Visualization methods for metric studies

FABRICE ROSSI[1]

1 Inria Rocquencourt Projet AxIS
Domaine de Voluceau, Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex France

Abstract

Metric studies are based on complex, voluminous and heterogeneous data. In order to obtain meaningful results, human guided analysis is therefore needed and can be achieved with information visualization methods. In this paper, we survey visualization methods traditionally used in informetrics and present recent achievements in this domain. We also outline some potentially interesting visualization tools from machine learning.

1. Introduction

Information visualization [1] uses the high processing capabilities of the human vision system to enable interactive exploration of abstract data. Humans are extremely good at spotting patterns in images. This skill is supported by the low level visual system which has pre-attentive processing capabilities (see e.g. [2]): some specific features, such as colors or shapes, are detected and recognized without effort and very quickly, generally in less than 200ms, even in a large image. Recent experiments have shown that pre-attentive processing scales up to one million items [3].

Human vision is however limited by several factors. The most obvious is that vision is physically restricted to three dimensional displays (3D). In addition, 3D has many specific problems, such as occlusion, disorientation, scarcity of stereovision hardware (despite its low cost), etc. As shown in [4], effectiveness of graphical features is also a complex issue: for instance color hue is an efficient and accurate way of displaying nominal variables whereas it is not adapted to quantitative value representation. Moreover fast pre-attentive processing is limited to two or three combined features (see [2]).

In practice therefore, original data are mapped to a two dimensional (2D) layout that exploits a few pre-attentive features to represent the characteristics of the studied objects. Interaction methods such as zooming, linking-and-brushing, dynamic distortion, etc. enable the user to manipulate and dynamical modify the image: she can focus on some part of the data, filter irrelevant aspects, compare different points of view (especially in 3D), etc.

In general, layout strategies leverage some particularities of the data. Scientific visualization, for instance, represents real world phenomena (such as fluid dynamics) and can therefore use their natural 3D interpretation. Another example comes from the general abstract data model called the *Table Data Model* [1,5]: each object is described by a fixed (and common) number of attributes. While there is no natural 3D embedding for such type of data, dimension reduction techniques (see e.g. [6] for a recent survey) can be used to map the high dimensional data to a low dimensional representation.

However, general abstract data have seldom a natural table or vector representation. Even if some successful models, such as Salton's Vector Space Model for text [7], have been proposed, they generally induce some distortion or some information loss. Visualization methods must therefore rely on some specific layout techniques. This is especially the case for data from metric studies. Citation data, for instance, correspond to graphs that have no meaningful vector representation.

We give in this paper an overview of the major visualization methods used in metric studies. General surveys of this field can be found in [8,9] and we won't try to have comprehensive coverage as those articles. Examples of state-of-the-art methods can be found in the winning entries of the InfoVis 2004 contest whose goal was to display the history of information visualization based on bibliographic information [10,11]. Our goal is to present the main layout strategies used in informetrics and to outline some possible research opportunities in this area by listing recent machine learning achievements that could be used for visual representation. Section 2 recalls the types of data commonly considered by metric studies. Section 3 is dedicated to dissimilarity methods, based on the transformation of the original data into a dissimilarity matrix. Section 4 presents methods adapted to the vector space model used for, e.g., abstract or full text paper analysis. Section 5 briefly outlines graph based methods that deal with citation data and other networks from metric studies.

2. Metric studies data

Informetrics is based on the analysis of complex, heterogeneous and interlinked data with different level of details. Bibliographic data include for instance articles whose full text is generally described by some meta-data: title, authors, publication mean (journal, conference proceedings, etc.), keywords, abstract, cited papers, etc. Articles are also nodes in a graph whose arcs correspond to citations (author or mixed graphs can also be considered). Analysis can be conducted at the article level, but also at higher levels such as author, journal, research field, etc. This general pattern applies to patents, web sites, etc. Its complexity is increased by the lack of consistency of data sources: older articles might miss full text or keywords, web page meta-data are frequently non-existent or even bogus, etc.

Because of their complex nature, metric studies data are generally transformed to simpler but manageable representations. A standard solution is to consider the graph structure induced at the article level by (co)-citation or at the author level by co-authoring and (co)-citation (higher level can also be analyzed). The obtained graphs can be enriched by node annotations (content of the article, number of articles per author, etc.) and arc annotations (number of co-citations for instance). It should be noted that this graph base representation doesn't generally imply information loss. It's in fact the natural representation for web sites, for instance. Another complementary approach consists in relying more on the textual part of the data, generally by using a vector space model [7] or more directly by focusing on meta-data with high semantic content such as keywords or categories.

