Identifying overlapping thematic structures in networks of papers: A comparison of three approaches

Alexander Struck[1], Jochen Gläser[2], Michael Heinz[1] and Frank Havemann[1]

[1] {Alexander.Struck | Michael.Heinz | Frank.Havemann}@ibi.hu-berlin.de
Institut für Bibliotheks- und Informationswissenschaft, Humboldt-Universität zu Berlin, 10099 Berlin (Germany)

[2] Jochen.Glaser@ztg.tu-berlin.de
Zentrum Technik und Gesellschaft, Technische Universität Berlin, 10623 Berlin (Germany)

Introduction

The delineation of scientific fields is a pertinent problem of science studies in general and bibliometrics in particular. Bibliometric research has shown that clusters in networks of papers do not have 'natural' boundaries (Zitt and Bassecoulard, 2006). This is why fields must be delineated by applying thresholds for parameters, which are chosen arbitrarily in terms of 'good structures' for the purposes of the analysis at hand.

However, the problem of delineation might be a consequence of the overlap of thematic structures. In other words, delineation may not be impossible in principle but rather depend on tools that enable the identification of overlapping fields and topics.

The aim of this poster is to compare three recently proposed approaches to the identification of overlapping communities in networks. A first approach starts from hard clusters obtained by any clustering method and fractionally assigns the nodes at the borders between clusters to these clusters (e.g. Wang, Jiao and Wu, 2009). Another approach is based on a hard clustering of links between nodes into disjoint modules, which makes nodes members of all modules that their links belong to (e.g. Ahn, Bagrow and Lehmann, 2010). The third approach constructs *natural communities* of all nodes, which can overlap with each other, by applying a greedy algorithm that maximises local fitness (e.g. Lancichinetti, Fortunato and Kertesz, 2009).

Data

Algorithms for all three approaches were imple-mented and applied to a benchmark graph (the karate club, Zachary, 1977) and to a network of 492 bibliographically coupled papers published 2008 in six information science journals (for details on data see Havemann, Heinz, Struck and Gläser, 2011). We used information science papers in order to be able to assess the clustering solutions and the overlap of modules. The communities obtained and overlaps produced by the three algorithms were assessed on the basis of our knowledge of information science and bibliometrics, the assignment of papers to topics by authors via keywords, and the position of key papers. We have selected the topics *h-index*, *webometrics*, and *bibliometrics* as first examples for evaluation. We start with the bipartite graph of papers and their cited sources and project it on a paper network based on Salton's cosine of bibliographic coupling.

Natural communities of nodes

The construction of a paper's natural community can be interpreted as the construction of its thematic environment from its own 'scientific perspective'. This idea is attractive from a conceptual point of view because it mimics the way in which scientists apply their individual perspectives when constructing their fields.

Our algorithm uses an approach to local fitness proposed by Lancichinetti et al. (2009). A module fitness function relates the summed weight of internal links to the weight of all links of the module. Nodes are added if this enhances the module's fitness. The fitness function depends on a resolution parameter which allows revealing the hierarchical structure of the network. While the algorithm proposed by Lancichinetti et al. (2009) is based on a numerical solution, our own algorithm is based on an analytical alternative (Havemann et al., 2011). It is parameter-free, calculates community-changing resolution levels exactly, and automatically detects the hierarchy of the graph by merging overlapping natural communities (MONC algorithm). Branches in MONC dendrograms do not represent disjoint sets of nodes but overlapping communities which merge when resolution decreases. We found that using cliques instead of nodes as seeds leads to better results. Relevant resolution levels can be identified by searching for intervals in which large resolution changes do not lead to a growth of communities. Leaving a systematic analysis of graph hierarchies for future work we proceeded heuristically to identify the branches in the MONC dendrogram which correspond to one of our preselected topics.

Link Clustering

The clustering of bibliographic links is theoretically attractive because a link between a paper and a cited source is probably the conceptually most homogeneous bibliometric unit. Although there are many cases in which a paper may cite a source for several different reasons, many bibliographic links will represent one theme that links the citing to the cited publication. Based on the assumption of thematic homogeneity, citation links can be hard-clustered, which leads to induced overlapping clusters of papers. The membership degree of a paper to a module corresponds to the part of outgoing citation links of this paper in this link cluster.

