrates the output of the technique just described: a time sequence of content-based similarity maps depicting a collection of articles. The collection show corresponds to publications by researcher Alessandro Vespignani from 1995 to 2010, as collected from ISI Web of Science (`http://www.isiknowledge.com`). Each individual map displays the articles published from the initial year up to the year indicated at the bottom. In the maps each circle represents an article, circle color indicates publication year and size indicates the global citation count. The edges represent the bibliographic coupling between two articles, and their color is interpolated from the color of the vertices. The pre-processing steps described in Section 3 have been applied to generate a vector space model of the collection, and similarity is computed with the cosine metric applied to the vector representation extended with information on the references.

A user may observe the sequence of maps generated from a collection, stepping forward or backward in time, or alternatively observe an animated view in which document positions are interpolated between consecutive time steps (see video available at `http://lcadfs2.lcad.icmc.usp.br/~aretha/timeaware`). In the latter case the visualization somewhat resembles the visualizations provided by Gapminder (`http://www.gapminder.org`), which show the temporal evolution of the relationship between two data variables. The time oriented LSP approach, however, attempts to capture the temporal evolution of the relationship amongst several data variables, and interpretation is not as straightforward, as screen positions (x,y) are indicative of relative proximity in the data space, with no direct mapping to two variables.

In Figure 1(b) one observes an enlarged view of the map corresponding to the collection up to year 2010 and some associated functionalities, e.g. resources to modify visual mappings and show topics related to groups of documents. Topics are obtained with a topic extraction technique based on terms covariance [5] to extract topics that summarize the major themes addressed by a group of selected documents .

Topic extraction works in the following manner: the document collection must be clustered, either manually or automatically, into groups of content-related (similar) documents. Then, sets of meaningful terms are extracted from the selection by first identifying the two terms with the highest covariance in the vector representations of the selected documents. An initial topic is defined that consists of these two terms. For each remaining term, the mean of the covariance taking the two terms selected in the first step is obtained, and if this mean is higher than a user-defined threshold α (a value in $[0,1]$), the term is added to the topic label. This method is highly affected by the choice of the two initial terms and there may be other pairs of terms with high covariance. In order to overcome this problem, whenever the ratio of the covariance of any pair of terms and the largest covariance in the matrix is above a certain threshold β (also a value in $[0,1]$), these terms are also considered to construct another topic label. The choice of α interferes on the number of terms added to the topic (the lower the threshold, the more terms are likely added), whereas the choice of β affects the number of different topics identified (again, more topics created for lower thresholds).

Figure 2 shows a comparison between maps generated with the proposed approach for three specific time stamps of the Alessandro Vespignani data, and the corresponding maps obtained with the conventional LSP. As expected, the standard LSP maps have their layout orientation changed significantly from one time instance to the next, visual groupings occupy different areas in different maps and are groups are more spread throughout the map area.

