NMF: Network Mining Framework using Topological Structure of Complex Networks

Kouhei Sugiyama, Hiroyuki Ohsaki, and Makoto Imase
Graduate School of Information Science and Technology, Osaka University
{k-sugi,oosaki,imase}@ist.osaka-u.ac.jp

Takeshi Yagi, and Junichi Murayama
NTT Information Sharing Platform Laboratories, NTT Corporation
{yagi.takeshi,murayama.junichi}@lab.ntt.co.jp

1 Introduction

In recent years, several researches using topological structure of complex networks represented by link mining have been performed actively.

Many link mining methods have been proposed, but they have been in differing fields and have each been researched independently [1]. At present, it is therefore difficult to utilize these link mining methods in a mutually coordinated manner.

In this paper, a variety of link mining methods are handled in an integrated manner with previous data mining methods, and a Network Mining Framework (NMF) is proposed.

This paper also presents example usages of the NMF, and discusses how high-level network mining are achieved by using the NMF.

2 NMF (Network Mining Framework)

2.1 Outline

The Network Mining Framework (NMF) is a framework for network mining which is capable of handling a variety of link mining methods and data mining methods in an integrated manner. In network mining, the objects of data mining not only include the structure of data, but also its quantity and semantics.

The NMF is composed of seven types of function block which have a unified input/output format (see Fig. 1). It is possible to implement a variety of network mining methods by freely connecting these function blocks. Since the input/output format of the function blocks is unified, it is possible to freely combine each of the function blocks.

2.2 Input/Output format

The input and output of the NMF function blocks take one of the following formats.

- Graph expressing the topological structure of the network $G = (V, E)$
- Set of vectors $\mathbf{x}$ expressing the attributes of the nodes
- Set of vectors $\mathbf{x}$ expressing the attributes of the links
- Vector $\mathbf{x}$ expressing general values

2.3 Function blocks

- Network analysis block
  Network analysis block is intended for data mining which uses the topological structure of a network (i.e., link mining).

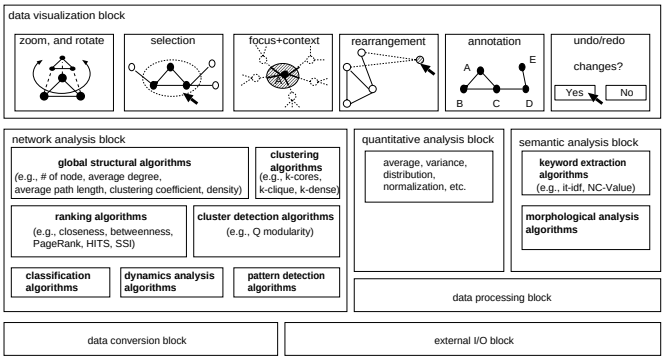

Figure 1: The seven types of function blocks constituting the NMF

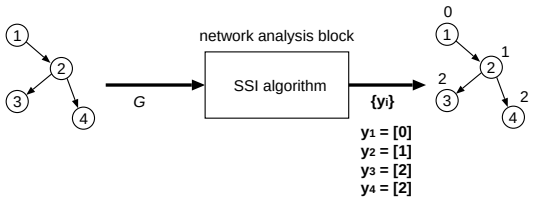

Figure 2: Example network analysis block

  For example, Fig. 2 shows a network analysis block which obtains the SSI (Structural Superiority Index) value [3] for each node from the topological structure of a network. The topological structure of the network, G, is taken as input and the SSI value of each node is output as a set of node attributes $\mathbf{y}_i$.
- Quantitative analysis block
  Quantitative analysis block is intended for data mining focused on the volume of data.
- Semantic analysis block
  Semantic analysis block is intended for data mining focused on the meaning of data.
  For example, Fig. 3 shows a semantic analysis block which extracts keywords from link attributes by means

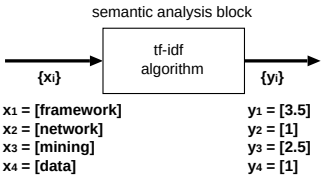

Figure 3: Example semantic analysis block

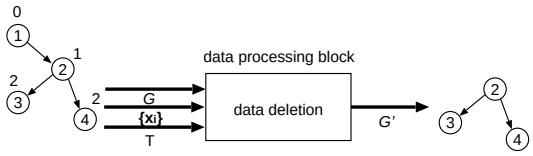

Figure 4: Example data processing block

of the tf-idf method. A set of node attributes $\mathbf{x}_i$ is taken as input, and the tf-idf value of each word is output as a set of link attributes $\mathbf{y}_i$.

