A Survey on Graph-Based Topic Visualization

Riva Malik*
FAST-NUCES

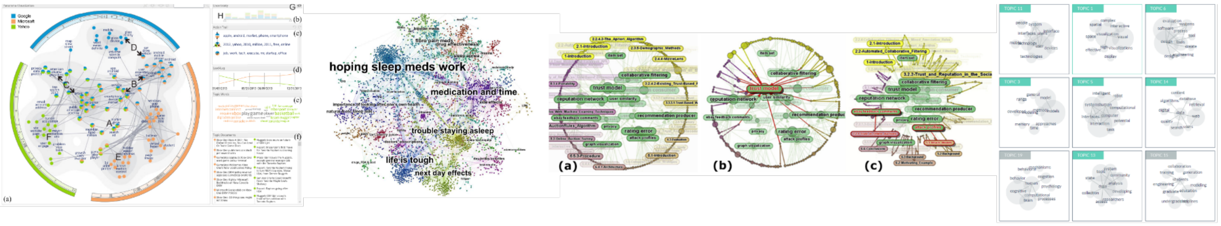

Figure 1: An Overview of Various Graph-Based Topic Visualization Techniques

Abstract

Text mining is the process of discovering knowledge from textual data. Topic extraction is a sub-field of text mining. The results of topic extraction depends on how the raw data is represented for input to topic extraction model. Graph-based representations are effective in capturing features of text. This research aims to explore visualization methods employed by various topic extraction models that employ graph representation techniques.

1 Introduction

Text Mining has recently emerged as sub-field of data mining, which involves extracting unknown,implicit, hidden, and potentially valuable information from unstructured text data. Entity relation recognition,topic extraction, text summarization, sentiment analysis, text categorization, and text clustering are some of the major text mining tasks. Topic extraction also called topic modelling is a probabilistic method of text mining that discovers abstract topics from documents

The representation of text for as input for topic modelling plays an important role in the quality of results.Multiple approaches towards topic modelling employ different text representation techniques. Each representation has the ability to capture different features of document. Table. 1 shows the text representation approaches and the features of the document they are able to capture.

Table 1: Text Representation Techniques

Model	Feature Coverage			
	Term Existence	Term Frequency	Term co-occurrence	Term Ordering
Bag of Words	✓			
Vector Space	✓	✓		
Graph-Based	✓	✓	✓	✓

[1] In the recent studies graph has come forward as a mean of representing text for topic extraction models, this is due to the ability of graph to capture intricate features of document. These researches [2], [3] utilize several types of nodes and edges configurations to encode features of documents at various granularity levels.

Visualization of these complex representations and models is a challenging task, yet it is vital for better understanding. Several studies have been carried out for topic visualization and graph visualization techniques.

1.1 Graph Visualization

Graph visualization helps viewer gain insights of data by converting the data elements and their internal relationships into graphs. Graph visualization leverages the human visual system to support knowledge discovery. The major challenges in graph visualization are visual cluttering, readability and layout setting. In graph visualization node-link layout is considered the most widely used, apart from that space division layout, space nested layout, and 3D layout are also used in some applications [4].

1.2 Topic Visualization

Visual representation of topics and topic models helps viewer in evaluating the model, effectively analyzing the results of the model and facilitating the interpretation of hidden patterns. Various topic visualization techniques include word cloud, theme river, topic river, bar charts, and heat maps.

The aim of this paper is to review work done on visualizing topic extraction using graph based representation approaches. It combines the concepts of topic visualization and graph visualization to create effective visualizations.