Because metric studies are based on very heterogeneous data, aggregate measures are frequently preferred over raw data. Basic statistics such as journal impact factors lead to simple and classical visualization (bar charts, time series, etc.). Aggregation can also be used to define similarities (or dissimilarities) between the studied objects. For instance the Jaccard index applied to cited articles defines a similarity between articles: articles that cite similar papers are considered to be similar. This rationale can be applied at different abstraction levels and to different low level data (authors cited together in articles, articles that use common words, etc.).

To summarize informetrics can be based on three major types of data: annotated graphs, vector data and similarity data (by similarity data, we mean data described by the matrix of their pairwise similarities).

3. Similarity based approaches

3.1. Multidimensional scaling

When data are described only through pairwise similarities, the main goal of visualization is to respect as closely as possible those similarities: close objects according to the similarity matrix should be represented by close points or icons on the 2D (or 3D) layout. The main tool for this ordination problem is the family of Multidimensional Scaling (MDS) methods, introduced by Torgerson [12]. The main idea of MDS methods is to find a low dimensional representation of the data such that Euclidean distances between points in the low dimensional space are good approximations of the original dissimilarities (which are obtained via simple transformations from the similarities).

Torgerson's classical MDS applies to dissimilarities defined by high dimensional data in an Euclidean space, a situation that doesn't correspond to metric studies data (in fact classical MDS is equivalent to Principal Component Analysis). In more general situations, one relies on Kruskal's MDS [13,14] which is based on the iterative optimization of a stress function that measures the discrepancies between the low dimensional Euclidean structure and the original dissimilarity matrix. This form of MDS is one of the most common tool in metric studies and more generally for mapping non vector data (see e.g. [8,9] for some references in metric studies). The type of MDS and, as a consequence, of visualization, depends on the optimized stress function. Some examples of modified stress include Sammon's non linear mapping [15] which tries to respect more small dissimilarities (in the original space) than large ones and Curvilinear Component Analysis (CCA [16]) which tries to make sure that small distances in the low dimensional representation correspond to small dissimilarities in the original space. To our knowledge, those variants of MDS are seldom use in metric studies, despite their advantages (CCA has low computational requirements for instance).

MDS used to have a major limitation, its high computational cost. Standard MDS algorithms scale in $O(n^3)$ or $O(n^4)$ for n objects. A lot of optimization work has been done in order to reduce this cost. A very sophisticated solution, based on sampling, interpolation and a force directed model is proposed in [17,18] and scales in $O(n^{5/4})$. This implementation achieves impressive scalability: 100 000 objects can be embedded in less than 6 minutes on a standard PC (Pentium IV 2.4 Ghz). Other optimized implementations include hierarchical grid methods proposed in [19]. It should be noted that those methods use approximations to solve the stress minimization problem and therefore that the final value of the stress should be checked in order to make sure that the mapping is acceptable.

Another optimization possibility consists in using spectral analysis (as in the classical scaling) either of the original dissimilarity matrix (discarding or not negative eigenvalues, see [20]) or of some pre-processed version such as in laplacian eigenanalysis (used in VxInsight [21,22]). Those solutions strongly benefit from research in eigenanalysis, especially because some similarity matrices obtained in metric studies are sparse. As for iterative MDS, classical MDS can be accelerated by approximating some calculation. The well known FastMap algorithm [23] for instance is in fact an approximate classical MDS [24]. It should be noted however, that methods based on classical MDS don't minimize the same stress function as non classical MDS and can lead to less interesting embeddings, mainly because even if the underlying similarity is metric, a perfect Euclidean embedding might be impossible: knowing that mapping errors are unavoidable, stress functions are designed to limit some type of errors while disregarding others. CCA [16], for instance, is known to be able to unfold complex structures by “tearing” the high dimensional surface.

A limitation MDS that is still valid is the fact that global analysis is difficult, partly because axes of the mapping don't have generally any obvious meaning. This might be considered as a consequence of the non linear nature of non classical MDS as opposed to Principal Component Analysis for instance. Another consequence of non linearity is the difficulty to trust proximities in the 2D representation. The valued of the stress function gives only a rough estimation of the quality of the mapping. Better global measures have been proposed [25,26] and local distortions can be visualized on the layout itself [27], but those techniques have not been applied yet to metric studies.