6. RESULTS

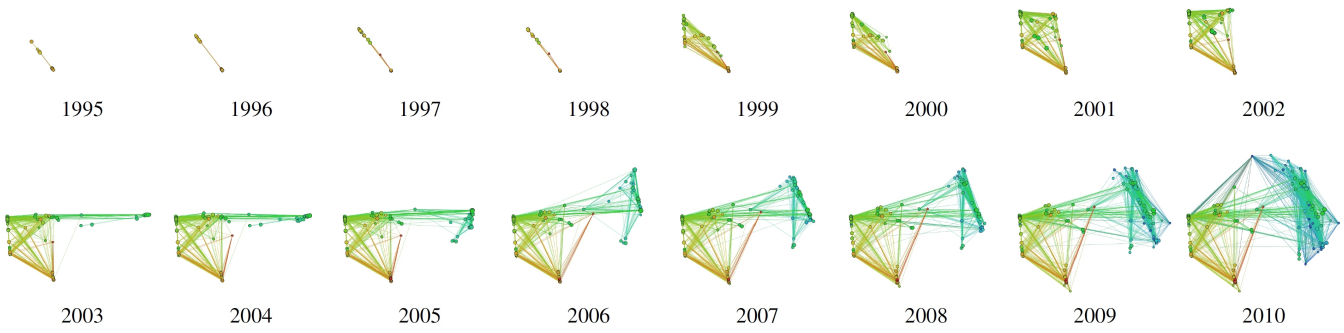
We employed the proposed time-based projection technique to observe the research trajectory of the physicist Alessandro Vespignani, based on his reported scientific production indexed by Thomson Reuters Web of Knowledge (`http://www.webofknowledge.com`). On the time-stamped maps presented in this section we applied the topic extraction technique based on terms covariance to identify the topics addressed by highly related articles as indicated by the visual groupings in the map. Due to space constraints, we do not show the complete sequence of time-stamped similarity-based maps generated which may be found at `http://lcadfs2.lcad.icmc.usp.br/~aretha/timeaware`. At the same site we show the sequence of maps obtained for two other scholars, namely Albert-László Barabási and Osvaldo Novais de Oliveira Jr.

Alessandro Vespignani is an Italian physicist and Professor of Informatics and Computing and adjunct professor of Physics and Statistics at Indiana University, USA, where he is also the director of the Center for Complex Networks and Systems Research (CNetS) and associate director of the Pervasive Technology Institute. He obtained his Ph.D in Physics at the University of Rome "La Sapienza" in 1993. Following postdoctoral research at Yale University and Leiden University from 1993 to 1996, he worked at the International Center for Theoretical Physics in Trieste from 1997 to 2002. He also worked briefly at the University of Paris-Sud from 2002 to 2004, before moving to Indiana University in 2004.

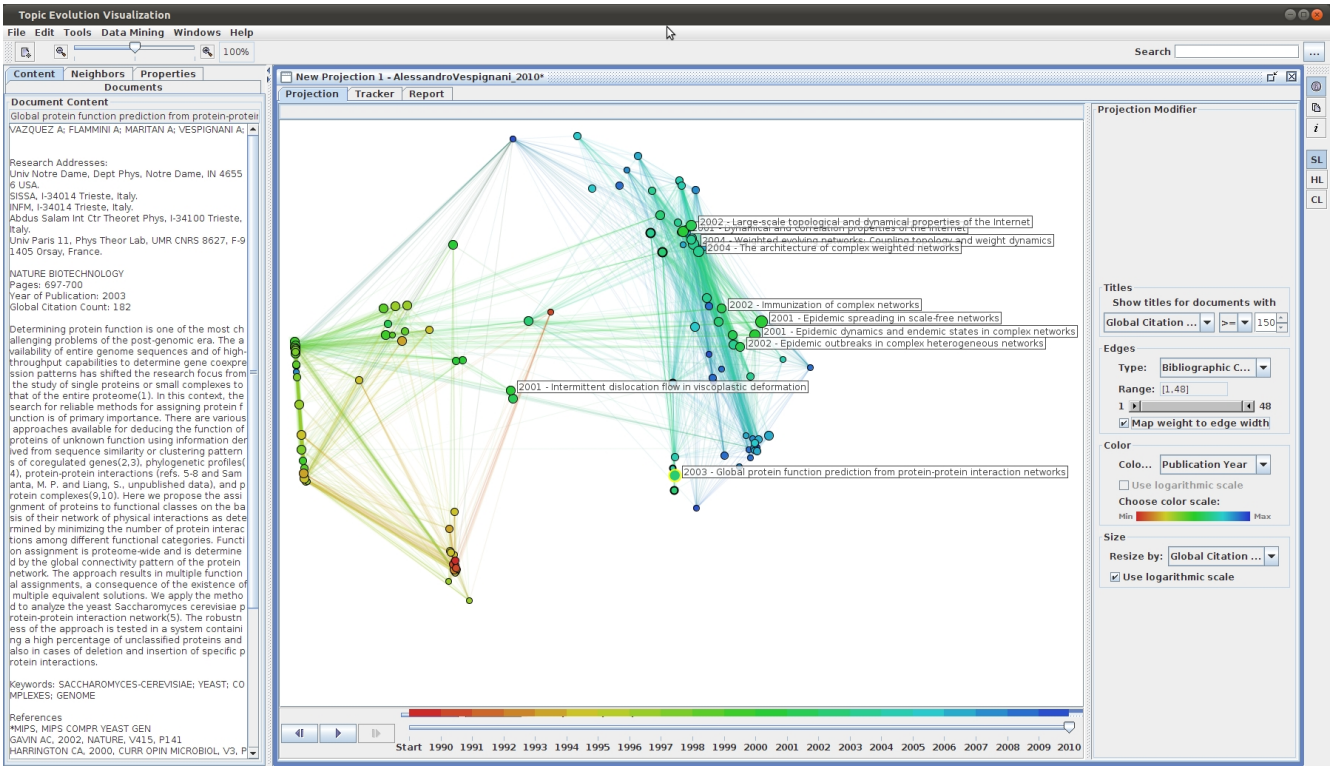
Vespignani has contributions in several areas of Physics, including characterization of non-equilibrium phenomena and phase transitions. He is best known, however, for his work on complex networks. His current research focuses on the interdisciplinary application of statistical and numerical simulation methods to the analysis of epidemics and spreading phenomena, and also on the study of biological, social and technological networks.

