

Network analysis of repositories

Vladimir Batagelj, Iztok Kavkler, and Matija Lokar

University of Ljubljana, FMF, Department of Mathematics,
Jadranska ulica 19, 1000 Ljubljana, Slovenia,
vladimir.batagelj@fmf.uni-lj.si,
WWW home page: <http://vlado.fmf.uni-lj.si>

Abstract. In the paper an approach to analysis of the structure of repositories based on network analysis is proposed. The repository metadata can be transformed into several (one-mode or two-mode) networks. Using the methods of network analysis interesting substructures in these networks can be detected. The proposed approach is illustrated with some analyses of SIO (Slovenian Educational Network) repository.

1 Networks from repositories

A metadata in repository consist of records describing each resource. The resource can be stored in the repository or only described and stored elsewhere. Each resource has a unique **ident** and is described by different properties such as: author(s), 'publication' date, language, subject field(s), keywords, ... (see LRE [7]).

Let *Props* be the set of values (range) of selected property *P*. If this set is not discrete or is too large we discretize it according to some partition into classes. Then we can produce the corresponding two-mode (defined on two different sets) network

$$[Idents, Props] = \{(r, p) : p \text{ is a value of } P \text{ in the description of resource } r\}$$

Since so obtained networks have the same first set *Idents* they are compatible and therefore, using network multiplication \star (Batagelj, Mrvar, 2006 [4]), additional networks can be derived from them.

For example, assume that we produced from the repository the two-mode networks $[Idents, Keywords]$, $[Idents, Subjects]$ and $[Idents, Ages]$. Then we can get using network multiplication the following derived networks:

- network of coappearances of keywords:

$$[Keywords, Keywords] = [Idents, Keywords]^T \star [Idents, Keywords]$$

the weights of the arcs are frequencies of coappearances – the number of resources that have both keywords in common;

- network of coappearances of keywords and subjects:

$$[Keywords, Subjects] = [Idents, Keywords]^T \star [Idents, Subjects]$$

the weights of the arcs are frequencies of coappearances of a keyword and subject – the number of resources for selected subject containing also given keyword;

- network of coappearances of subjects and ages:

$$[Subjects, Ages] = [Idents, Subjects]^T \star [Idents, Ages]$$

the weights of the arcs are frequencies of coappearances of subject and age – the number of resources on the subject for a given age of kids.

The obtained networks can be visually inspected (if they are of moderate size) or analyzed using different methods of network analysis (Wasserman and Faust, 1994 [9]; de Nooy, Mrvar, Batagelj, 2005 [5])

2 Example: Some analyses of SIO Repository

For illustration we produced the $[Idents, Keywords]$ network from the SIO (Slovenian Educational Network) Repository [8]. In the June 2007 version it contained 4470 resources that were described by 7076 keywords.

This two-mode network can be analyzed using different direct methods such as: 4-rings weights islands, two-mode cores, and two-mode hubs and authorities (Ahmed et al., 2007 [1]).

We describe here the 4-rings weights islands approach. A good indicator of dense parts in two mode network are the 4-rings weights – to each arc the number of 4-rings (closed chains of length 4) containing the arc is assigned as its weight. The islands algorithm (Zaveršnik, Batagelj, 2004 [10]) identifies islands – subnetworks with a minimal spanning tree that has all values larger than the values on the arcs linking the island to the remaining network.

Using the program *Pajek* (Batagelj, Mrvar, 2007 [2]) we first determined the 4-rings weights and afterward applied the islands algorithm with parameters: the smallest size of an island is $k = 4$ and the largest size of an island is $K = 159$. We obtained 238 such islands; 25 of the of size at least 10.

The largest island deals with (recreational) mathematics. The other topics of the islands are: photography, Slovenian language, Soča river front (First world war), foreign language learning, psychology, numbers, anatomy, programming languages, operation systems, drugs, Carst, insects, library, projects and competitions, dance, astronomy, ships, ...

Multiplying the two-mode network with itself produces two one-mode networks: the network of coappearances of keywords $[Keywords, Keywords]$ but also the network of associations between resources

$$[Idents, Idents] = [Keywords, Idents]^T \star [Keywords, Idents]$$

The weight of an arc counts the number of keywords that the corresponding vertices – resources have in common.

To analyze the keywords network we first computed the p_S -cores values (Batagelj, Zaveršnik, 2002 [3]). A subset of vertices C determines a p_S -core at level t iff for each vertex $v \in C$ the sum of weights of lines that have v as an endpoint and also the other endpoint belongs to C is at least t ; and C is a maximal such set. The largest value of t such that a p_S -core at level t exists containing vertex v is called the p_S -core number of vertex v .

Afterward we determined the vertex islands for p_S -cores values ($k = 4, K = 150$). There are 60 such islands. The largest has 129 vertices (see Figure 1). The remaining islands are much smaller: 16, 10, $3 \times 8, 3 \times 7, 9 \times 6, 12 \times 5, 30 \times 4$.

