Reducing Social Network Dimensions Using Matrix Factorization Methods

Václav Snášel, Zdeněk Horák, Jana Kočíbová
VSB Technical University Ostrava
Czech Republic
E-mail: vaclav.snasel@vsb.cz
zdenek.horak.st4@vsb.cz
jana.kocibova.st1@vsb.cz

Ajith Abraham
Center of Excellence for Quantifiable Quality of Service
Norwegian University of Science and Technology
Norway
E-mail: ajith.abraham@ieee.org

***Abstract*—Since the availability of social networks data and the range of these data have significantly grown in recent years, new aspects have to be considered. In this paper we address computational complexity of social networks analysis and clarity of their visualization. Our approach uses combination of Formal Concept Analysis and well-known matrix factorization methods. The goal is to reduce the dimension of social network data and to measure the amount of information which is lost during the reduction.**

***Keywords*-concept lattice, two-mode social network, matrix factorization, correlation dimension**

I. Introduction

As a **social network** we denote a set of subjects which are linked together by some kind of relationship. Social networking – in the sense of providing services to persons to stay in touch, communicate and express their relations – received great attention in the recent years. The fact that the name of one such service has been contender in the choice for the 2007 Word of the year[1] is an evidence of it.

Freeman in [6] underlines the needs for Social Networks Visualization and provides overview of the development of their visualization. The development from hand drawn images to complex computer-rendered scenes is evident. Also the shift from classical sociograms to new approaches and methods of visualization is evident. What remains is the need for clarity of such visualization.

But the social network data are not limited to friendship between people. Many different relations, like ties between animals [14], Web pages and links between them [1], or more generally whole Websites [10], can be also considered. Because of this, the availability and range of social network data increased dramatically in recent years. As a specific kind of network data can be considered so-called **two-mode network data**. This data consists of two sets – set of subjects and set of events which are, or are not, connected. Paper [7] introduces the usage of Formal Concept Analysis (FCA), a well-known general data analysis method, in this area of social networks and reviews the motivation for finding relations hidden in data that are not covered by simple graph visualization. The paper shows that the **Galois lattice** is capable of capturing all three scopes of two-mode network data – relation between subjects, relation between events and also the relation between subjects and events.

A. Complexity aspects

As can be seen both from the mentioned paper and experiments presented below – with the increasing range of input data, the Galois lattice becomes soon very complicated and the information value decreases. Also the computational complexity grows quickly.

Comparison of computational complexity of algorithms for generating Galois lattice can be found in [11]. As stated in the paper, the total complexity of lattice generation depends on the size of input data as well as on the size of output lattice. This complexity can be exponential. Important aspect of these algorithms is their time delay complexity (time complexity between generating two concepts). Recently published paper [4] describes linear time delay algorithm. In many applications it is possible to provide additional information about key properties interesting to the user which can be used to filter unsuitable concepts during the lattice construction [2]. In some applications it is also possible to select one particular concept and navigate through its neighbourhood. These approaches allow us to manage larger scale of data, but cannot provide the whole picture of the lattice.

Many social network data can be seen as object-attribute data or simply matrix (binary and fuzzy). Therefore they can be processed using matrix factorization methods which have been proven to be useful in many data mining applications dealing with large scale problems. In our paper we present results of using Formal Concept Analysis and dimension reduction methodson social network data. Our aim is to allow processing of larger amount of data. Our approach is compatible with the approaches mentioned in the previous paragraph.

Clearly, some bit of information has to be forgotten, but we want to know, how close or far from the original result we are. The paper [15] introduces a method for measuring so-

[1]The word Facebook was selected into the word list by both the American Dialect Society and MerriamWebster dictionary.

978-0-7695-3689-7/09 $25.00 © 2009 IEEE
DOI 10.1109/ASONAM.2009.48

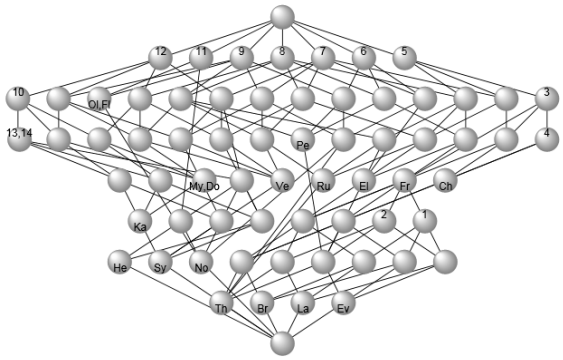

Figure 1. Galois lattice before reduction

called normalized correlation dimension (ncd) which can be seen as the number of independent variables in the dataset.

