GraphRep: Boosting Text Mining, NLP and Information Retrieval with Graphs

Michalis Vazirgiannis
École Polytechnique, France
AUEB, Greece
mvazirg@lix.polytechnique.fr

Fragkiskos D. Malliaros
CentraleSupélec and Inria Saclay
France
fragkiskos.me@gmail.com

Giannis Nikolentzos
École Polytechnique
France
nikolentzos@lix.polytechnique.fr

ABSTRACT

Graphs have been widely used as modeling tools in Natural Language Processing (NLP), Text Mining (TM) and Information Retrieval (IR). Traditionally, the unigram bag-of-words representation is applied; that way, a document is represented as a multiset of its terms, disregarding dependencies between the terms. Although several variants and extensions of this modeling approach have been proposed, the main weakness comes from the underlying term independence assumption; the order of the terms within a document is completely disregarded and any relationship between terms is not taken into account in the final task. To deal with this problem, the research community has explored various representations, and to this direction, graphs constitute a well-developed model for text representation. The goal of this tutorial is to offer a comprehensive presentation of recent methods that rely on graph-based text representations to deal with various tasks in Text Mining, NLP and IR.

CCS CONCEPTS

• **Information systems → Data mining**; **Information retrieval**; *Retrieval models and ranking*;

KEYWORDS

Graph Mining, Natural Language Processing, Information Retrieval

ACM Reference Format:
Michalis Vazirgiannis, Fragkiskos D. Malliaros, and Giannis Nikolentzos. 2018. GraphRep: Boosting Text Mining, NLP and Information Retrieval with Graphs. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3269206.3274273

1 INTRODUCTION

Traditionally, in text analytics tasks, the unigram *bag-of-words* representation is applied [1]; a document is represented as a multiset of its terms, disregarding dependencies between the terms. Although several variants and extensions of this model have been proposed (e.g., the *n*-gram model), the main weakness comes from the underlying term independence assumption. The order of the terms within a document is completely ignored and any relationship between terms is not taken into account in the final task (e.g., text categorization).

Nevertheless, as the heterogeneity of text collections is increasing (especially with respect to document length and vocabulary), the research community started exploring different document representations aiming to capture more fine-grained contexts of co-occurrence between different terms, challenging the well-established unigram bag-of-words model. To this direction, graphs constitute a well-developed model that has been adopted for text representation. More precisely, a graph $G = (V, E)$ consists of a set of vertices V and a set of edges E that connect different vertices. Due to the strong modeling capabilities of graphs, vertices and edges can capture a plethora of linguistic units [14]:

- The **vertices** can correspond to *paragraphs*, *sentences*, *phrases*, *words* and *syllables*.
- The **edges** of the graph can capture various types of relationships between two vertices, including *co-occurrence* within a window over the text, *syntactic* relationship as well as *semantic* relationship.

Depending on the task and the granularity level that we are interested in, the graph itself can represent different entities, such as a sentence, a single document, multiple documents or even the entire document collection. Furthermore, the edges on the graphs can be *directed* or *undirected*, as well as associated with *weights* or not. For example, in the case where the vertices correspond to terms of the text and the edges capture co-occurrence relations, a directed graph is able to preserve actual flow on the text, while in the case of undirected one, an edge captures co-occurrence of two terms whatever the respective order between them is. If we are also interested to take into account of the number of co-occurrences of two terms in the document, we can consider weighted edges, where the weight of each edge will be equal to the number of co-occurrences.

In addition to the rich modeling capabilities of graphs, the scientific fields of graph theory and graph mining, has to offer plenty of sophisticated algorithms that can have direct applications in NLP, IR and Web Mining in general. That way, a plethora of text analytics and NLP tasks (e.g., web search, text categorization and keyword extraction), have been addressed combining a graph-based representation of text with graph mining algorithms.

1.1 Scope of the Tutorial

The goal of this tutorial is to offer a comprehensive presentation of recent methods that rely on graph-based text representations to deal with various tasks in Web mining, NLP and IR. We will describe basic as well as novel graph theoretic concepts and we will

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3274273

examine how they can be applied in a wide range of text-related application domains.

2 OUTLINE OF THE TUTORIAL

1. Introduction
– Basics on IR and NLP [1]
– Highlights on graph-based document representations
– Overview of the topics that will be covered in the tutorial
– What the tutorial is not about

2. Graph-theoretic Concepts
– Basic graph definitions
– Node centrality criteria (e.g., closeness, betweenness) and community structure
– PageRank and HITS
– Graph degeneracy (k-core and K-truss decompositions)
– Frequent subgraph mining
– Basics on graph kernels

3. Graph-based Text Representations
– How to construct a graph from a single document or a collection of documents
– Graph-of-words concept
– Semantics of nodes and edges
– Edge directionality and edge weight
– Graph construction trade-offs

4. Information Retrieval [2, 16]
– Graph-based term weighting in IR
– TW and TW-IDF weighting functions

5. Keyword - Keyphrase Extraction and Text Summarization [3–5, 8, 12, 17, 21, 22]
– Clustering-based methods – TextRank and PageRank-based approaches for single topic keyword extraction
– HITS algorithm for keyword extraction
– Node centrality criteria for keyword and keyphrase extraction
– Graph degeneracy-based methods
– Combining graph degeneracy and submodularity for unsupervised extractive summarization – Keyphrase annotation
– Software demonstration

