On Social Networks Reduction*

Václav Snášel[1], Zdeněk Horák[1], Jana Kočíbová[1], and Ajith Abraham[2]

[1] VSB Technical University Ostrava, Czech Republic
{vaclav.snasel,zdenek.horak.st4,jana.kocibova.st1}@vsb.cz
[2] Center of Excellence for Quantifiable Quality of Service
Norwegian University of Science and Technology, Norway
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 use combination of Formal Concept Analysis and well-known matrix factorization methods to address computational complexity of social networks analysis and clarity of their visualization. The goal is to reduce the dimension of social network data and to measure the amount of information, which has been lost during the reduction. Presented example containing real data proves the feasibility of our approach.

1 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.

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.

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 purpose data analysis method, in the 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.

* This research was supported in part by Czech Science Foundation (GACR) project 201/09/0990.

J. Rauch et al. (Eds.): ISMIS 2009, LNAI 5722, pp. 533–541, 2009.

1.1 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 concept 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 [1]. 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. Our aim is to allow processing of larger amount of data and our approximation approach is compatible with the two 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-called normalized correlation dimension which can be seen as the number of independent variables in the dataset. This idea comes from the field of fractal dimension. Another way could be to directly compare the results from the original and reduced datasets. [3] introduces the modification of classical Lorenz curve to describe dissimilarity between presence-absence data.

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.

2 Preliminaries

2.1 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 lattice**). For more details, see [9], [8].

Galois lattice may be visualized by so-called Hasse diagram. In this diagram, every node represents one formal concept from the lattice. Nodes are usually labeled 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.

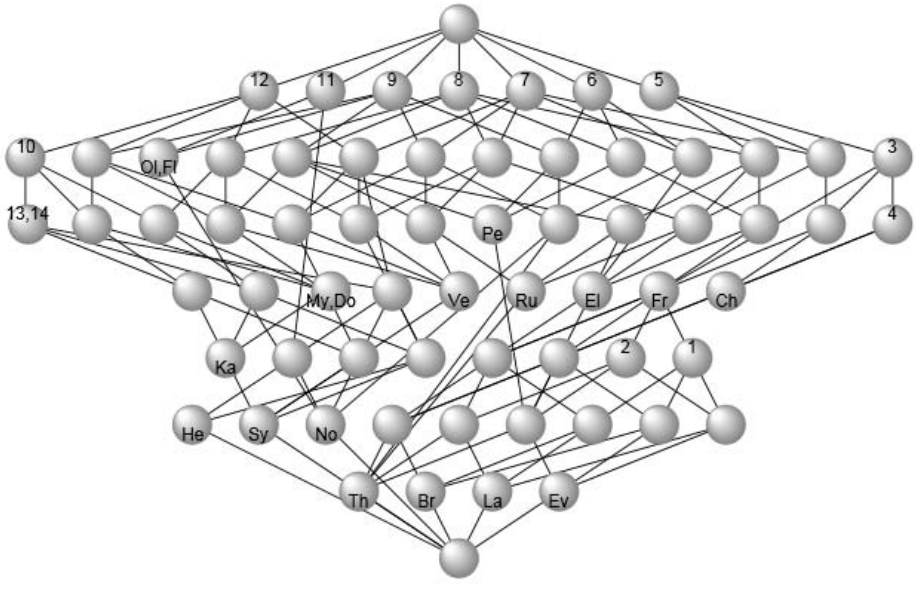

Fig. 1. Concept lattice before reduction

2.2 Non-negative Matrix Factorization

Matrix factorization methods decompose one – usually huge – matrix into several smaller. Non-negative matrix factorization differs by the use of constraints that produce non-negative basis vectors, which make possible the concept of a parts-based representation.