3.2. Other methods

While the family of MDS methods dominates the field of ordination algorithms, other solutions are used. VxInsight uses for instance a sophisticated type of force directed placement (FDP), VxOrd [21,22], that scales in $O(n)$. The algorithm is related to MDS but the optimized stress is quite different. As pointed out in [9], VxInsight and similar FDP methods have been used in many metric studies.

Older methods include for instance the triangulation technique proposed in [28] which provides perfect preservation of some of the original dissimilarities (the dissimilarities between any object and its two closest neighbors are exactly represented which corresponds to *2(n-3)+3* perfectly represented values).

Kohonen's Self Organizing Map (SOM [29]) have been used in DIVA [30] to provide ordinations of similarity data from metrics studies (DIVA maps documents according to citations and words based similarities). One of the limitations of the proposed solution is that it transforms the similarity matrix into a vector model: two objects are close according to this model if they have the same pattern of similarity to all other objects. In a way, this emphasizes global distances rather than local ones. A similar approach is used in [31] for mapping authors based on co-citation similarities. While this approach can lead to good results, its theoretical properties are not well known. Some extensions of the SOM to dissimilarity data that try to give a proper semantic to the algorithm have been proposed [32-35], but none have been applied to metric studies. There computational costs tend to be quite high even if recent advances have been made [36].

4. Vector model

When the data can be accurately represented in a vector space, most of the standard visualization tools can be applied (more precisely methods that target the table data model [1,5]). In metric studies, the vector space approach has been mainly applied to textual data, via Salton's model [7] combined with some term weighting strategy such as inverse document frequency. For full text analysis, this generates very high dimensional vectors: simple linear dimension reduction techniques, such as Principal Component Analysis (PCA) and Latent Semantic Analysis (LSA [37]) are generally used to limit noise and to enhance semantic content.

The obtained data (raw vector data or preprocessed vector data) are sometimes cast into a similarity matrix (via the Euclidean norm or more generally a Minkowsky norm) and submitted to a non linear MDS algorithm.

4.1. The Self Organizing Map

Apart from this approach, the most used vector space tool in metric studies is Kohonen's SOM [29]. Pioneering work in this area was done by X. Lin et al. in [38] which is generally considered as the first application of the SOM to abstract information visualization (document spaces). The SOM has several advantages over MDS that include its scalability to huge data sets (the impressive Websom application was conducted on almost seven millions of patent abstracts [39]), the simultaneous clustering and projection operations that provide both overview and detailed representation, the numerous visualization methods [40-42] that give a complete view of the SOM's results, etc.

The standard SOM algorithm has been modified and enhanced for text database visualization and more generally for document space analysis. Chen et al. have proposed for instance a type of hierarchical SOM (layered SOM in their words [43]) that enables to recursively refine the mapping by training a SOM on some area of the results of a previous SOM. This model was further enhanced in [44] by taking advantage of the sparsity of the vector model for text analysis (see also [45] for recent achievements of the same team). Other hierarchical SOM methods have been proposed for similar tasks, for instance the Treeview SOM [46] which targets web pages. A more complex extension of SOM which is in a way related to hierarchical SOM but with communications between SOMs, the multi-SOM [47], has also been applied to the mapping of document bases [48].

The SOM appears has a meeting point for many communities. For example, its excellent visualization capabilities have been a motivation for researchers of the machine learning community to leverage concepts and methods from information visualization. The general concept of “focus + context” [49] has been for instance applied to the SOM by Ritter in [50] and further extended to hierarchical growing SOM by Ontrup and Ritter in [51] (see also [52] for dynamic distortion techniques applied to the SOM).

Because it generates a map, the SOM has also been exploited and analyzed by researchers from the geographic information system (GIS) community. Skupin for instance has studied the geographic and cartographic metaphors for displaying information spaces, originally with MDS [53] and latter with the SOM [54,55]. The display of the SOM results by GIS software provide very interesting results with might be easier to comprehend by non expert users with the help of a strong cartographic metaphor. A survey of the link between information visualization and cartography is available in [56].

4.2. Promising recent methods

To our knowledge, metric studies visualizations based on vector space data have been limited to MDS like methods and to derivatives of the SOM and haven't yet take advantage of recent advances from machine learning in the field of manifold learning [6] and latent variable models [57].