Singular value decomposition has already been used in the field of social network data ([5]) to determine the position of nodes in the network graph. Next chapter of this paper reviews some basic notions of aforementioned theories. In the third chapter we describe our experiments in detail.

II. Preliminaries

A. Formal concept analysis

Formal concept analysis (shortly FCA, introduced by **Rudolf Wille** in 1980) is well known method for object-attribute data analysis. The input data for FCA we call **formal context** C which can be described as $C = (G, M, I)$ – a triplet consisting of a set of objects G and set of attributes M, with I as relation of G and M. The elements of G are defined as objects and the elements of M as attributes of the context.

For a set $A \subseteq G$ of objects we define $A^{'}$as the set of attributes common to the objects in A. Correspondingly, for a set $B \subseteq M$ of attributes we define $B^{'}$as the set of objects which have all attributes in B. A **formal concept** of the context (G, M, I) is a pair (A, B) with $A \subseteq G$, $B \subseteq M$, $A^{'} = B$ and $B^{'} = A$. $\mathcal{B}(G, M, I)$ denotes the set of all concepts of context (G, M, I) and forms a complete lattice (so called Galois or concept lattice). For more details, see [8].

Galois lattice may be visualized using so-called Hasse diagram. In this diagram, every node represents one formal concept from the lattice. Nodes are usually labelled by attributes (above the node) and objects (below the node) possessed by a concept. For the sake of clarity it is sometimes used so-called reduced labeling (see fig. 1 for illustration) which means that attributes are shown only at the first node (concept) they appear in. This holds reciprocally for objects. These two labelings are equivalent. Detailed explanation is provided in the Experiment section.

B. Singular value decomposition

Theorem 1: Let A is an $m \times n$ rank-r matrix. Let $\sigma_1 \geq \ldots \geq \sigma_r$ be the eigenvalues of a matrix $\sqrt{AA^T}$. Then there are orthogonal matrices $U = (u_1, \ldots, u_r)$ and $V = (v_1, \ldots, v_r)$, whose column vectors are orthonormal, and a diagonal matrix $\Sigma = diag(\sigma_1, \ldots, \sigma_r)$. The decomposition $A = U\Sigma V^T$ is called *singular value decomposition* of matrix A and numbers $\sigma_1, \ldots, \sigma_r$ are *singular values* of the matrix A.

Using this theorem we can obtain a decomposition of the original matrix A. We have at most r non-zero singular numbers, where rank r is the smaller of the two matrix dimensions. Because the singular values usually fall quickly, we can take only k greatest singular values and create a k-reduced singular decomposition of A. We call $A_k = U_k E_k V T_k$ a k-reduced singular value decomposition (rank-k SVD).

Theorem 2: (Eckart-Young) Among all $m \times n$ matrices C of rank at most k, A_k is the one that minimises $||A_k - A||_F^2 = \sum_{i,j}(A_{i,j} - C_{w,j})^2$.

Briefly said, SVD allows us to decompose one matrix into several others. By multiplying them back we can obtain the original matrix. Another choice is to remove some part of decomposed matrices before the multiplication, which will give us matrix similar to the original one. For more details please see [13]. Description of Non-negative matrix factorization (NMF) and Semidiscrete decomposition method (SDD) is omitted due to the lack of space. For the purpose of our paper, this method works in a similar way as the two mentioned above. Detailed explanation can be found in [12], [9].

C. Correlation dimension

The idea behind the Correlation dimension comes from the theory of Fractal dimension and is based on studying the distance between two random data points. Suppose we have a binary dataset D containing $|D|$ objects and K attributes. Consider random variable (denoted by Z_D) whose value is L_1 distance (attributes used as coordinates) between two randomly chosen objects from D. The distance varies from 0 (objects have the same attributes) to K (objects differ in all attributes). Now we can define the function $f : \mathbb{N} \to \mathbb{R}$ as $f(r) = \mathbb{P}(Z_D < r)$. For a given dataset D, we can compute the set of points $\mathcal{I}(D, r_1, r_2, N)$ by $\left\{(\log r, \log f(r)) \mid r = r_1 + \frac{i(r_2 - r_1)}{N}, i = 0 \ldots N\right\}$. The correlation dimension $\mathrm{cd}_{\mathrm{R}}(D, r_1, r_2, N)$ for a binary dataset D and parameters r_1, r_2 is the slope of the least-squares linear approximation $\mathcal{I}$. One would expect that the dimension of dataset with K independent attributes is K. To achieve this, we can normalize the result using random binary dataset having K independent variables such that the probability of i-th variable being one is equal to