The link clustering method suggested by Ahn et al. (2010) has not yet been tested with bipartite graphs. Our tests with the paper-source network showed that we need a similarity measure taking into account that there are many more cited sources than citing papers. We compute the similarity of two citation links as a linear combination of Salton's cosine measures of bibliographic coupling and of co-citation. The coefficients used compensate for the paper-source asymmetry. Ahn et al. (2010) calculate similarities only for link pairs which have a node in common, we for all pairs.

Using the modified similarity measure and applying the Louvain algorithm (Blondel, Guillaume, Lambiotte and Lefebvre, 2008) results in disjoint clusters of citation links and overlapping modules of papers.

Fuzzification of hard clusters

This approach is based on the assumption that hard clustering is able to correctly identify thematic cores and only unable to cope with overlapping boundaries of thematic structures. If this is true, using an established hard clustering algorithm and enhancing it with a fuzzification process might suffice to make it applicable to our networks.

The implemented algorithm takes any hard cluster solution as input and recalculates membership of cluster border nodes. Like Wang et al. (2009) we use the fitness function proposed by Lancichinetti et al. (2009) for the recalculation. However, our algorithm is different from that of Wang et al. in that its recalculation of memberships is independent from the order of nodes. A node's membership degree is determined by its fitness balance with respect to a module. Negative balance means zero membership. The fitness-inherent resolution parameter controls the extent of the overlap, where lower values cause a wider border area to be included into the former hard cluster.

The Louvain algorithm was the first used to generate the input for the fuzzification procedure. Our fuzzification test results show interesting additions to the initial sets which improve the thematic cluster coverage according to keyword and title observation in the fuzzy overlap.

Comparison and discussion

All three tested algorithms result in overlapping clusters of papers and in varying membership degrees. This enables a detailed comparison of assignments to the preselected topics and their overlaps based on the three algorithms and on keywords. An article-title based validation also shows differences and similarities of the results.

The three algorithms have different strengths and weaknesses. Link clustering and fuzzification crucially depend on the performance of the hard clustering solution. The Louvain algorithm maximises modularity i.e. does not operate locally in a strict sense. It also fails in giving a detailed picture of a network's hierarchy (we obtained only two levels of clustering for the paper network as for the link clustering). In further tests of link clustering and fuzzification we will also test local and hierarchical hard-cluster algorithms. Fuzzification also has the potential to fundamentally 'revise' the hard clustering solution if turned into a recursive procedure. The strict locality of MONC makes it a highly appropriate algorithm for paper clustering. Both MONC and link clustering can be used in a strict bottom-up process, i.e. for building networks of papers by starting from single papers.

Acknowledgments

We thank Michael Bärtl for technical assistance with the classification of papers. This work is part of a BMBF project (http://www.bmbf.de/en/) on methods for measuring the diversity of research (http://141.20.126.172/~div).

References

Ahn, Y.-Y., Bagrow, J. P. & Lehmann, S. (2010). Link communities reveal multi-scale complexity in networks. *Nature,* 466, 761–764. arXiv:0903.3178

Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.*, 2008, P10008. arXiv:0803.0476v2

Havemann, F., Heinz, M., Struck, A. & Gläser, J. (2011). Identification of overlapping communities and their hierarchy by locally calculating community-changing resolution levels. *J. Stat. Mech.,* 2011, P01023. arXiv:1012.1269

Lancichinetti, A., Fortunato, S. & Kertesz, J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New J. Phys.*, 11, 033015. arXiv:0802.1218

Wang, X., Jiao, L. & Wu, J., (2009). Adjusting from disjoint to overlapping community detection of complex networks. *Physica* A, 388, 5045–5056.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *J. Anthropol. Res.,* 33, 452–473.

Zitt, M., Bassecoulard, E., (2006). Delineating complex scientific fields by an hybrid lexical citation method: An application to nanosciences. *Inf. Proc. Man.* 42 (6), 1513–1531.