Accepted Manuscript

A novel parallel distance metric-based approach for diversified ranking on large graphs

Jin Li, Yun Yang, Xiaoling Wang, Zhiming Zhao, Tong Li

PII: S0167-739X(18)30176-6
DOI: https://doi.org/10.1016/j.future.2018.05.031
Reference: FUTURE 4206

To appear in: *Future Generation Computer Systems*

Received date : 26 January 2018
Revised date : 24 April 2018
Accepted date : 14 May 2018

Please cite this article as: J. Li, Y. Yang, X. Wang, Z. Zhao, T. Li, A novel parallel distance metric-based approach for diversified ranking on large graphs, *Future Generation Computer Systems* (2018), https://doi.org/10.1016/j.future.2018.05.031

A Novel Parallel Distance Metric-based Approach for Diversified Ranking on Large Graphs

Jin Li[a,b,c], Yun Yang[a,b,c,*], Xiaoling Wang[d], Zhiming Zhao[e], Tong Li[b,c]

[a] School of Software, Yunnan University, Kunming, China
[b] Key Laboratory of Software Engineering of Yunnan Province, Kunming, China
[c] Kunming Key Laboratory of Data Science and Intelligent Computing, Kunming, China
[d] Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai, China
[e] University of Amsterdam, Amsterdam, Netherlands

Abstract

Diversified ranking on graphs (DRG) is an important and challenging issue in researching graph data mining. Traditionally, this problem is modeled by a submodular optimization objective, and solved by applying a cardinality constrained monotone submodular maximization. However, the existing submodular objectives do not directly capture the dis-similarity over pairs of nodes, while most of algorithms cannot easily take full advantage of the power of a distributed cluster computing platform, such as Spark, to significantly promote the efficiency of algorithms. To overcome the deficiencies of existing approaches, in this paper, a generalized distance metric based on a subadditive set function over the symmetry difference of neighbors of pairs of nodes is introduced to capture the pairwise dis-similarity over pairs of nodes. In our approach, DRG is formulated as a Max-Sum *k*-dispersion problem with metrical edge weights, which is NP-hard, in association with the proposed distance metric, a centralized linear time 2-approximation algorithm GA is then developed to significantly solve the problem of DRG. Moreover, we develop a highly parallelizable algorithm for DRG, which can be easily implemented in MapReduce style parallel computation models using GA as a basic reducer. Finally, extensive experiments are conducted on real network datasets to verify the effectiveness and efficiency of our proposed approaches.

Keywords: graph algorithms, diversified ranking, distance metric, parallel computing, MapReduce