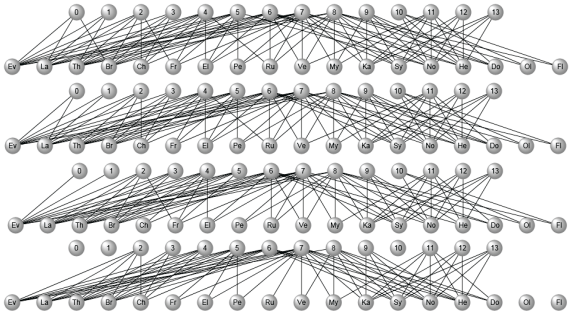

Figure 2. Social network - before and after reduction to ranks 8, 5, 3

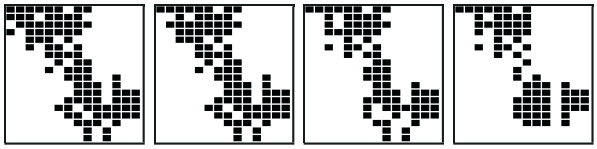

Figure 3. Context visualization (original, rank 8, 5, 3)

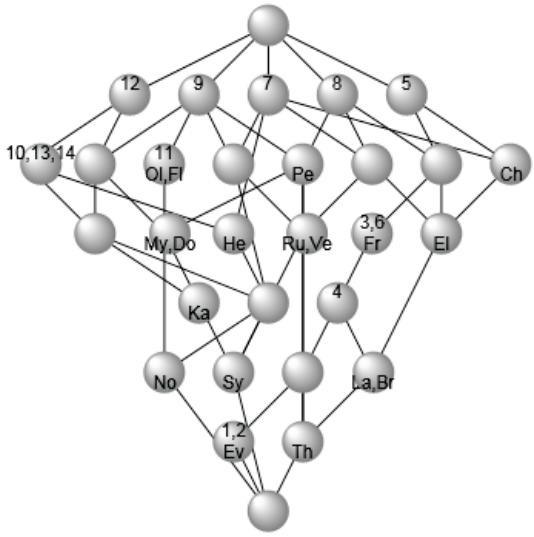

Figure 4. Galois lattice at rank 5

the probability of randomly chosen object from dataset D having i-th attribute. For more details see [15].

III. Experiments

A. Real-world experiment

In our first example, we will use well known dataset from [3]. It contains information about participation of 18 women in 14 social events during the season. This participation can be considered as two-mode network or as formal context (binary matrix with rows as women and columns as social events). Visualization of this network as bipartite graph can be seen in the upper part of figure 2. Events are represented by nodes on the first row. These nodes are labelled by the event numbers. The second row contains nodes representing women. These nodes are labelled by two first letters of their names – Evelyn, Laura, Theresa, Brenda, Charlotte, Frances, Eleanor, Pearl, Ruth, Verne, Myrna, Katherine, Sylvia, Nora, Helen, Dorothy, Olivia, Flora. Participation of the woman in the event is represented by edge between corresponding nodes. Illustration of the formal context can be seen in the left part of figure 3. Rows correspond to women (in the order mentioned earlier), columns to events. Black rectangle denotes participation.

Now, let's describe the computed Galois lattice (figure 1). Each node in the graph represents one formal concept. Every concept is a set of objects (women in this case) and set of attributes (events). Edges express the ordering of concepts. Reduced labelling is used here. Therefore if concept has an attribute (event), every connected concept lying under the labelled one contains the attribute too, and vice versa.

The lattice contains all combinations of objects and attributes present in the data. One can easily read that Sylvia participated in all events that Katherine did. Also everyone who participated in the events 13 and 14, also participated in the event 10. The reasons for these nodes to be separate, are the women Dorothy and Myrna that took part in the event 10, but not in the events 13 and 14.

Due the high number of nodes and edges, many interesting groups and dependencies are hard to find. Now we will try to reduce the formal context to lower dimension and observe the changes. We have performed reduction (using NMF method) of original 18x14 context to ranks 13, ..., 1 and computed corresponding Galois lattices. Modified contexts can be seen in the remaining part of figure 3. Visualization of network into bipartite graph (fig. 2) reveals some changes (e.g. cutting the events 0 and 1 off), but is still too complicated. The Galois lattice can give us better insight. Detailed look at the reduced lattice (fig. 4 for rank 5) shows that the general layout has been preserved as well as the most important properties (e.g. mentioned implication about Sylvia and Katherine). The reduction to rank 5 caused merging of nodes previously marked by attributes 10, 13, 14 (which we have discussed earlier).

B. Synthetic datasets

To analyse results of described approach on larger data, we have generated synthetic binary dataset with 400 rows, 40 attributes and 20% density. This corresponds to two-mode network with 400 subjects and 40 events. Each subject participated at average in 8 events.

