Network Creation: Overview

Christian Borgelt

European Centre for Soft Computing
Calle Gonzalo Gutiérrez Quirós s/n, E-33600 Mieres (Asturias), Spain
christian.borgelt@softcomputing.es

Although networks are a very natural and straightforward way of organizing heterogeneous data, as argued in the introductory chapters, few data sources are in this form. We rather find the data we want to fuse, connect, analyze and thus exploit for creative discoveries, stored in flat files, (relational) databases, text document collections and the like. As a consequence, we need, as an initial step, methods that construct a network representation by analyzing tabular and textual data, in order to identify entities that can serve as nodes and to extract relevant relationships that should be represented by edges.

Rather than simply connect all (named) entities for which there is evidence that they may be related in some way, it is clearly desirable that these methods should try to select edges that have a higher chance of being part of a bisociation (or should at least try to endow such edges with higher weights) and should try to identify nodes that have a higher chance of being a bridging concept. In this way the created networks will be better geared towards the goal of creative information discovery. In addition, we need a representational formalism that allows us to reason about graph relationships, in order to support the network analysis and exploration methods described in Parts III and IV, respectively.

Contributions

Most of the following chapters deal with constructing BisoNets from text document collections, like web pages, (scientific) abstracts and papers, or news clippings. In order to process such data sources, the authors all start with standard text mining techniques for keyword extraction in order to obtain an initial set of node candidates. These candidates may then be filtered in order to identify potential bridging concepts or at least to rank them higher than other terms.

In more detail, the first chapter by Segond and Borgelt [1] simply takes the extracted keywords as the node set of the BisoNet that is to be constructed and focuses on the task of selecting appropriate edges. Since standard measures for the association strength of terms turn out to be of fairly limited value, the authors suggest a new measure, which has become known as "Bison measure" or "bisociation index". This measure is based on the insight that for selecting appropriate edges the similarity of the term weights is at least as important as, if not more important than, the magnitude of these weights.

In contrast to this, the chapter by Juric *et al.* [2] concentrates on the selection and ranking of terms and keywords in order to identify bridging concepts. Starting with a more detailed description of the used text mining techniques and

document representations, the authors provide a thorough overview of a variety of approaches to compute term weights and of several distance measures between vectors that represent documents in a bag-of-words vector space model. From these the authors derive heuristics to rank terms based on their occurrence in two or more document collections. Included here are heuristics relying on classifiers that are trained to distinguish between the document collections, and for which the misclassified terms are interpreted as potential bridging concepts. They show that in this way a significantly higher number of bridging concepts appear at the top of the ranking list than can be expected in a chance ranking.

The chapter by Hynönen *et al.* [3] again emphasizes the relation between terms, but rather than selecting edges for a BisoNet the authors try to identify terms that are connected in a document even though they are usually not in the underlying document collection as a whole. The core idea is that such unusually correlated terms can indicate a new development or a new insight that is described in the corresponding document(s). In order to measure the connection strength, the authors introduce two new aspects: the first consists in measures for the term pair frequency to assess the strength of correlation in a document and the term pair uncorrelation to describe the background of the document collection to which it is compared. The second aspect is that they take the document apart into sentences in order to achieve more fine-grained assessments.

The second chapter by Segond and Borgelt [4] presents a new item set mining technique, which may be applied to text document mining by seeing each term as an item and each document as a transaction of the terms that occur in it. The core idea of the approach is to go beyond terms pairs and to find correlations between multiple terms, which correspond to possible hyperedges in a BisoNet. However, since the standard measure for the selection of item sets, the support (number of transactions containing all items) is not well suited to assess the association of terms, the authors introduce an approach based on the similarity of item covers (sets of transactions containing the items) and develop an efficient algorithm to mine item sets with several such similarity measures.

Finally, the chapter by Kimmig *et al.* [5] discusses a representation and reasoning framework for graphs with probabilistically weighted edges that relies on the ProbLog language. The authors demonstrate how both graphs and graph patterns can conveniently be described in a logical framework and how deductive, abductive and inductive reasoning are supported, as is shown with several precise examples. In addition, modifications of the knowledge base can easily be expressed, including graph simplification, subgraph extraction, abstraction etc. Finally, the authors demonstrate how probabilistic edge weights (interpreted as the probability that an edge is present) can be incorporated and how all discussed logical concepts can be transferred and extended to probabilistic graphs.

Conclusions

Whether graphs are described by explicit graph data structures or in a logical framework (endowed with probabilistic edge weights or not), they are a powerful framework for knowledge representation. However, creating them from

heterogeneous and, in particular, from unstructured data like documents, is a challenging task, especially if one wants to support creative information discovery. Even though standard text mining techniques form the starting point for all of the approaches discussed in this part, they are not sufficient for creating useful BisoNets. As the following chapters demonstrate, several enhancements of the selection of both nodes and edges can increase the chance of obtaining edges that support bisociative discoveries and of identifying nodes that are potential bridging concepts. It has to be conceded, though, that the methods are not perfect yet and that there is a lot of room for improvement. However, the described methods are highly promising and they could be shown to produce significantly better results than known techniques.

References

1. Marc Segond and Christian Borgelt. Selecting the Links in BisoNets Generated from Document Collections. In: Michael R. Berthold, ed. *Bisociative Knowledge Discovery*, LNAI 7250:53–64. Springer-Verlag, Heidelberg, 2012
2. Matjaž Juršič, Borut Sluban, Bojan Cestnik, Miha Grčar, and Nada Lavrač. Bridging Concept Identification for Constructing Information Networks from Text Documents. In: Michael R. Berthold, ed. *Bisociative Knowledge Discovery*, LNAI 7250:65–89. Springer-Verlag, Heidelberg, 2012
3. Teemu Hynönen, Sebastien Mahler, and Hannu Toivonen. Discovery of Novel Term Associations in a Document Collection. In: Michael R. Berthold, ed. *Bisociative Knowledge Discovery*, LNAI 7250:90–102. Springer-Verlag, Heidelberg, 2012
4. Marc Segond and Christian Borgelt. Cover Similarity based Item Set Mining. In: Michael R. Berthold, ed. *Bisociative Knowledge Discovery*, LNAI 7250:103–120. Springer-Verlag, Heidelberg, 2012
5. Angelika Kimmig, Esther Galbrun, Hannu Toivonen, and Luc De Raedt. Patterns and Logic for Reasoning with Networks. In: Michael R. Berthold, ed. *Bisociative Knowledge Discovery*, LNAI 7250:121–142. Springer-Verlag, Heidelberg, 2012