Common approaches to NMF obtain an approximation of V by computing a (W, H) pair to minimise the Frobenius norm of the difference $V - WH$. Let $V \in R^{m \times n}$ be a non-negative matrix and $W \in R^{m \times k}$ and $H \in R^{k \times n}$ for

$0 < k \ll min(m, n)$. Then, the objective function or minimisation problem can be stated as $min\|V - WH\|^2$ with $W_{ij} > 0$ and $H_{ij} > 0$ for each i and j. There are several methods for computing NMF. We have used the multiplicative method algorithm proposed by Lee and Seung [13], [12].

Description of Singular Value Decomposition (SVD) and Semidiscrete decomposition method is omitted due to the lack of space. For the purpose of our paper, these methods work in a similar way as the two mentioned above. Detailed explanation can be found in [14] and [10].

2.3 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 $f(r) = \mathbb{P}(Z_D < r)$ and the set of points

$$\mathcal{I}(D, r_1, r_2, N) = \left\{ (\log\, r, \log\, f(r)) \mid r = r_1 + \frac{i(r_2 - r_1)}{N}, i = 0 \ldots N \right\}.$$

The correlation dimension $\mathrm{cd_R}(D, r_1, r_2, N)$ 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 ith variable being one is equal to the probability of randomly chosen object from dataset D having ith attribute. For more details see [15].

2.4 Lorenz Curves

To evaluate similarity we can use Lorenz Curves, an approach well-known from economy, in the way proposed in [3]. Let's suppose we have two presence-absence (binary) arrays $r = (x_i)_{i=1,\ldots,N}$ and $s = (y_i)_{i=1,\ldots,N}$ of dimension N. In the same manner as we normalize vectors, we can create arrays a_i and b_i by dividing each element of the array by their total sum. Formally $a_i = \frac{x_i}{T_r}, b_i = \frac{y_i}{T_s}, \forall i = 1, \ldots, N;$, where $T_r = \sum_{j=1}^{N} x_j$ and $T_s = \sum_{j=1}^{N} y_j$. Next, we can compute difference array $d = (d_i)_i = 1, \ldots, N$ as $d_i = a_i - b_i$, ordered from the largest value to the smallest one. By putting $c_i = \sum_{j=1}^{i} d_j$ we obtain the coordinates of the Lorenz similarity curve by joining the origin $(0, 0)$ with the points of coordinates $(\frac{i}{N}, c_i)_{i=1,\ldots,N}$.

3 Experiments

3.1 Real-World Experiment

In our first example, we will use well known dataset from [2]. 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 3. Events are represented by nodes on the first row. These nodes are labeled by the event numbers. The second row contains nodes representing women and are labeled by two first letters of their names. Participation of the woman in the event is represented by edge between corresponding nodes. Illustration of the formal context (resp. binary matrix) can be seen in the left part of figure 4.

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 corresponding attributes (events). Edges express the ordering of concepts. Aforementioned reduced labeling is used here. 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.