Manifold learning aims at finding low dimensional structures in high dimensional data, such as a sphere in a 3D space. The main idea is to build in the high dimensional space a model of the local structure of the data (via its k-neighbor graph for instance) chosen so as to be easily embeddable in a low dimensional space. Isomap [58] for instance builds a geodesic distance between objects based on the k-neighbor graph and maps the obtained dissimilarity matrix to a low dimensional representation via classical MDS. Curvilinear Distance Analysis (CDA [59]) uses the same idea but replaces MDS with CCA [16]. Locally Linear Embedding (LLE [60]) represents the local structure by a hyperplane and finds a low dimensional embedding that has a similar linear structure. Other related methods have been defined (see [6] and [61]) and tend to produce very interesting results, even in very high dimensional spaces (such as image spaces). It remains however to be seen whether the underlying structure of vector spaces built from metric studies data can be modeled as a low dimensional manifold that can be easily embedded in a low dimensional space. The emphasize put on locality by manifold learning methods might be for instance incompatible with the requirements of metric studies (manifold learning methods do not respect long distances at all because they are discarded in the early phase of the algorithms). It should be noted that many of those methods, especially Isomap, CDA, and LLE, can be applied to dissimilarity matrices.

Latent variable models take in a way the opposite point of view: the high dimensional observed data are supposed to be generated from corresponding low dimensional unobserved (or latent) variables [57] through a known probabilistic model of the general form $t=y(x;w)+e$, where t denote the high dimensional data, x the latent variables, e the noise and y the parametric form of the model with parameters w. The oldest latent variable model is the one of factor analysis (see e.g. [6]) in which y is linear. With some assumptions on the distribution of x and e, it corresponds to PCA (see e.g. [62]). It allows one to define probabilistic PCA [62], which in turns gives mixture of probabilistic PCA [63]: in this approach, the data are simultaneously clustered and linearly projected to a low dimensional representation. This allows one to overcome the limitation of one single linear projection, while preserving a simple model. An interactive hierarchical version of this model has been proposed in [64] and gave interesting visual results. The Generative Topographic Mapping (GTM [65]) is a SOM like non linear latent variable model that has visualization capabilities quite similar to those of the SOM, with the advantage of being based on a simple probabilistic model. An interactive hierarchical version of the GTM has been proposed in [66].

The main difficulty in designing latent variable models is to come up with meaningful probabilistic high dimensional models. The standard GTM for instance is not adapted to discrete data and has to be modified for this type of task as in [67], or more recently with the Latent Trait Model (LTM [68]). It has been shown in [69] that combining LTM with the hierarchical GTM of [66] provides very interesting visualization possibility for text corpus represented by a binary vector space model. Many recent works on latent variable models try to define generative models for different types of data. The model of [70] targets for instance text streams, while [71] is designed for symbolic sequences, such as web log data. Latent variable models are therefore applicable to metric studies data such as text collections, but broader applications, e.g., to graph structures, will be possible only if meaningful generative models can be build for those type of data (based for instance on random graph models, see e.g. [72] for a survey).

5. *Graph based methods*

The natural representation of metric studies data is generally a graph with annotations and early visualization of such data were done by manual layout of the graphs (see e.g. [73]). Latter works represent simplified graphs (clustered graphs) via MDS (see e.g. [74]). This type of visualization is based on a transformation of the graph into a dissimilarity matrix and therefore doesn't directly leverage the graph structure of the data. On the contrary, the work of Chen (Generalized Similarity Analysis, GSA [75,76]) is truly based on the links between analyzed objects. Chen's method is based on the extraction from the graph of a Pathfinder network [77]: this network acts as a summary of the original graph by retaining only some of the links between nodes. The obtained sub-graph depends on the parameters of the algorithm, but they can be chosen such that the pathfinder network corresponds to the union of all possible minimum spanning trees of the original graph. In this situation, the geodesic dissimilarity between nodes of the pathfinder network is a distance and is therefore easier to embed in 2D than an arbitrary non metric dissimilarity. Chen uses the Kadama and Kawai's force directed placement method [78] to layout the pathfinder network. GSA has been applied to different metric studies, for instance author co-citation in [79] and for document co-citation in [9]. The main limitation of GSA is the high cost of pathfinder network calculation that scales in $O(n^4)$. It is therefore tempting to use a Minimum Spanning Tree (MST) for which simple algorithms such Kruskal's or Prim's scale in $O(e \log n)$ where e is the number of edges in the graph. While this allows one to obtain interesting results [80], comparative studies have shown that pathfinder networks provide better summaries than MSTs (see e.g. [81]).