The dataset of articles authored and co-authored by Alessandro Vespignani contains a total of 133 documents. Figures 3(b), 3(c), 3(d) and 3(e) detail four time-stamps of the map sequence generated for the Alessandro Vespignani collection for years 2002, 2006, 2008 and 2010, respectively. Circle color maps its corresponding article's publication year, according to the rainbow color scale shown at Figure 3(a). Circle size is proportional to the global citation count of the article, in the logarithmic scale. The edges are based on the bibliographic coupling between two articles: an edge is added between two articles if they have at least one reference in common, and edge color is an intermediate between the colors of the circles it connects.

To automatically identify relevant sub-groups of similar papers at each time-stamped map, we employed the Density Based Spatial Clustering of Applications with Noise (DBSCAN) clustering algorithm [6]. DBSCAN has been designed to discover clusters of arbitrary shape, relying on density information for doing so. As such, it is suitable to identify high-density regions in our similarity-based maps. It requires only two parameters, namely: ε, how close two points must be to be determined as part of the same cluster; and *minPts*, the minimum number of points required to form a cluster. We applied the same parameters values ($\varepsilon = 0.07$ and $minPts = 3$) for all time-stamped maps. Some articles are not included in clusters because they were identified as noise by the clustering algorithm, i.e., from their placement in the map they do not bear strong similarity to any cluster. After identifying the clusters, we extracted topic labels for each cluster with the covariance tech-



(a) Sequence



(b) Screenshot

Figure 1: (Top) Sequence of time-stamped document maps of the collection of articles by scholar Alessandro Vespignani, as reported at ISI; and (Bottom) a larger view of one particular map in the sequence. Each point in the maps represents a scientific article, point color indicates publication year and size indicates the value of the global citation count. Edges represent the bibliographic coupling between articles.

nique described in Section 5. The topics shown were obtained with parameters $\alpha = 0.4$ and $\beta = 0.8$.

