2015-08

VizLattes: a tool for relevance analysis from scientific co-authorship networks

Conference on Graphics, Patterns and Images, XXVIII; Workshop on Visual Analytics, Information Visualization and Scientific Visualization, VI, 2015, Salvador.
http://www.producao.usp.br/handle/BDPI/49442

VizLattes: a tool for relevance analysis from scientific co-authorship networks

Markus Diego S. S. Dias, Moussa R. Mansour, Luzia M. Romanetto, Maria Cristina F. Oliveira, Luis Gustavo Nonato
Departamento de Matemática Aplicada e Estatística
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo

Fig. 1. Vizlattes layout: The relevant words of each topic on the left. The most relevant authors for each topic below. Each bar is a community and each color is a different topic. On the bounding box the relevant authors for the topic on that comunnity.

***Abstract*—Social network data typically carry attribute information associated with the individuals and to their relationships. Such interconnected information can be useful to identify groups of individuals sharing common attribute properties and also to investigate the behaviour of particular individuals in a global network scenario. Different approaches have been introduced to extract and identify information of interest in social networks, and community identification is one of them. Some methods focus on identifying groups or communities of individuals based on their relationships, while others try to identify groups of individuals based on the common information they share. Integrating both approaches is not straightforward, as different mathematical and computational must be implemented and integrated into a unified framework. In this paper we approach this problem and propose a new method to identify underlying communities in a network, while highlighting the information shared by their components. Our solution relies on a single unified mathematical method. As a proof-of-concept, we have applied the proposed method to scientific co-authorship networks extracted from the well-known Lattes Platform made available by CNPq, the Brazilian national science funding agency. We use textual information on the co-authors and their papers as focus attributes. We show that the method supports both community detection and also the identification of thematic paths, underlying topics and relevant authors characterizing distinct academic communities. The results presented show that this approach can be quite useful for exploration and understanding of academic collaboration networks.**

***Keywords*-Thematic paths, Scientific co-authorship networks, Non-Negative Matrix Factorization.**

I. Introduction

Graphs are an important tool for modeling, analyzing and visualizing social networks, co-authorship networks, biological and patent networks, among others. However, as networks described by massive graphs become ever more common so do the additional mathematical and computational challenges that must be faced in handling them. For instance, when such graphs are rendered as standard node-link diagrams the large number of edges and nodes introduce severe visual cluttering, hampering visualization-based data exploration tasks. The high computational cost involved in the analysis of large networks poses an additional challenge.

Many alternatives have been proposed to overcome the drawbacks involved in large network analysis, such as node and edges prunning [1] or clustering [2]. Some techniques such as matrix decomposition [3], spectral clustering [4], and edge compression [5] have been shown to perform well on large

networks, especially for community detection. Techniques based on *edge clustering* and *edge-bundling* [6] have also been introduced, mainly for visualization purposes. Although useful, most techniques only take into account geometrical and topological information on the networks, ignoring additional available information. In fact, social networks usually carry information that could be used to improve community analysis and assist their interpretation. Such complementary information has been explored by techniques such as Multivis [7], which relies on email information to analyze networks through matrix decomposition. VATT [8] and Utopian [9] also employ matrix decomposition coupled with network information to analyze and find relevant topics in networks of textual documents. Multivis, VATT, and Utopian accomplish network analysis using basically information stored on the network nodes, ignoring graph topological information such as paths and subnetworks.

In this work we investigate alternatives to analyze large networks that take into account information associated with paths of the network graph, improving community and topic detection while still enabling the identification of prominent nodes. Our approach relies on a single mathematical formulation based on non-negative matrix factorization. We apply the method to academic co-authorship networks built from information extracted from the Lattes database, which records the full academic CVs of researchers based in Brazil. When analyzing these co-authorship networks we aim at answering questions such as:

- Which are the most important communities in a given network and the most prominent researchers in those communities?
- Which are the most expressive research topics?
- Is it possible to identify researchers from distinct communities working on similar topics?

Aiming to answer these questions we introduce VisLattes, a new method to analyze and visualize co-autorship networks. VisLattes relies on paths defined on the network and on the non-negative matrix factorization formulation [10] to identify similar topics while preserving the capability of finding related communities. Our method relies on an inovative visual metaphor that clearly reveals the communities, their underlying topics and most prominent authors within each community.