More recently, metric studies data visualizations have started to use general purpose graph visualization tools (see [82,83] for surveys on graph visualization), especially software such as Pajek [84], originally designed for social network analysis. Two of the four winning submissions to the InfoVis 2004 contest [10] made intensive use of graph visualization ([85] used Pajek and [86] was based on WilmaScope [87]), while five among the eight second place submissions do the same. One of the most original works from this contest is [88] which uses a hierarchical decomposition of the graph to be visualized based on the small-world structure of many social networks [89].

6. *Conclusion*

Metric studies data are complex and heterogeneous; with the growing availability of interlinked numerical libraries, the volume of those data is becoming huge. Even if automated knowledge discovery tools can be applied to informetrics, interactive visualization tools remain extremely useful to help practitioners to understand those voluminous data.

Recent works in the visualization field try to display the complex networks that give a natural and complete description of some important aspects of metric studies data, e.g. co-citation author networks in bibliometrics. Progresses of the graph visualization fields have allowed the layout of large graphs and provide the infrastructure of those advanced knowledge domain exploration methods.

In the future, dissimilarity based methods, which remain useful even if they loose part of the original structure by summarizing it into a simple numerical matrix, could benefit from recent advances such as efficient Multidimensional Scaling algorithms, modified Self Organizing Maps for dissimilarity data, and Manifold Learning methods like Isomap or LLE.

The standard vector model for text remains a valuable tool for document analysis, especially in conjunction with the Self Organizing Map. Better results might be achieved nevertheless via latent variable models (such as the Generative Topographic Mapping), especially because good generative models have been recently proposed for text data.

References

1. S. K. Card, J. D. Mackinlay, and B. Shneiderman, editors. Readings in Information Visualization: Using Vision to Think, Morgan Kaufmann, San Francisco, 1999.
2. C. G. Healey, K. S. Booth, and J.T. Enns. Visualizing real-time multivariate data using preattentive processing. ACM Transactions on Modeling and Computer Simulation, 5(3):190-221, July 1995.
3. J.-D. Fekete and C. Plaisant. Interactive information visualization of a million items. In Proceedings of IEEE Symposium on Information Visualization 2002 (InfoVis 2002), pages 117-124, Boston (USA), 2002.
4. J. D. Mackinlay. Automating the design of graphical presentations of relational information. ACM Transactions on Graphics, 5(2):110-141, April 1986.
5. P. E. Hoffman. Table Visualizations: A Formal Model and Its Applications. PhD thesis, University of Massachusetts at Lowell, 1999.
6. C. J. C. Burges. Geometric methods for feature extraction and dimensional reduction. In L. Rokach and O. Maimon, editors, Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. Kluwer Academic Publishers, 2005.
7. G. Salton, C. Yang, and A. Wong. A vector space model for automatic indexing. Communications of the ACM, 18(11):613-620, 1975.
8. H. D. White, and K. W. McCain. Visualization of literatures. In M. E. Williams, editor, Annual Review of Information Science and Technology, 32:99-168. Medford, NJ: Information Today, 1997.
9. K. Börner, C. Chen, and K. Boyack. Visualizing Knowledge Domains. In B. Cronin, editor, Annual Review of Information Science & Technology, 37:179-255. Medford, NJ: Information Today, 2003.
10. J.-D. Fekete, G. Grinstein, and C., Plaisant. IEEE InfoVis 2004 Contest, the history of InfoVis, http://www.cs.umd.edu/hcil/iv04contest/, 2004.
11. J.-D. Fekete and C. Plaisant. Les leçons tirées des deux compétitions de visualisation d'information. In Proceedings of IHM2004, pages 7-12, Namur, Belgium, September 2004.
12. W. S. Torgerson. Multidimensional scaling: I. theory and method. Psychometrika, 17:401-419, 1952.
13. J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 29:1--27, 1964.
14. J. B. Kruskal. Nonmetric multidimensional scaling: a numerical method. Psychometrika, 29:115--129, 1964.
15. J. W. Sammon. A nonlinear mapping for data structure analysis. IEEE Transactions on Computer, C-18(5):401-409, May 1969.
16. P. Demartines and J. Hérault. Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets. IEEE Transactions on Neural Networks, 8(1):148--154, 1997.
17. A. Morrison, G., Ross, and M. Chalmers. Multidimensional Scaling through Sampling, Springs and Interpolation, Information Visualization 2(1): 68-77, March 2003.
18. A. Morrison, and M. Chalmers. A Pivot-Based Routine for Improved Parent-Finding in Hybrid MDS, Information Visualization 3(2):109-112, 2004.
19. M. M. Bronstein, A. M. Bronstein, R. Kimmel, and I. Yavneh, Multigrid multidimensional scaling, Numerical Linear Algebra with Applications, 13(2-3):149-171, 2006.
20. J. Laub, and K.-R. Müller. Feature Discovery in Non-Metric Pairwise Data, Journal of Machine Learning Research, 5:801-818, July 2004.
21. G. S. Davidson, B. Hendrickson, D. K. Johnson, C. E. Meyers, and B. N. Wylie. Knowledge Mining With VxInsight: Discovery Through Interaction, Journal of Intelligent Information Systems, 11:259-285, 1998.
22. G. S. Davidson, B. N. Wylie, and K. W. Boyack. Cluster stability and the use of noise in interpretation of clustering. Proc. IEEE Information Visualization, pages 23-30, 2001.
23. C. Faloutsos and K.-I. Lin. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In Proceedings ACM SIGMOD'95, pages 163-174, 1995.