2 Graph-Based Topic Visualizations

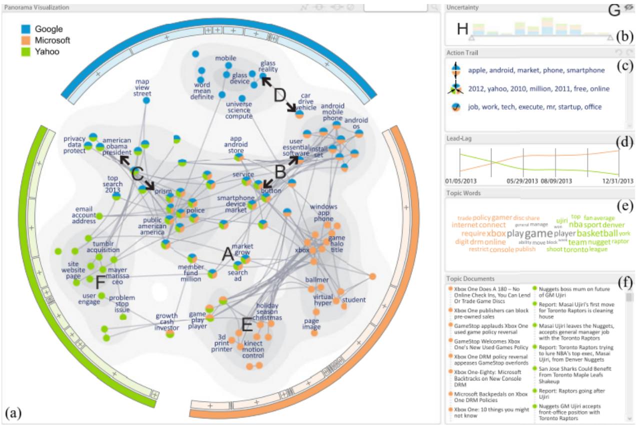

Figure 2: Full Picture of TopicPanorama [5]

Main Idea News articles, blogs, and micro-blogs carry information related to several topics. The topics discussed in these multiple sources are of two types: a set of distinct topics discussed individually by each document and a set of mutual topics. To represent

*e-mail: i201208@nu.edu.pk

complete information in the documents both aspects of the topics need to be presented. In order to gain a comprehensive understanding of the whole picture, the user has to repeatedly switch back and forth between different views which decreases the user experience. To facilitate such an analysis and increase user experience, it is important that the user gets complete picture. The aim of this paper is to gather separate pieces of information about these topics, scattered among different documents, and reconstruct the full picture.

Application Domain The research can be used by journalists, media houses to examine the topics of interest and their correlation, to analyze the temporal patterns and relationship between different corpora with respect to selected topic, and to explore clusters at different granularity [5].

Visualization Technique The TopicPanorama(see Fig. 2) is designed to visualize hierarchical models. The interface consists of three components: a visualization panel, an information panel and a control panel. The implementation of the visualization is also divided into different parts.

The first step is designing the topic hierarchy. The documents contain large number of topics. These topics are represented as hierarchies of topic graph based on Bayesian Rose Tree model. The constraint ensures that the hierarchy of different graphs have similar structure. Hierarchy of each topic graph is generated and then iteratively refined based on hierarchy of other graphs. A radial icicle plot is employed to display topical hierarchies. The angle of the radial plot encodes the number of topics in different topics of each corpus.

The second step is the creation of a density-based graph. A density based graph visualization is evolved that mixes a node-link diagram with density maps to display nodes at chosen level of topic hierarchy. Representative nodes from each cluster are selected and other nodes closed to them are assigned. The node-link diagram is used to illustrate the relationship between nodes and density is utilized to represent the global context. The topic nodes belonging to different corpora are encoded using different colors and the nodes belong to multiple corpora are encoded using pie chart where each slice correspond to a particular corpus. The topic nodes mutual in multiple corpora are placed near center of the layout whereas distinct topic nodes are placed in the area corresponding to the corpus near edges.

The step involves the creation of other panels. The information panel contains additional information about topics and documents related to topics. It displays topic information as word cloud. A lead-lag chart displays the temporal change of topics.

Interactivity is added in form of zooming into the clusters, highlighting topic nodes belonging to particular cluster.Lasso effect is available to select group of keywords to examine their distribution.

Strengths and Weaknesses The visualization has the ability to display the full picture of different corpora at one glance. The channels such as colour and area has been effectively utilized. The weakness is that the hierarchical structure cannot be changed.

Main Idea Exploring large collection of unstructured data, like text is a difficult task. For this purpose analysis tools can be utilized to explore the data. Topic modelling based tools helps in analysis of terms and topics in the textual data. These tools can also prove helpful in determining the correctness of the model by analyzing the results. Existing work on topic modelling use the bag of word representation. The topic model presents the corpus as an unrelated set of topics where each topic is a probability distribution over words. Thus the analysis tools also use word cloud representation of topics. The word cloud representation lacks in describing the relationship between the words [6].