The strongest subgroup in the main island is a tetrahedron (mathematics, Derive, learning sheet, time). The upper part of the main group contains different subjects (music, physics, geography, ...) and at the left side computer programming keywords (internet, informatics, programming, computer science, java, manual). It is connected to the subgroup at the right side through another tetrahedron (addition, subtraction, multiplication, division). The subgroup contains mainly the keywords from elementary mathematics. The group in the lower part of the picture contains keywords related to advanced topics in elementary mathematics and geometry.

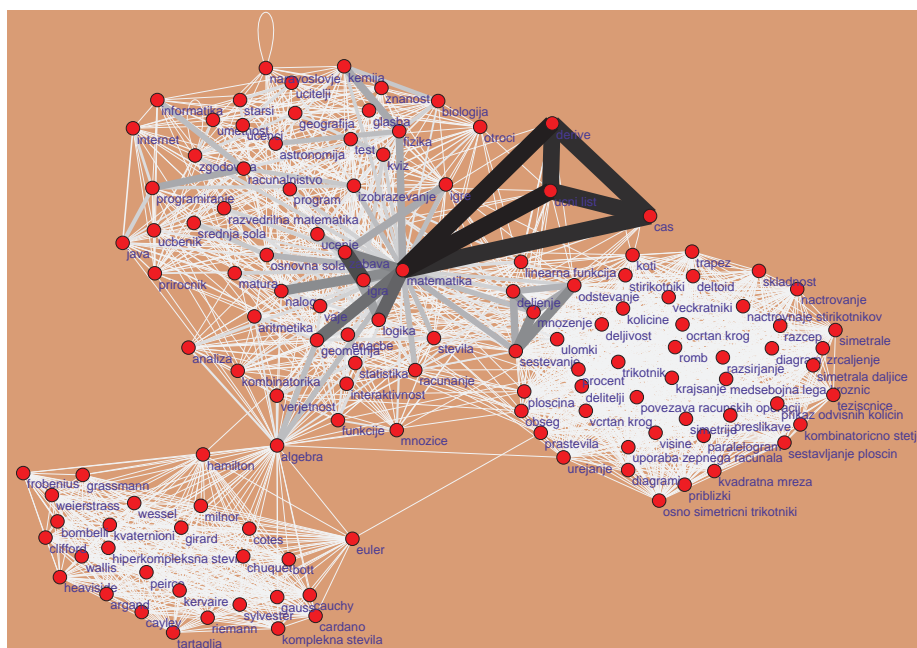


Fig. 1. The largest vertex island for p_S -cores values in keywords network of SIO Repository

3 Conclusions

In the paper we proposed a network analysis approach of the structure of repositories. As an illustration we applied it to the two-mode network [*Idents, Keywords*] of SIO. In repositories there are many other properties for which the corresponding networks can be derived and analyzed.

An interesting question for further research is which of these results can be used to help the user when searching for resources, and how.

4 Acknowledgment

The work presented in this paper is partially supported by the European Commission under the Information Society Technologies (IST) program of the 6th FP for RTD – as part of the CALIBRATE project, contract IST-28025. The authors are solely responsible for the content of this paper. It does not represent the opinion of the European Commission, and the European Commission is not responsible for any use that might be made of data appearing therein.

References

1. Ahmed, A., Batagelj, V., Fu, X., Hong, S.-H., Merrick, D., and Mrvar, A.: Visualisation and analysis of the Internet movie database. Asia-Pacific Symposium on Visualisation 2007, 17–24.
2. Batagelj V., Mrvar A.: Pajek – program for analysis and visualization of large networks, 1996-2007. <http://vlado.fmf.uni-lj.si/pub/networks/pajek/>
3. Batagelj V., Zaveršnik M.: Generalized Cores. <http://arxiv.org/abs/cs.DS/0202039>
4. Batagelj V., Mrvar A.: Multiplication of networks, 2006. submitted.
5. de Nooy, W., Mrvar, A. and Batagelj V.: Exploratory Social Network Analysis with Pajek, Cambridge University Press, 2005.
6. Kleinberg J.: Authoritative sources in a hyperlinked environment. In Proc 9th ACM/IEEE Symposium on Discrete Algorithms, 1998, p. 668–677.
7. LRE – Learning Resource Exchange. http://insight.eun.org/ww/en/pub/insight/interoperability/monthlyinsight/lre_presentations.htm
8. Slovensko izobraževalno omrežje – Slovenian Educational Network. <http://sio.edus.si>
9. Wasserman S., Faust K.: Social Network Analysis: Methods and Applications. Cambridge University Press, 1994.
10. Zaveršnik M., Batagelj V.: Islands. Slides from *Sunbelt XXIV, Portorož, Slovenia, 12.-16. May 2004* <http://vlado.fmf.uni-lj.si/pub/networks/doc/sunbelt/islands.pdf>