24. J. C. Platt. FastMap, MetricMap, and Landmark MDS are all Nystörm algorithms. In Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics, pages 261-268, 2005.

25. J. Venna and S. Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In Proceedings of ICANN 2001, pages 485-491, Berlin, 2001.

26. S. Kaski, J. Nikkila, M. Oja, J. Venna, P. Toronen, and E. Castren. Trustworthiness and metrics in visualizing similarity of gene expression. BMC Bioinformatics:4, 2003.

27. M. Aupetit. Visualizing distortion in continuous projection techniques. In Proceedings of XIIth European Symposium on Artificial Neural Networks (ESANN 2004), pages 465-470, Bruges (Belgium), April 2004.

28. R. C. T. Lee, J. R. Slagle, and H. Blum. A triangulation method for the sequential mapping of points from N-Space to Two-Space. IEEE Transactions on Computer (26):288-292, 1977.

29. T. Kohonen. Self-Organizing Maps, volume~30 of Springer Series in Information Sciences. Springer, third edition, 2001.

30. S. A. Morris, C. DeYong, Z. Wu, S. Salman, and D. Yemenu. DIVA: a visualization system for exploring document databases for technology forecasting. Computers & Industrial Engineering 43:841-862, 2002.

31. X. Lin, H. D. White, and J. Buzydlowski. Real-time author co-citation mapping for online searching. International Journal of Information Processing & Management, 39(5):689-706, 2003.

32. C. Ambroise and G. Govaert. Analyzing dissimilarity matrices via Kohonen maps. In Proceedings of 5th Conference of the IFCS 1996, volume 2, pages 96-99, Kobe (Japan), 1996.

33. T. Kohonen and P. J. Somervuo. Self-organizing maps of symbol strings. Neurocomputing, 21:19-30, 1998.

34. T. Graepel and K. Obermayer. A stochastic self-organizing map for proximity data. Neural Computation, 11(1):139-155, 1999.

35. A. El Golli, B. Conan-Guez, and F. Rossi. A self organizing map for dissimilarity data. In Proceedings of IFCS 2004, pages 61-68, Chicago, Illinois (USA), 2004.

36. B. Conan-Guez, F. Rossi, and A. El Golli. Fast algorithm and implementation of dissimilarity self-organizing maps. Neural Networks, to be published in 2006.

37. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. A. Harshman. Indexing by latent semantic analysis, Journal of the American Society for Information Science, 41(6):391-407, 1990.

38. X. Lin, D. Soergel, and G. Marchionini. A self-organizing semantic map for information retrieval. Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval, pages 262-269, 1991.

39. T. Kohonen, S. Kaski, K. Lagus, J. Salöjarvi, J. Honkela, V. Paatero, and A. Saarela. Self organization of a massive text document collection. IEEE Transactions on Neural Networks, 11(3):574-585, May 2000.

40. J. Vesanto. SOM-based data visualization methods. Intelligent Data Analysis, 3(2):111--126, 1999.

41. J. Vesanto. Data Exploration Process Based on the Self-Organizing Map. PhD thesis, Helsinki University of Technology, Espoo (Finland), May 2002.

