Graph-of-word: Boosting Text Mining with Graphs

Michalis Vazirgiannis
Laboratory of Informatics, Ecole Polytechnique, France
mvazirg@lix.polytechnique.fr

ABSTRACT
The Bag-of-words model has been the dominant approach for IR and Text mining for many years assuming the word independence and the frequencies as the main feature for feature selection and for query to document similarity. Although the long and successful usage, bag- of-words ignores words' order and distance within the document – weakening thus the expressive power of the distance metrics. We propose graph-of-word, an alternative approach that capitalizes on a graph representation of documents and challenges the word independence assumption by taking into account words' order and distance. We applied graph-of-word in various tasks such as ad-hoc Information Retrieval, Single-Document Keyword Extraction, Text Categorization and Sub-event Detection in Textual Streams. In all cases the the graph of word approach, assisted by degeneracy at times, outperforms the state of the art base lines in all cases.

Author Keywords
Graph Mining, Text Mining, Graph Degeneracy

BIOGRAPHY
Dr. Vazirgiannis is a Professor at LIX, Ecole Polytechnique in France. He has conducted research in GMD-IPSI, Max Planck MPI (Germany), in INRIA/FUTURS (Paris). He has been a teaching at AUEB (Greece), Ecole Polytechnique, Telecom-Paristech, ENS (France), Tsinghua (China) and in Deusto University (Spain). His current research interests are on machine learning and combinatorial methods for Graph analysis (including community detection, graph clustering and embeddings, influence maximization), Text mining including Graph of Words, word embeddings with applications to web advertising and marketing, event detection and summarization. He has active cooperation with industrial partners in the area of data analytics and machine learning for large scale data repositories in different application domains. He has supervised fourteen completed PhD theses. He has published three books and more than a 140 papers in international refereed journals and conferences. He has organized large scale conferences in the area of Data Mining and Machine Learning (such as ECML/PKDD) while he participates in the senior PC of AI and ML conferences – such as AAAI and IJCAI, He has received the ERCIM and the Marie Curie EU fellowships and since 2015 he leads the AXA Data Science chair.

REFERENCES
1. Christos Giatsidis, Dimitrios M. Thilikos, Michalis Vazirgiannis: D-cores: Measuring Collaboration of Directed Graphs Based on Degeneracy. IEEE - ICDM 2011: 201-210
2. Matching Node Embeddings for Graph Similarity, Giannis Nikolentzos, Polykarpos Meladianos and Michalis Vazirgiannis, **AAAI2017**
3. Konstantinos Skianis, François Rousseau, Michalis Vazirgiannis: Regularizing Text Categorization with Clusters of Words. EMNLP 2016: 1827-1837
4. A. J.-P. Tixier, Fragkiskos D. Malliaros, Michalis Vazirgiannis: A Graph Degeneracy-based Approach to Keyword Extraction. EMNLP 2016: 1860-1870
5. GoWvis: a web application for Graph-of-Words-based text visualization and summarization, AJP Tixier, K Skianis, M Vazirgiannis ACL 2016, 151 2015
6. J. Kim, F.Rousseau and M.Vazirgiannis, Convolutional Sentence Kernel with Word Embedding for Short Text Categorization, proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP '15).
7. F. Rousseau, E. Kiagias and Michalis Vazirgiannis, Text Categorization as a Graph Classification Problem, Proceedings of the 52th Annual Meeting of the Association for Computational Linguistics and the 6th International Joint Conference on Natural Language Processing (ACL-IJCNLP '15).
8. F. Rousseau and Michalis Vazirgiannis. K-core on Graph-of-words for Single-Document Keyword Extraction, European Conference on Information Retrieval, Vienna, Austria, 2015
9. P. Meladianos, G. Nikolentzos, F. Rousseau, Y. Stavrakas, M. Vazirgiannis, "Degeneracy-Based Real-Time Sub-Event Detection in Twitter Stream", in the proceedings of the AAAI-ICWSM 2015 conference

WWW 2017 Companion, April 3-7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3055362