The time-stamped map depicted in Figure 3(b) discloses the articles published from 1990 to 2002. At the bottom of each figure we show the topic labels extracted for the automatically extracted clusters, visually indicated by the gray boundaries. In our implementation, a user may observe the topics in the map itself, by placing the mouse over the region corresponding to a cluster. A major area is observed in this map, which includes mainly articles related to topics in Physics, the first research area pursued by Alessandro Vespignani. Given the topic labels extracted for cluster (1) and the fact that this cluster is formed by articles published from 1990 to 1997, we infer that it includes articles reporting scientific results obtained during the scholar's Ph.D. in Physics obtained in 1993, entitled "*Fractal Growth and Self-Organized Criticality*". Clusters identified as (2), (4) and (5) are also related to topics referring to early work in Physics. Cluster (3), despite being close to other clusters related to Physics, includes 7 articles on Complex Networks. As we shall see in the subsequent map, shown in Figure 3(c), this cluster will move to the right region, where articles related to the Complex Networks area start to appear.

In the map generated from the articles published from 1990 to 2003, shown in Figure 3(c), we extracted topic labels for the 5 clusters visually identified in this particular time-stamp of the sequence. The topics related to Physics – clusters (1), (2), (4) and (5) – persist in the same relative position and display similar topical labels. At the same time, a second cluster of articles emerges at the right side of the map, which represents the initial strengthening of the

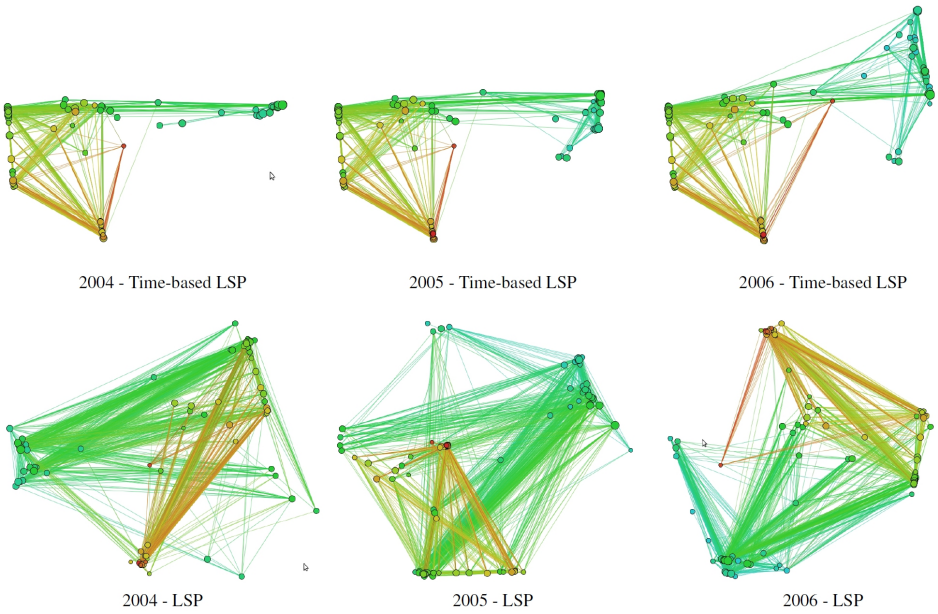


Figure 2: Timewise sequence maps obtained with the proposed approach and corresponding maps from the same data stamps generated with the original LSP.

topics related to what used to be cluster (3) in Figure 3(b). This occurs because Alessandro Vespignani's research slowly changed focus from his previous areas in Physics to Complex Networks.

The time-stamped map depicted in Figure 3(d) represents the articles published from 1990 to 2008. In this map, there are three new clusters (3), (4) and (5) that originated from previous cluster (3) in Figure 3(c). A major current research line by Dr. Vespignani is the application of network models to describe the interactions governing spreading phenomena taking place within populations and technological environments Vespignani, which is represented by cluster (3). Cluster (4) is formed mainly by articles addressing the application of complex network models to study properties of the Internet, e.g., the content of one of the articles in this cluster is about the automatic extraction of semantic information from text and links in web pages in order to assess the similarity among pairs of web pages. Cluster (5) is composed by articles related to employing complex networks to determine the function of proteins. Another event is the merging of clusters (4) and (5) in Figure 3(c) into cluster (6) in Figure 3(d).