42. J. Himberg. From Insights to Innovations: Data Mining, Visualization, and User Interfaces. PhD thesis, Helsinki University of Technology, Espoo (Finland), November 2004.

43. H. Chen, C. Schuffels, and R. Orwig. Internet Categorization and Search: A Self-Organizing Approach. Journal of Visual Communication and Image Representation, Special Issue on Digital Libraries, 7(1):88-102, 1996.

44. D. Roussinov, and H. Chen. A Scalable Self-Organizing Map Algorithm for Textual Classification: A Neural Network Approach to Automatic Thesaurus Generation. Communication and Cognition in Artificial Intelligence Journal (CC-AI), 15(1-2):81-111, 1998.

45. T. Ong, H. Chen, W. Sung, and B. Zhu. Newsmap: a knowledge map for online news. Decision Support Systems, 39:583-597, 2005.

46. R. T. Freeman, and H. Yin. Tree view self-organization of web content. Neurocomputing, 63:415-446, 2005.

47. J.-C. Lamirel. Application d'une approche symbolico-connexionniste pour la conception d'un système documentaire hautement interactif, Thèse de l'Université de Nancy 1 Henri Poincaré, 1995.

48. X. Polanco, C. Francois, and J. C. Lamirel. Using artificial neural networks for mapping of science and technology: A multi-self-organizing-maps approach. Scientometrics, 51(1): 267-292, 2001.

49. J. Lamping, R. Rao, and P. Pirolli. A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In Proceedings of the ACM Conference on Human Factors in Computing Systems, CHI, pages 401-408, 1995.

50. H. Ritter. Self-organizing maps in non-euclidean spaces. In E. Oja and S. Kaski, editors, Kohonen Maps, pages 97-108. Elsevier, Amsterdam, 1999.

51. J. Ontrup and H. Ritter. A hierarchically growing hyperbolic self-organizing map for rapid structuring of large data sets. In Proceedings of the 5th Workshop on Self-Organizing Maps, Paris (France), 2005.

52. C. Yang, H. Chen and K. Hong. Visualization of large category map for Internet browsing. Decision Support Systems, 35(1):89-102, April 2003.

53. A. Skupin, and B. P. Buttenfield. Spatial Metaphors for Visualizing Very Large Data Archives. Proceedings GIS/LIS'96. Bethesda: American Society for Photogrammetry and Remote Sensing, pages 607-617, 1996.

54. A. Skupin. From Metaphor to Method: Cartographic Perspectives on Information Visualization. In: Roth, S.F., and Keim, D.A. (Eds.) Proceedings IEEE Symposium on Information Visualization (InfoVis 2000), Salt Lake City, Utah, pages 91-97, 2000.

55. A. Skupin. A Cartographic Approach to Visualizing Conference Abstracts. IEEE Computer Graphics and Applications, 22(1):50-58, 2002.

56. A. Skupin, and S. I. Fabrikant. Spatialization Methods: A Cartographic Research Agenda for Non-Geographic Information Visualization. Cartography and Geographic Information Science, 30(2):99-119, 2003.

57. C. M. Bishop. Latent variable models. In M. I. Jordan editor, Learning in Graphical Models, pp. 371–403. MIT Press, 1999.

58. J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, 290(22):2319-2323, December 2000.

59. J. A. Lee, A. Lendasse, and M. Verleysen. Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis. Neurocomputing, 57:49-76, March 2004.

60. S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. Science, 290(22):2323-2326, December 2000.

61. L. K. Saul, K. Q. Weinberger, F. Sha, J. Ham, and D. D. Lee. Spectral methods for dimensionality reduction. In B. Schlkopf, O. Chapelle, and A. Zien, editors, Semisupervised Learning. MIT Press, Cambridge, MA, 2006.

62. M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. Journal of the Royal Statistical Society, Series B, 61(3):611-622, 1999.

63. M. E. Tipping and C. M. Bishop. Mixtures of probabilistic principal component analyzers. Neural Computation, 11(2):443-482, 1999.

64. C. M. Bishop and M. E. Tipping. A hierarchical latent variable model for data visualization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(3):281-293, 1998.

65. C. M. Bishop, M. Svensén, and C. K. I. Williams. GTM: The generative topographic mapping. Neural Computation, 10(1):215-34, 1998.

66. P. Tino and I. Nabney. Hierarchical GTM: Constructing localized non-linear projection manifolds in a principled way. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5):639-656, 2002.