After reduction. 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

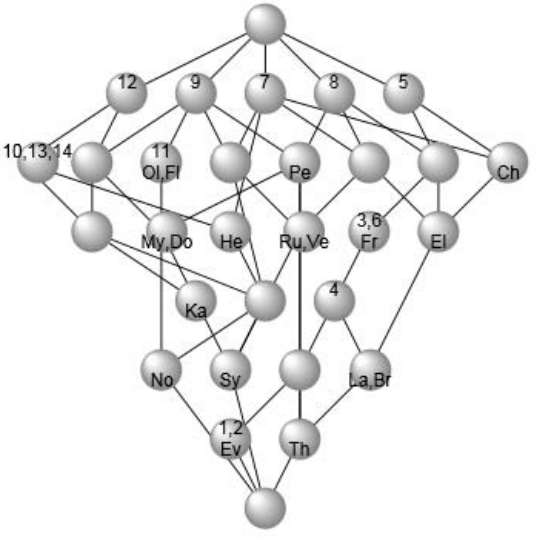

Fig. 2. Concept lattice at rank 5

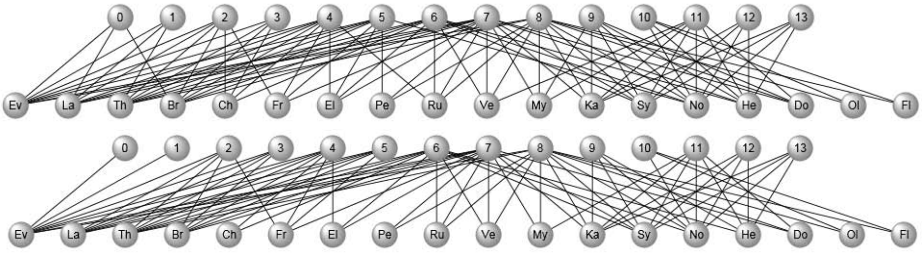

Fig. 3. Social network - before and after reduction to rank 5

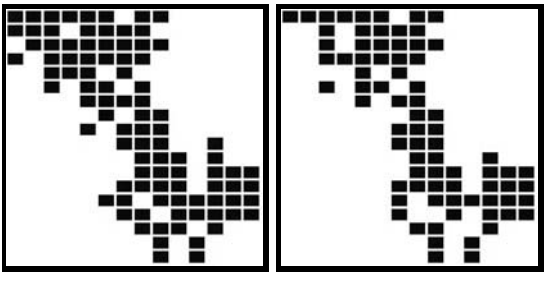

Fig. 4. Context visualization (original, rank 5)

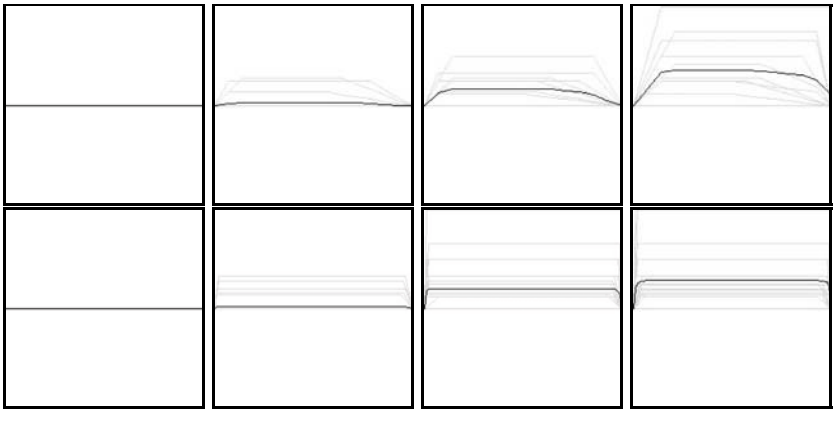

Fig. 5. Lorenz curve comparing contexts (first row) and lattices (second row) - before (first column) and after reduction to ranks 8, 5, 3 (remaining columns)

context to lower dimension and observe the changes. We have performed reduction of original 18x14 context to lower ranks and computed corresponding concept lattices. For illustration we have selected results obtained for rank 5 using NMF method. Modified context can be seen in the remaining part of figure 4. Visualization of network into bipartite graph (fig. 3) reveals some changes, but is still too complicated. The concept lattice can give us better insight. Detailed

look at the reduced lattice (fig. 2 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).

To illustrate the amount of reduction, we can compute similarity between the original and the reduced context and draw Lorenz curves (see first row of fig. 5). A larger area under the curve means higher dissimilarity (lower similarity). Because we compare the context using object-by-object approach, we obtain several curves (drawn using gray color on the figure). To simplify comparison, we have averaged these curves (result drawn using black color). In the same manner, we have computed these curves for formal concepts (second row of fig. 5).

3.2 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 1 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 about original data). Second column contains normalized correlation dimension (ncd).

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. Since the

Table 1. Reduction progress for synthetic dataset (400x40)

	NMF		SVD		SDD	
	$\mid \mathcal{B} \mid$	ncd	$\mid \mathcal{B} \mid$	ncd	$\mid \mathcal{B} \mid$	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

probabilistic nature of ncd computation, we can expect more precise results for datasets containing larger number of objects.