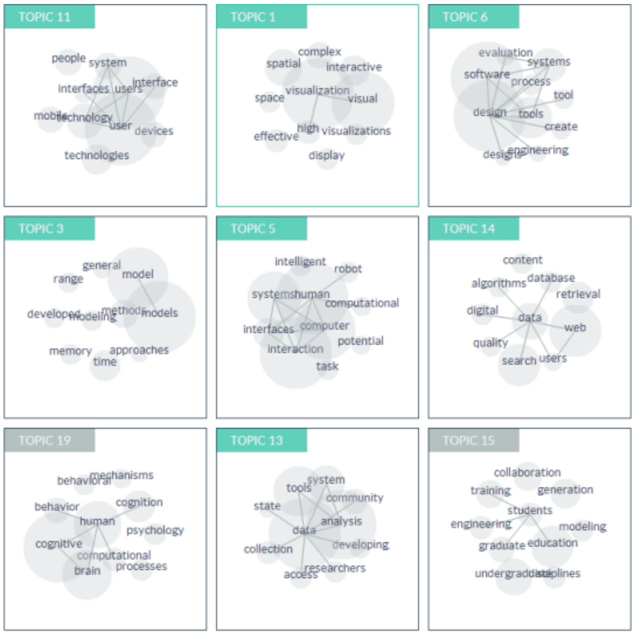

Figure 3: Full Picture of Graph-In-Box Layout [6]

Application Domain This research can be directly applied to quickly overview the results and correctness of topic models. It provides information about words related to a certain topic, a certain topic and similar topics, topic co-variance, existence of same word in different topics [6].

Visualization Technique The visualization [6] is inspired from GIB(graph-In-Box) layout. Each topic is represented as a force directed graph. The node of graph represents word and edge between nodes represents frequent co-occurrence of the word nodes. The probability of each word in a given topic is encoded through the size of node in the topic graph. The layout of each topic graph represents the topical co-variance. The most connected topic is placed at the center of layout and least connected topics at the corners. The topic co-variance represents topics which occur together in the same document.

Interactivity is added in form of hover function. Hovering over a word node highlights the same word in other topics.

Strengths and Weaknesses By mapping word probability to the area of the nodes instead of the height of node the layout becomes independent of word length. Furthermore, circles can overlap without affecting a user's ability to visually separate them, and lead to more compact and less cluttered visual layout.

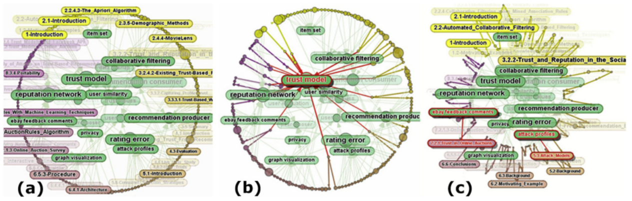

Figure 4: Full Picture of TopicNets [7]

Main Idea TopicNets [7] is designed for visualization and interactive analysis of large collection of text documents through web interface. The main idea is to support interactive thematic modelling in real-time.

Application Domain This research serves as a real-time web application for analyzing and exploring any type of corpora.

Visualization Technique The first step in creating the graph is topical modelling of documents.The distribution of topics are calculated, if the probability of topic occurring in document exceeds a user defined threshold the document and topic is connected by an edge. The thickness of the edge represents the probability calculated earlier. The threshold is modified using a slider available in the interface.

The color of the document nodes is determined either by a set of colors defined by user or a color range generated with the help of metadata.

Two layout techniques are used in TopicNets [7] one is based on topic similarity and other preserves the structural aspect of nodes. The topic similarity layout is computed using symmetric Kullback-Leibler divergence between every pair of word topic distributions. The resulting dissimilarity matrix is then used as an input to a Multi-Dimensional Scaling algorithm this algorithm determined the position of each topic node. After fixing the positions of topic nodes a standard force-directed layout algorithm is applied to place the document nodes in this topic space.

The other layout has the ability to preserve the order of document. This layout is especially useful if the user wants to extract sequential information from documents.For this purpose the documents are places around the circumference of a circle.In this layout, document nodes are first fixed in place, and a force-directed layout is applied to connected topic nodes.