67. C. M. Bishop, M. Svensén, and C. K. I. Williams. Developments of the generative topographic mapping. Neurocomputing, 21:203-224, 1998.

68. A. Kabán and M. Girolami. Combined Latent Class and Trait Model for the Analysis and Visualisation of Discrete Data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(8):859-872, 2001.

69. I. Nabney, Y. Sun, P. Tino, and A Kabán. Semisupervised Learning of Hierarchical Latent Trait Models for Data Visualization. IEEE Transactions on Knowledge and Data Engineering, 17(3), 2005.

70. A. Kabán and M. Girolami. A Dynamic Probabilistic Model to Visualize Topic Evolution in Text Streams. Journal of Intelligent Information Systems, special issue on Automated Text Categorization, 18(2/3):107-125, 2002.

71. P. Tino, A. Kabán, and Y. Sun. A Generative Probabilistic Approach to Visualizing Sets of Symbolic Sequences. Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - (KDD04), August 22-25, Seattle, Washington, USA, pages 701-706, 2004.

72. M. E. J. Newman. The structure and function of complex networks. SIAM Review, 45:167-256, 2003.

73. E. Garfield, I. H. Sher, and R. J. Torpie. The Use of Citation Data in Writing the History of Science. Published by The Institute for Scientific Information, December 1964.

74. H. Small, and E. Garfield. The geography of science: disciplinary and national mappings. Journal of Information Science, 11(4) :147-159, 1985.

75. C. Chen. Structuring and visualizing the WWW with Generalized Similarity Analysis. In 8th ACM Conference on Hypertext (Hypertext '97), Southampton, UK, ACM Press, pages 177-186, 1997.

76. C. Chen. Generalized Similarity Analysis and Pathfinder Network Scaling. Interacting with Computers, 10 (2):107-128, 1998.

77. R. W. Schvaneveldt. Pathfinder Associative Networks: Studies in Knowledge Organization. Norwood, NJ. Ablex Publishing, 1990.

78. T. Kamada, and S. Kawai. An algorithm for drawing general undirected graphs. Information Processing Letters, 31(1):7-15, 1989.

79. C. Chen. Visualizing semantic spaces and author co-citation networks in digital libraries. Information Processing and Management, 35(2):401-420, 1999.

80. S. Noel, C.-H. H. Chu, and V. Raghavan. Co-Citation Count versus Correlation for Influence Network Visualization, Information Visualization, 2(3), 2003.

81. C. Chen, and S. Morris. Visualizing evolving networks: Minimum spanning trees versus Pathfinder networks. In IEEE Symposium on Information Visualization, Seattle, Washington, IEEE Computer Society Press, pages 67-74, 2003.

82. G. Di Battista, P. Eades, R. Tamassia, and I. G. Tollis. Graph Drawing: Algorithms for the Visualization of Graphs. Prentice Hall, 1999.

83. I. Herman, G. Melançon, and M. Scott Marshall. Graph visualization and navigation in information visualization. IEEE Transactions on Visualization and Computer Graphics, 6(1):24--43, 2000.

84. V. Batagelj, and A. Mrvar. Pajek: Program Package for large network analysis (http://vlado.fmf.uni-lj.si/pub/networks/pajek/). XVII Sunbelt Social Networks Conference San Diego, February 13-16, 1997.

85. W. Ke, K. Börner, and L. Viswanath. Major Information Visualization Authors, Papers and Topics in the ACM Library. IEEE Symposium on Information Visualization (INFOVIS'04), 2004.

86. A. Ahmed, T. Dwyer, C. Murray, L. Song, Y. X. Wu. WilmaScope Graph Visualisation. IEEE Symposium on Information Visualization (INFOVIS'04), 2004.

87. T. Dwyer and P. Eckersley. WilmaScope - a 3D Graph Visualization System. In Graph Drawing Software, M. Junger and P. Mutzel, editors, series "Mathematics and Visualization", Springer Verlag, 2003.

88. M. Delest, T. Munzner, D. Auber, J.-P. Domenger. Exploring InfoVis Publication History with Tulip. IEEE Symposium on Information Visualization (INFOVIS'04), 2004.

89. D. Auber, Y. Chiricota, F. Jourdan, and G. Melançon. Multiscale Visualization of Small World Networks, IEEE Symposium on Information Visualization (INFOVIS'03), pages 75-81, 2003.