4 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.

This approach has potentially many uses - for example generating fast preview of large social network data, approximate analysis of huge World Wide Web data where exact analysis is computationally unmanageable, etc. As we have mentioned in the introduction, the complexity of many algorithms involved in concept lattice analysis bears (e.g. linearly or exponentially) upon the number of concepts. Therefore the ratio of concepts reduction gives us directly (knowing the complexity of used algorithm) the speed up of computation.

Important fact is, that the progress of reduction is gradual. Every small change in the formal context made during the reduction can be seen as a small change in the corresponding concept lattice. User may be also involved to decide what amount of reduction is suitable to his purpose. Taking the changes in the context into account, we are also able to (considering all the changes made in affected objects and attributes) reconstruct the unreduced concept lattice or its part. This fact maybe useful when more precise result are needed at the detailed level.

From the results it may look like SVD being the best method for reduction, because for fixed rank it gives more concepts than other methods. However, the situation is not that simple. For example the NMF, due the mentioned parts-based representation, uses more natural and independent factors and therefore gives more intuitive results. Thus in our future work we would like to analyse the effects of different reduction methods in detail and illustrate the usage on practical problems.

References

1. Belohlavek, R., Sklenar, V.: Formal concept analysis constrained by attribute-dependency formulas ICFCA. In: Ganter, B., Godin, R. (eds.) ICFCA 2005. LNCS (LNAI), vol. 3403, pp. 176–191. Springer, Heidelberg (2005)
2. Davis, A., Gardner, B.B., Gardner, M.R.: Deep South: A Social Anthropological Study of Caste and Class. University of Chicago Press, Chicago (1965)
3. Egghe, L., Rousseau, R.: Classical retrieval and overlap measures satisfy the requirements for rankings based on a Lorenz curve. Information Processing and Management 42, 106–120 (2006)
4. Farach-Colton, M., Huang, Y.: A linear delay algorithm for building concept lattices. In: Ferragina, P., Landau, G.M. (eds.) CPM 2008. LNCS, vol. 5029, pp. 204–216. Springer, Heidelberg (2008)
5. Freeman, L.C.: Graphical Techniques for Exploring Social Network Data. In: Models and Methods in Social Network Analysis (2005)

6. Freeman, L.C.: Visualizing social networks. Journal of social structure 1 (2000)
7. Freeman, L.C., White, D.R.: Using Galois Lattices to Represent Network Data. Sociological Methodology 23, 127–146 (1993)
8. Ganter, B., Wille, R.: Applied Lattice Theory: Formal Concept Analysis. In: Grätzer, G.A. (ed.) General Lattice Theory, pp. 592–606. Birkhäuser, Basel (1997)
9. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, New York (1997)
10. Kolda, T.G., O'Leary, D.P.: A semidiscrete matrix decomposition for latent semantic indexing information retrieval. ACM Transactions on Information Systems (TOIS) 16, 322–346 (1998)
11. Kuznetsov, S.O., Obedkov, S.A.: Comparing Performance of Algorithms for Generating Concept Lattices. Journal of Experimental and Theoretical Artificial Intelligence 14, 189–216 (2002)
12. Lee, D., Seung, H.: Algorithms for Non-Negative Matrix Factorization. Advances in Neural Information Processing Systems 13, 556–562 (2001)
13. Lee, D., Seung, H.: Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791 (1999)
14. Letsche, T., Berry, M.W., Dumais, S.T.: Computational methods for intelligent information access. In: Proceedings of the 1995 ACM/IEEE Supercomputing Conference (1995)
15. Tatti, N., Mielikainen, T., Gionis, A., Mannila, H.: What is the dimension of your binary data? In: Proceedings of the Sixth International Conference on Data Mining, pp. 603–612 (2006)