The final projection map shown in Figure 3(e) includes all articles in the dataset, from 1990 to 2010. One identifies two major areas that represent the research topics in Physics, shown in red-green at the left region, and more recent Complex Networks research in green-blue at the right region. Observe that the Physics area in Figure 3(e) remains practically the same as the one appearing in Figure 3(b), which is explained by the fact that Alessandro Vespignani almost did not publish any articles in Physics after 2003. This behavior illustrates the capability of the proposed technique of preserving the spatial coherence along the sequence of time-stamped maps. Within the area of Complex Networks, the previous cluster (3) in Figure 3(d) split up into clusters (3) and (7) in Figure 3(e). Cluster (3) represents epidemic studies on scale-free networks that have the peculiar property of being prone to the spreading of infections. Cluster (7) is formed by articles that use metapopulation models – a theoretical framework used to describe population dynamics whenever the spatial structure of populations is known to play a key role in the system's evolution – to study the dynamics of epidemic process.

We generated sequence maps for two other scholars, namely Albert-László Barabási (169 papers published from 1989 to 2010), and Osvaldo Novais de Oliveira Jr (332 papers published from 1984 to 2010). Results are available at `http://lcadfs2.lcad.icmc.usp.br/~aretha/timeaware`. Again, representations proved useful to summarize and visually highlight important turning points in their research trajectories, such as the introduction of new research topics and the strengthening of particular research subjects.

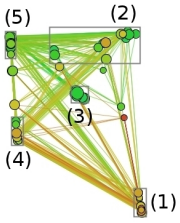
7. CONCLUSIONS

We introduced a temporally-aware Least-Square Projection technique to generate a sequence of similarity-based maps that conveys the evolution of a collection of documents along time, and illustrated its usefulness to analyze the research trajectory of a research scholar. We believe this approach can also support users in analyzing the evolution of a research topic or area from its reported scientific production and, in fact, conveying temporal changes in the similarity relationships of any kind of time-stamped data that can be embedded in a multidimensional space or for which similarity relationships may be derived.

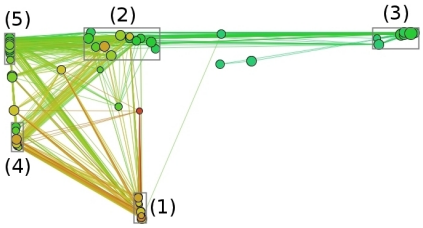
A major issue that remains to be addressed is scalability: we only applied this approach to relatively small collections of scientific papers related to one individual, whereas it would be desirable to handle larger collections, e.g., of conference proceedings or groups of individuals. In this case we anticipate alternative visual metaphors may be required to achieve multiple levels of visual summarization.

As future work, we would like to further investigate its application to analyzing other collections of scientific articles. One may consider whether other alternative models of the collection that combine content similarity with article attributes allow obtaining better maps in terms of similarity preservation. Moreover, alternative topic detection strategies may be investigated in this context and compared regarding their output, and the sequence of temporal topics obtained from the sequence of document maps could be