- Data processing block
  Data processing block is intended for processing the structure among data (i.e., links), node attributes and link attributes.
  For example, Fig. 4 shows a data processing block which deletes nodes whose node attribute values are in the T% lower ranks. Inputs to the data processing block are the topological structure of the network, G, a set of node attributes $\mathbf{x}_i$, and T, output from the data processing block is topological structure of the network, G'.
- Data conversion block
  Data conversion block is intended for converting external data into a format that can be handled by the NMF, and alternatively, for exchanging the internal data handled by the NMF into other formats.
- Data visualization block
  Data visualization block is intended for visualizing the structure, volume and semantics of data.
- External I/O block
  External I/O block is intended for inputting and outputting external data in real time.

3 Example Usages

3.1 Recursive estimation of node characteristics

The general outline of the RENC (Recursive Estimation of Node Characteristics) method presented in [3] is as follows. Previous node characteristic estimation methods [2] use the complete topological structure of a network when estimating node characteristics. In the recursive estimation method for node characteristics, RENC, the complete topological structure of a network is not used, and node characteristics are estimated using only those nodes and links which are meaningful for node characteristic estimation. Firstly, characteristic estimation is conducted using the complete topological structure of the network. The results are used to eliminate the nodes which are thought to constitute noise during characteristic estimation, as well as all the links attached to these nodes. Node characteristic estimation is once again applied recursively to the partial network obtained in this way.

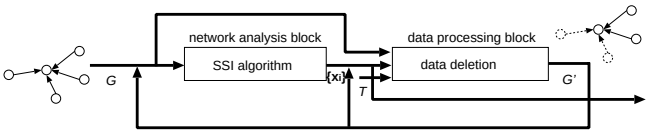

Figure 5: Example usage of the NMF (recursive estimation of node characteristics)

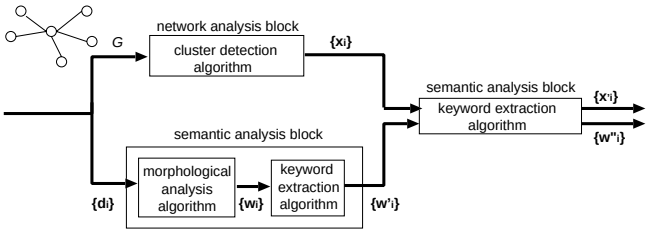

Figure 6: Example usage of the NMF (keyword extraction from network clusters)

Link mining methods such as this may be implemented by connecting a network analysis block which estimates node characteristics with a network analysis block that eliminates nodes thought to be unimportant (Fig. 5).

3.2 Keyword extraction from network clusters

The outline of the method for extracting keywords from network clusters proposed in [4] is as follows. Firstly, for weblog track-back networks, clusters of weblog articles are extracted by means of a cluster detection method which uses the topological structure of a network. For each cluster, keywords are then extracted from the bodies of articles by means of natural language processing methods.

These kinds of link mining and data mining methods can be implemented by connecting a network analysis block which detects network clusters in parallel with a semantic analysis block which extracts keywords from natural language (Fig. 6). In this case, the semantic analysis block which extracts keywords from natural language may be implemented by linking a semantic analysis block which performs morphological analysis with a semantic analysis block which extracts keywords. The implementation is thus possible by linking the blocks connected in parallel as described above to a semantic analysis block which extracts keywords (Fig. 6).

References

[1] L. Getoor and C. P. Diehl. Link mining: A survey. *SIGKDD Explorations*, 7(2):3–12, Dec. 2005.

[2] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[3] K. Sugiyama, H. Ohsaki, and M. Imase. RENC: Recursive estimation of node characteristics using topological structure of complex networks. In *Proceedings of the 7th Asia-Pacific Symposium on Information and Telecommunication Technologies (APSITT 2008)*, pages 59–64, Apr. 2008.

[4] M. Uchida, N. Shibata, and S. Shirayama. Identification and visualization of emerging trends from Blogosphere. In *Proceedings of International Conference on Weblogs and Social Media*, pages 305–306, Mar. 2007.