The table I contains results of this experiment. We have tested three different reduction methods - NMF, SVD and SDD. First column of each group contains the number of formal concepts in computed lattice. Different rows correspond to different ranks of reduction (first one contains information

	NMF		SVD		SDD	
	$\lvert\mathcal{B}\rvert$	ncd	$\lvert\mathcal{B}\rvert$	ncd	$\lvert\mathcal{B}\rvert$	ncd
original	15477	39	15477	39	15477	39
rank 35	10672	43	15459	39	7750	31
rank 30	5429	35	15127	38	4747	23
rank 25	2665	30	14621	28	2824	23
rank 20	1016	25	11831	29	1377	17
rank 15	348	21	7288	19	514	14
rank 10	149	17	3322	10	169	8
rank 5	56	6	526	6	4	4

Table I
SYNTHETIC DATASET REDUCTION

about original data). Second column contains normalized correlation dimension.

Since the original data have been created as uncorrelated, their normalized correlation dimension is close to the number of columns. The reduction tries to resemble the original data maximally, so it often preserves repeatedly appearing patterns. Therefore we expect ncd to decrease during the rank reduction. Computed results verify this expectation.

To estimate roughly the ratio of reduction, one does not have to compute the whole original lattice. The normalized correlation dimension – which is computed more rapidly and using formal context only – can be used to do this.

IV. CONCLUSIONS

We have seen that Galois lattice is suitable for displaying dependencies in two-mode network data. The restrictive factor is the size and inner structure of input data. Using matrix factorization methods, we can simplify the structure to allow better insight into the data, but still to retain the most important properties.

During the experiments we have observed worsen results of normalized correlation dimension when dealing with contexts with low rows/columns ratio (near square contexts). This can be explained by the probabilistic nature of independent context generation (variables can be easily reassigned after column permutations). This influence can be partially eliminated by averaging the results among several tries.

In our future work we would like to analyse the effects of different reduction methods. It would be also interesting to find some rules and limitations of reduction.

REFERENCES

[1] L. A. Adamic, N. Glance: The political blogosphere and the 2004 US election: divided they blog, Proceedings of the 3rd international workshop on Link discovery, pp. 36–43 (2005)

[2] R. Belohlavek, V. Sklenar: Formal concept analysis constrained by attribute-dependency formulas ICFCA, vol. 3403, pp. 176–191 (2005)

[3] A. Davis, B. B. Gardner, M. R. Gardner: Deep South: A Social Anthropological Study of Caste and Class, University of Chicago Press (1965)

[4] M. Farach-Colton, Y. Huang: A linear delay algorithm for building concept lattices, Combinatorial Pattern Matching: 19th Annual Symposium (2008)

[5] L. C. Freeman: Graphical Techniques for Exploring Social Network Data, Models and Methods in Social Network Analysis (2005)

[6] L. C. Freeman: Visualizing social networks, Journal of social structure, vol. 1 (2000)

[7] L. C. Freeman, D. R. White: Using Galois Lattices to Represent Network Data, Sociological Methodology, vol. 23, pp. 127–146 (1993)

[8] B. Ganter, R. Wille: Formal Concept Analysis: Mathematical Foundations, Springer-Verlag, New York (1997)

[9] T. G. Kolda, D. P. O'Leary: A semidiscrete matrix decomposition for latent semantic indexing information retrieval, ACM Transactions on Information Systems (TOIS), vol. 16, pp. 322–346 (1998)

[10] M. Kudelka, V. Snasel, Z. Horak, A. E. Hassanien: Web Communities Defined by Web Page Content, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (2008)

[11] S. O. Kuznetsov, S. A. Obedkov: Comparing Performance of Algorithms for Generating Concept Lattices, Journal of Experimental and Theoretical Artificial Intelligence, vol. 14, pp. 189–216 (2002)

[12] D. Lee, H. Seung: Learning the parts of objects by non-negative matrix factorization, Nature, vol. 401, pp. 788–791 (1999)

[13] T. Letsche, M. W. Berry, S. T. Dumais: Computational methods for intelligent information access, Proceedings of the 1995 ACM/IEEE Supercomputing Conference (1995)

[14] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, S. M. Dawson: The bottlenose dolphin community of Doubtful Sound features a large proportion of long-lasting associations. Can geographic isolation explain this unique trait?, Behavioral Ecology and Sociobiology 54, pp. 396–405 (2003)

[15] N. Tatti, T. Mielikainen, A. Gionis, H. Mannila: What is the dimension of your binary data?, Proceedings of the Sixth International Conference on Data Mining, pp. 603–612 (2006)