(a) Color Legend



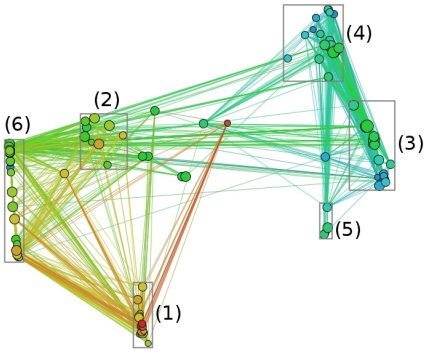
(b) 1990 - 2002



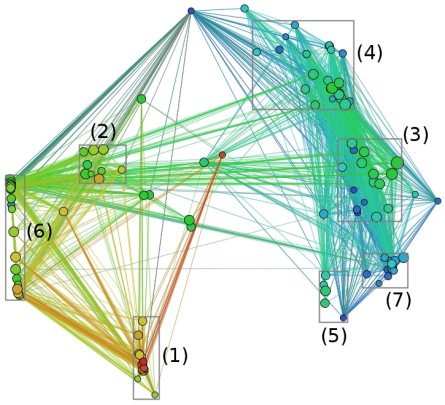
(c) 1990 - 2003

Group	Topic labels
(1)	(fractal, scale, growth, fixed, invariant, critical, irreversible) (dla, parallel)
(2)	(dislocation, deformation, acoustic, ice, creep, single, motion, viscoplastic, interacting) (model, phase, transition, nonequilibrium, cultural) (breakdown, scaling, field, fracture, nucleation, electrical) (organized, behaviour)
(3)	(connectivity, immunization, complex, show, high) (internet, dynamical, time, correlation, growth, structure, topological, growing) (finite, size, scale, epidemic, free, small, threshold, present, bounded, cutoff) (spreading, computer)
(4)	(critical, renormalization, group, scheme, nonequilibrium, driven, dynamical) (point, nu)
(5)	(absorbing, phase, critical, state, transition, density, percolation) (scaling, zhang)

Group	Topic labels
(1)	(fractal, scale, growth, fixed, invariant, critical, irreversible) (dla, parallel)
(2)	(dislocation, deformation, acoustic, ice, creep, single, motion, viscoplastic, interacting) (model, phase, transition, nonequilibrium, cultural) (breakdown, scaling, field, fracture, nucleation, electrical) (organized, behaviour)
(3)	(epidemic, threshold, scale, free, connectivity, spreading, small, finite, size, bounded, cutoff) (internet, dynamical, correlation, time, real, growth, structure, topological, growing) (complex, show)
(4)	(critical, renormalization, group, scheme, nonequilibrium, driven, dynamical) (point, nu)
(5)	(absorbing, phase, critical, state, transition, density, percolation) (scaling, zhang)



(d) 1990 - 2008



(e) 1990 - 2010

Group	Topic labels
(1)	(fractal, scale, growth, fixed, invariant, critical, irreversible) (dla, parallel)
(2)	(model, phase, transition, cultural, nonequilibrium) (breakdown, scaling, fracture, field, nucleation, electrical) (law, power, emission, acoustic, organized, criticality, microfracturing, behaviour, tau) (disordered, order)
(3)	(pandemic, av, global, containment, influenza, rate, high, reproductive, worldwide, therapeutic) (scale, free)
(4)	(similarity, semantic, web, measure) (graph, statistical)
(5)	(protein, function) (interaction, complex)
(6)	renormalization, group, critical, dynamical, nonequilibrium, driven, scheme, dynamically) (absorbing, phase, state, transition, energy, density, conserved, percolation) (model, scaling)

Group	Topic labels
(1)	(fractal, scale, growth, fixed, invariant, critical, irreversible) (dla, parallel)
(2)	(model, phase, transition, cultural, nonequilibrium) (breakdown, scaling, fracture, field, nucleation, electrical) (law, power, emission, acoustic, organized, criticality, microfracturing, behaviour, tau) (disordered, order)
(3)	(scale, free, epidemic, connectivity, degree, small, threshold, size) (backbone, multiscale)
(4)	(similarity, semantic, web, measure) (graph, statistical)
(5)	(protein, function) (person, face)
(6)	(renormalization, group, critical, dynamical, nonequilibrium, driven, scheme, dynamically) (absorbing, phase, state, transition, energy, density, conserved, percolation) (model, scaling)
(7)	(model, based, metapopulation, data, agent) (pandemic, av)

Figure 3: Time-based similarity maps of articles authored or co-authored by Alessandro Vespignani.

analyzed to identify correlations and detect meaningful vocabulary changes. Another desired facility is integrating exhibition of the topics over the maps, while visually emphasizing strong topical changes along time - these could be indicated by regions of great variation over successive time-stamped maps.