In summary, the main contributions of our work are:

- A path-based mechanism to analyze co-authorship networks.
- A unified method based on Non-negative Matrix Factorization (NMF) to detect communities, topics, and prominent authors.
- A visual metaphor to display the co-authorship communities and their topics.

II. Technical background

In this section we present the proposed NMF-based method for visualization and analysis of social networks.

A. Scientific co-authorship networks

Let $G = (V, E, C,)$ be a network, with V the set of nodes, E the set of edges and C a set of textual documents, each document carrying information linked to a node in V.

Given a social network G, we preprocess the textual documents to create a term frequency matrix of the collection, thus obtaining a matrix A of size n (number of nodes) by m (number of words). Each row represents a document (node) and each column represents a word. Element a_{ij} in matrix A informs the frequency of word j in document (node) i. Each document is described by an m-dimensional vector a_i.

For instance, in our Lattes co-authorship network models, each node represents a researcher and is associated with a text formed by the titles of all her published papers. An edge between two nodes indicates that the respective researchers are co-authors.

We employed the scriptLattes system [11] to derive the target co-authorship networks with all the required information associated with authors and their publications.

B. Relevant paths

Paths are associated with textual information (obtained from their nodes). Each path in the network can represent the equivalence in research areas among researchers. In this context, we can define the concept of relevant path as the shortest path that connects researchers that work in the similar research area.

To determine the most relevant paths, the cosine similarity measure was initially applied for text comparison, and all pairs shortest paths were determined using the Floyd-Warshall Algorithm [12]. We filter these shortest paths based on their size. The relevant paths will be the shortest paths with size smaller than a given constant k.

Given the relevant paths, we create a new term frequency matrix P of size $m \times p$, where p is the number of remaining relevant paths after the filtering based on size. Each term p_{ij} registers the frequency of occurrence of word i in the documents associated with path j.

In the co-authorship network application, our assumption is that the NMF decomposition of the paths will reveal groups of similar or related research topics. Clustering the paths allows a node to be assigned to multiple clusters after grouping.

A transfer function is applied to balance the weights of the word frequency matrix P. The transfer function is given by a function $f : [0, 1] \times [0, 1] \rightarrow [0, 1]$ created using a Bezier surface formulation (Fig. 2).

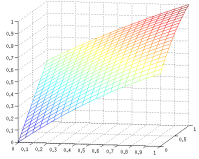

Fig. 2. *Word frequency weighting transfer function*

Given a path c and a word t, the inputs x and y to the transfer function $f(x, y)$ are defined as follows:

- $x = N_p/N_t$ where N_p is the frequency of the word t in the path c and N_t is the highest frequency of t in a path.
- $y = N_{nt}/N_{nc}$, where N_{nt} is the number of document nodes in the path c that include the word t and N_{nc} is the total number of nodes in c.

A new matrix X with size $m \times p$ is obtained, where each term x_{ij} gives the transfer function output for a word i and a path j.

C. Topic extraction with NMF

In our implementation the NMF [10] method was applied for topic extraction, and is now briefly explained.

Let $X \in \mathbb{R}^{n,m}$ be a nonnegative matrix, the NMF method approximates $X \approx WH$, where $W \in \mathbb{R}^{+}_{n\times k}$, $H \in \mathbb{R}^{+}_{k\times m}$ and $k \ll n$ is a given value that defines the maximum rank of the factorization. Such a factoring is obtained from initial matrices W_0 and H_0, applying optimization techniques to an objective function subject to the constraint of nonnegative terms.

Each matrix resulting from the factorization has its own meaning. In the case of extraction threads, X is a matrix determined by the bag-of-words model constructed from the data, where each row corresponds to the description of a particular path and each column entry informs the weight contribution of a particular word in that path. Thus, W gives the topics extracted from data and H brings the association of these topics with path representations.

Using such matrix we came up with a visual metaphor capable of conveying an overview of the relevance of topics and communities identified in the co-authorship network.

D. Metrics

NMF is a tool for automated discovery of topics in large document corpora. Basically, it identifies and outputs a set of topics in the data, as well as a representation of data instances in terms of their topic composition. This is the most common view, but it is also possible to interpret each row in matrix H as a basis for the distribution of paths that contribute to a particular topic. From this information we can extract the relative importance of paths to each topic.