Strengths and Weaknesses TopicNets visualization is build in real time, during the rendering on web interface, thus it is a fast model. It is able to provide corpus specific as well as document specific views [7].

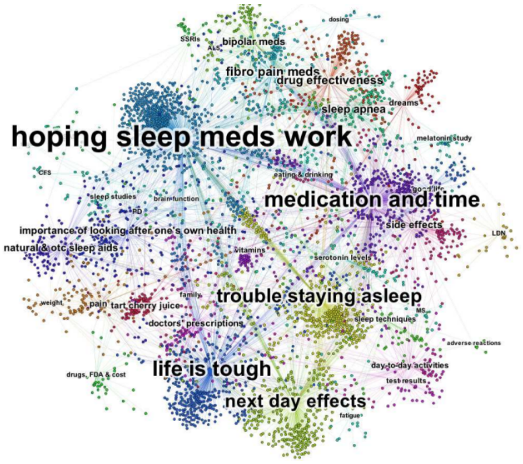

Figure 5: Full Picture of Gephi-Based System [8]

Main Idea Online platforms for health-related issues provide users a platform to connect with each other and seek ways to improve their health. However these platforms can be overwhelming for users due to large amount of information available. In order to solve this issue [8] has developed a platform that utilizes topic modelling and graph visualization to effectively aggregate the large literature and present a visualization understandable by user.

Application Domain This visualization [8] can be used by patients for easily skimming through the vast information available on internet about health related issues.

Visualization Technique A two-mode network is used to map the relationships between posts and topics.The nodes of the graph are based on topics and posts. An edge list contains post ID, topics and the weight of the edge is the proportion of posts that contain the topic. The node and edge lists fed to Gephi that creates a topic-post bipartite graph. The size of topic node is based on the number of posts to which each is connected, and the text in proportion to node size. To help the user in distinguishing topics community detection algorithm is employed to cluster nodes. The color of node represents the membership to cluster.

Strengths and Weaknesses The approach [8] uses a bipartite graph which effectively encodes the relationship between entities. The layout has not been properly defined due to automatic generation by Gephi.

3 Conclusion

In the past decade, text mining and its related tasks have become a focus of many studies. Topic modelling which is a task of text mining has been used in various domains to extract topics from textual document as well as corpora. Among different representations of text graph-based approach has also been employed in various researches. This survey provides an overview of domains of application as well as different visualization techniques employed to generate effective visualizations of graph-based topics.

References

[1] L. Stappen, J. Thies, G. Hagerer, B. W. Schuller, and G. Groh, "Graphtmt: Unsupervised graph-based topic modeling from video transcripts," 2021.

[2] H. Sayyadi and L. Raschid, "A graph analytical approach for topic detection," *ACM Trans. Internet Technol.*, vol. 13, no. 2, dec 2013. [Online]. Available: https://doi.org/10.1145/2542214.2542215

[3] K. Ghoorchian and M. Sahlgren, "Gdtm: Graph-based dynamic topic models," *Progress in Artificial Intelligence*, vol. 9, 05 2020.

[4] W. Cui and H. Qu, "A survey on graph visualization," *Clear Water Bay, Kowloon, Hong Kong*, vol. 145, 2007.

[5] X. Wang, S. Liu, J. Liu, J. Chen, J. Zhu, and B. Guo, "Topicpanorama: A full picture of relevant topics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 12, pp. 2508–2521, 2016.

[6] A. Smith, J. Chuang, Y. Hu, J. Boyd-Graber, and L. Findlater, "Concurrent visualization of relationships between words and topics in topic models," in *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 2014, pp. 79–82.

[7] B. Gretarsson, J. O'donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth, "Topicnets: Visual analysis of large text corpora with topic modeling," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 2, pp. 1–26, 2012.

[8] A. T. Chen, L. Sheble, and G. Eichler, "Topic modeling and network visualization to explore patient experiences," in *Visual Analytics in Healthcare Workshop 2013*, 2013.