8. ACKNOWLEDGMENTS

This work was supported by the Brazilian funding agencies FAPESP (grants 2008/00848-1 and 2008/04622-8), CAPES (grant 1271-10-5) and CNPq (305079/2009-3).

9. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003.

[2] C. Chen. Citespace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57:359–377, February 2006.

[3] A. M. Cuadros, F. V. Paulovich, R. Minghim, and G. P. Telles. Point placement by phylogenetic trees and its application to visual analysis of document collections. In *IEEE Symposium on Visual Analytics Science and Technology*, pages 99–106, Washington, DC, USA, 2007. IEEE Computer Society.

[4] S. Diehl and C. Görg. Graphs, they are changing - dynamic graph drawing for a sequence of graphs. In *Proceedings of Graph Drawing*, pages 23–31, London, UK, 2002. Springer-Verlag.

[5] D. M. Eler, F. V. Paulovich, M. C. F. d. Oliveira, and R. Minghim. Topic-based coordination for visual analysis of evolving document collections. In *International Conference Information Visualisation*, pages 149–155, Washington, DC, USA, 2009. IEEE Computer Society.

[6] M. Ester, H. peter Kriegel, J. S, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *International Conference on Knowledge Discovery and Data Mining*, pages 226–231. AAAI Press, 1996.

[7] 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(1):9–20, 2002.

[8] B. Herr, R. Duhon, K. Borner, E. Hardy, and S. Penumarthy. 113 years of physical review: Using flow maps to show temporal and topical citation patterns. In *International Conference on Information Visualisation*, pages 421–426, Los Alamitos, CA, USA, 2008. IEEE Computer Society.

[9] J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7:373–397, 2003. 10.1023/A:1024940629314.

[10] L. Leydesdorff and T. Schank. Dynamic animations of journal maps: Indicators of structural changes and interdisciplinary developments. *Journal of the American Society for Information Science and Technology*, 59:1810–1818, September 2008.

[11] S. Liu, M. X. Zhou, S. Pan, W. Qian, W. Cai, and X. Lian. Interactive, topic-based visual text summarization and analysis. In *ACM Conference on Information and Knowledge Management*, pages 543–552, New York, NY, USA, 2009. ACM.

[12] A. A. Lopes, R. Pinho, F. V. Paulovich, and R. Minghim. Visual text mining using association rules. *Computers and Graphics*, 31:316–326, June 2007.

[13] H. P. Luhn. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2:159–165, April 1958.

[14] F. V. Paulovich and R. Minghim. Hipp: A novel hierarchical point placement strategy and its application to the exploration of document collections. *IEEE Transactions on Visualization and Computer Graphics*, 14:1229–1236, 2008.

[15] F. V. Paulovich, L. G. Nonato, R. Minghim, and H. Levkowitz. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics*, 14:564–575, 2008.

[16] F. V. Paulovich, M. C. F. Oliveira, and R. Minghim. The projection explorer: A flexible tool for projection-based multidimensional visualization. In *Brazilian Symposium on Computer Graphics and Image Processing*, pages 27–36, Washington, DC, USA, 2007. IEEE Computer Society.

[17] PNNL. IN-SPIRE^{TM} Visual document analysis. http://in-spire.pnl.gov, 2008.

[18] M. F. Porter. An algorithm for suffix stripping. *Program: Electronic Library & Information Systems*, 40(3):211–218, 1980.

[19] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18:613–620, November 1975.

[20] Sci^2 Team. Science of Science (Sci^2) Tool. Indiana University and SciTech Strategies. http://sci2.cns.iu.edu, 2009.

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

[22] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. Tiara: a visual exploratory text analytic system. In *ACM International Conference on Knowledge Discovery and Data Mining*,pages 153–162, New York, NY, USA, 2010. ACM.