Based on this concept, we employed k-means clustering to group the paths in networks associated with the extracted topics, resulting in a set of networks $\mathcal{C}$.

Although each network community is associated with a major topic, NMF supports a fuzzy representation. In order to better capture and illustrate the multiple topics on a community we compute the relevance of each topic l in a network community c as follows:

$$r_{lc} = \sum_{j \in \mathcal{P}_c} h_{lj}, l = 1, \ldots, k \tag{1}$$

where $\mathcal{P}_c$ are the indices of the paths within a community $c \in \mathcal{C}$.

To assess the relative importance of authors in a particular network community, we consider the paths in this community that include this author. The relevance of each author i in a network community c relative to a topic l is given by:

$$r_{lc_i} = \sum_{j \in \mathcal{P}_{c_i}} h_{lj}, l = 1, \ldots, k \tag{2}$$

where $\mathcal{P}_{c_i}$ are the indices of the paths that include author i within a network community $c \in \mathcal{C}$.

Given these measures, the VizLattes application can show a summary overview of a given scientific co-authorship network, in terms of the coauthorship communities formed, their descriptive topics and the relative expressiveness of distinct topic communities and authors in the overall collaboration network.

E. VizLattes

The NMF was capable of identifying some topic communities in the original co-authorship network. VizLattes represents each community as a vertical bar (fig. 1), with the bar size proportinally mapping the size of its represented community. The size of a community is measured by the number of researchers.

Each bar is split into smaller color regions representing the topic distribution within the community. Each topic is identified by a different color and its relative importance is reflected in the total area of its assigned region. The relative importances of the topics within the community are computed employing the previously described metrics.

Interaction with the visualization is achieved by hovering over on each bar, which results in three kinds of information shown:

- The most representative authors for the topic given by the colored area under the mouse in this particular community are shown in a bouding box.
- The words most representative of this topic are shown on the left area of the visualization.
- The most representative authors for this topic, considering the network as a whole, are shown in the bottom area of the visualization.

III. Results

In order to validate the proposed method we conducted three experiments on data, described in the following:

A. First experiment

For an initial validation of the method and its implementation we created a synthetic dataset consisting of six topics in which each topic is characterized by ten words. We created textual documents by concatenating the topic words characteristic of each topic, and then configured a particular association of these six topics with five nodes.

Links between nodes were established so that nodes sharing topics are densely linked, whereas nodes associated with distinct topics are sparsely linked. We expect that applying our method to this artificial network will properly reveal the six topics with their respective words and prominent nodes.

As the number of topics is known, the NMF was set to cluster the data into six topics. As expected, VizLattes was

TABLE I
THIS TABLE SUMMARIZES MAJOR TOPICS AND RESPECTIVE PROMINENT AUTHORS EXTRACTED WITH THE PROPOSED METHOD FROM THE CO-AUTHORSHIP NETWORK OF THE ICMC FACULTY.

Topics								
Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
trees	bayesian	assisted	mutation	schemes	evolving	web	maps	safety-critical
classification	estimators	crossings	mutants	3d	retrieval	content	notes	aerial
multiclasses	regression	multidimensional	softwares	flows	indexing	materials	topology	unmanned
multi-objective	surviving	biosensing	coverage	free	caracterizao	accessible	genericity	vehicles
radial	measuring	spectroscopy	programs	smoothing	querying	interactive	manifolds	embeddings
algoritmos	poisson	highly	contexto	surfaces	university	videos	planes	lines
Most relevant authors on each topic								
A.C.P.L.F. Carvalho	J.A. Achcar	M.C.F. Oliveira	J.C. Maldonado	A. Castelo Filho	A.J.M. Traina	R.P.M. Fortes	M.A.S. Ruas	K.R.L.J.C. Branco
R.F. Mello	F. Louzada	S.M. Aluísio	M.C.F. Oliveira	J.A. Cuminato	C. Traina Jr.	R. Goularte	C. Biasi	P.C. Masiero
A.C.B. Delbem	F.A. Rodrigues	L.G. Nonato	E.F. Barbosa	L.G. Nonato	M.C.F. Oliveira	R.T.V. Braga	E.M. Zuffi	M.C.F. Oliveira
F.M.B. Toledo	M.C. Andrade Filho	P.C. Masiero	A.S. Simo	A.N. Carvalho	J.E.S. Batista Neto	R.F. Mello	R.N.A. Santos	R.T.V. Braga

capable of conveying the most important words in each topic and the most representative nodes associated with each topic.