6. Novelty and Event Detection in Text Streams [10, 11]
– Degeneracy-based sub-event detection in Twitter streams
– A graph optimization approach for sub-event detection and summarization in Twitter

7. Text Categorization (TC) [6, 7, 9, 13, 15, 18–20]
– Graph-based term weighting for TC
– Frequent subgraphs as categorization features
– Term graph models for TC
– Graph matching approaches
– Graph-based regularization for TC
– Graph representation learning methods

8. Open Problems and Future Research
– Graph kernels and network embeddings for document similarity and categorization
– Dense subgraphs for keyword selection
– Multi-topic keyword extraction

3 LEARNING OBJECTIVES

The learning outcomes of the tutorial consist in:

- A thorough presentation of novel graph-theoretic concepts and algorithms (including graph degeneracy, graph kernels and network representation learning methods), and their applications in text analytics.
- Demonstration of recent approaches of graph-based text representations in various web mining-related research areas, including information retrieval, text summarization, text categorization and their applications.

REFERENCES

[1] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
[2] Roi Blanco and Christina Lioma. 2012. Graph-based Term Weighting for Information Retrieval. *Inf. Retr.* 15, 1 (2012), 54–92.
[3] Florian Boudin. 2013. A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction (IJCNLP '13). In *Sixth International Joint Conference on Natural Language Processing*. 834–838.
[4] Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2013. TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction. In *IJCNLP*. 543–551.
[5] Adrien Bougouin, Florian Boudin, and Béatrice Daille. 2016. Keyphrase Annotation with Graph Co-Ranking. In *COLING*.
[6] Samer Hassan, Rada Mihalcea, and Carmen Banea. 2007. Random-Walk Term Weighting for Improved Text Classification.. In *ICSC*. 242–249.
[7] Chuntao Jiang, Frans Coenen, Robert Sanderson, and Michele Zito. 2010. Text classification using graph mining-based feature extraction. *Knowl.-Based Syst.* 23, 4 (2010), 302–308.
[8] Marina Litvak and Mark Last. 2008. Graph-based Keyword Extraction for Single-document Summarization. In *MMIES*. Association for Computational Linguistics, 17–24.
[9] Fragkiskos D. Malliaros and Konstantinos Skianis. 2015. Graph-Based Term Weighting for Text Categorization. In *ASONAM*. ACM, 1473–1479.
[10] Polykarpos Meladianos, Giannis Nikolentzos, François Rousseau, Yannis Stavrakas, and Michalis Vazirgiannis. 2015. Degeneracy-Based Real-Time Sub-Event Detection in Twitter Stream. In *ICWSM*.
[11] Polykarpos Meladianos, Christos Xypolopoulos, Giannis Nikolentzos, and Michalis Vazirgiannis. 2018. An Optimization Approach for Sub-event Detection and Summarization in Twitter. In *ECIR*. 481–493.
[12] Rada Mihalcea and Paul Tarau. 2004. TextRank: bringing order into texts. In *EMNLP*. Association for Computational Linguistics.
[13] Giannis Nikolentzos, Polykarpos Meladianos, François Rousseau, Yannis Stavrakas, and Michalis Vazirgiannis. 2017. Shortest-Path Graph Kernels for Document Similarity. In *EMNLP*. 1891–1901.
[14] Francois Rousseau. 2015. *Graph-of-words: mining and retrieving text with networks of features.* Ph.D. Dissertation. École Polytechnique.
[15] François Rousseau, Emmanouil Kiagias, and Michalis Vazirgiannis. 2015. Text Categorization as a Graph Classification Problem. In *ACL (1)*. 1702–1712.
[16] François Rousseau and Michalis Vazirgiannis. 2013. Graph-of-word and TW-IDF: new approach to ad hoc IR. In *CIKM*. ACM, 59–68.
[17] François Rousseau and Michalis Vazirgiannis. 2015. Main core retention on graph-of-words for single-document keyword extraction. In *ECIR*. Springer, 382–393.
[18] Konstantinos Skianis, Fragkiskos D. Malliaros, and Michalis Vazirgiannis. 2018. Fusing Document, Collection and Label Graph-based Representations with Word Embeddings for Text Classification. In *TextGraphs*. 49–58.
[19] Konstantinos Skianis, François Rousseau, and Michalis Vazirgiannis. 2016. Regularizing Text Categorization with Clusters of Words. In *EMNLP*. The Association for Computational Linguistics, 1827–1837.
[20] Jian Tang, Meng Qu, and Qiaozhu Mei. 2015. PTE: Predictive Text Embedding Through Large-scale Heterogeneous Text Networks. In *KDD*. 1165–1174.
[21] Antoine J.-P. Tixier, Fragkiskos D. Malliaros, and Michalis Vazirgiannis. 2016. A Graph Degeneracy-based Approach to Keyword Extraction. In *EMNLP*. The Association for Computational Linguistics, 1860–1870.
[22] Rui Wang, Wei Liu, and Chris McDonald. 2015. Corpus-independent Generic Keyphrase Extraction Using Word Embedding Vectors. In *Workshop on Deep Learning for Web Search and Data Mining (DL-WSDM '15).*