B. Second experiment

The second experiment was again conducted on an artificial dataset with the same six topics of the first experiment in which each topic is characterized by ten words.

The difference is the way that we constructed the textual documents of each node. Given a node, we want that this node belongs to one topic. The associated textual document of this node was created containing more words of that topic and fewer words of other topics.

Again, good results were obtained and VizLattes was capable of capturing the relevant communities and words for each topic.

C. Third experiment

The third experiment was performed on a real dataset, using the scriptLattes system to generate a target co-authorship network of the faculty members from ICMC-USP. Results are shown in Table I and figure 1.

IV. CONCLUSION

In this paper a new method has been introduced to detect semantic networks on scientific co-authorship networks. Employing this method we were able to extract latent topics and analyze the relevance of authors based on their collaboration patterns. This information has been mapped to visualization schemes that give data analysts a global contextual picture of the academic collaborations and highlight relevant authors and communities.

As future work we plan to integrate into our tool a supervised variation of the method so as to enable enhanced interactivity and user drilling-down to topics of interest. This would allow additional insight into the data and possibly support gradual topic identification and refinement. Another possible path for further work is to explore other methods based on NMF factorization to reduce sensitivity to data characteristics such as sparsity.

ACKNOWLEDGMENT

The authors are grateful to Jesus P. Mena-Chalco for his helpful assistance with the scripLattes tool. We also acknowledge the financial support of funding agencies CAPES, CNPq and FAPESP.

REFERENCES

[1] F. Zhou, S. Mahler, and H. Toivonen, "Network simplification with minimal loss of connectivity." in *ICDM*, G. I. Webb, B. L. 0001, C. Zhang, D. Gunopulos, and X. Wu, Eds. IEEE Computer Society, 2010, pp. 659–668.

[2] K. Dinkla, M. A. Westenberg, and J. J. van Wijk, "Compressed adjacency matrices: Untangling gene regulatory networks." *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2457–2466, 2012.

[3] R.-S. Wang, S.-H. Zhang, Y. Wang, X.-S. Zhang, and L. Chen, "Clustering complex networks and biological networks by nonnegative matrix factorization with various similarity measures." *Neurocomputing*, vol. 72, no. 1-3, pp. 134–141, 2008.

[4] S. White and P. Smyth, "A spectral clustering approach to finding communities in graphs," *SIAM International Conference on Data Mining*, 2005.

[5] T. Dwyer, N. H. Riche, K. Marriott, and C. Mears, "Edge compression techniques for visualization of dense directed graphs." *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 2596–2605, 2013.

[6] D. Holten and J. J. van Wijk, "Force-directed edge bundling for graph visualization," in *Proceedings of the 11th Eurographics / IEEE - VGTC Conference on Visualization*, ser. EuroVis'09. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2009, pp. 983–998.

[7] J. Sun, S. Papadimitriou, C.-Y. Lin, N. Cao, S. Liu, and W. Qian, "Multivis: Content-based social network exploration through multi-way visual analysis." in *SDM*. SIAM, 2009, pp. 1064–1075.

[8] D. Chu, D. A. Sheets, Y. Zhao, Y. Wu, J. Yang, M. Zheng, and G. Chen, "Visualizing hidden themes of taxi movement with semantic transformation," in *Pacific Visualization Symposium (PacificVis), 2014 IEEE*. IEEE, 2014, pp. 137–144.

[9] J. Choo, C. Lee, C. K. Reddy, and H. Park, "Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, no. 12, pp. 1992–2001, 2013.

[10] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[11] J. P. Mena-Chalco, C. Junior, and R. Marcondes, "Scriptlattes: an open-source knowledge extraction system from the lattes platform," *Journal of the Brazilian Computer Society*, vol. 15, no. 4, pp. 31–39, 2009.

[12] R. W. Floyd, "Algorithm 97: Shortest path," *Commun. ACM*, vol. 5, no. 6, pp. 345–, Jun. 1962.