Information Sciences 329 (2016) 619–637

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

A multi-class approach for ranking graph nodes: Models and experiments with incomplete data

Gianna M. Del Corso*, Francesco Romani

Dipartimento di Informatica, Università di Pisa, Largo Pontecorvo, 3, Pisa 56127, Italy

ARTICLE INFO

Article history:
Received 13 February 2014
Revised 16 September 2015
Accepted 26 September 2015
Available online 9 October 2015

Keywords:
Link analysis
Models for ranking graph nodes
Missing links
Web matrix reducibility and permutation

ABSTRACT

After the phenomenal success of the PageRank algorithm, many researchers have extended the PageRank approach to ranking graphs with richer structures in addition to the simple linkage structure. Indeed, in some scenarios we have to deal with networks modeling multi-parameters data where each node has additional features and there are important relationships between such features.

This paper addresses the need of a systematic approach to deal with multi-parameter data. We propose models and ranking algorithms that can be applied to a large variety of networks (bibliographic data, patent data, twitter and social data, healthcare data). We focus on several aspects not previously addressed in the literature: (1) we propose different models for ranking multi-parameters data and a class of numerical algorithms for efficiently computing the ranking score of such models, (2) we analyze stability and convergence of the proposed numerical schemes and we derive a fast and stable ranking algorithm, (3) we analyze the robustness of our models when data are incomplete. The comparison of the rank on the incomplete data with the rank on the full structure shows that our models compute consistent rankings whose correlation is up to 60% when just 10% of the links of the attributes are maintained.

1. Introduction

Ranking algorithms are essential tools for searching in large collections of data and they are becoming more and more important as the amount of available data gets bigger and richer. Following the introduction and the success of PageRank and other ranking algorithms [8,17], researchers have extended similar techniques to a multitude of domains [4,5,10,13,18].

With the advent of the semantic web, data containing many types of features and relationships are becoming common. Algorithms taking advantage of such additional information are needed and sought. Many analytical techniques have been proposed to better understand these data and their properties. Particularly important are ranking algorithms that evaluate objects on the basis of ranking functions measuring some characteristics of the objects. With such functions any two objects of the same type can be sorted in a partial order and compared either qualitatively or quantitatively. Ranking algorithms are widely applied in different network settings to get an overall view of the data.

In this paper we consider the setting in which the data consist of a collection of linked items, where each item has a set of additional attributes (features). In this setting it is natural to assume that the ranking of items with common attributes are mutually influenced. Many important problems are instances of this general framework.

* Corresponding author. Tel.: +39 050 2213118.
E-mail addresses: gianna.delcorso@unipi.it (G.M. Del Corso), romani@di.unipi.it (F. Romani).

http://dx.doi.org/10.1016/j.ins.2015.09.046
0020-0255/© 2015 Elsevier Inc. All rights reserved.

In *bibliographic ranking* items are scholarly papers with the linkage structure provided by their citations. To each paper it is natural to associate features such that: authors, publication venue, subject classification and so on. The ranking of scientific papers on the basis of the received citations and publication venue has become an increasingly popular topic since the late 80's due to its importance for recruiting, promotions, and funding.

In *patent data analysis* items are patents linked by the citations to older patents. To each patent we can associate inventors, firm, examiner, technologies, *etc*. Studies in marketing science utilize patents to examine different aspects of innovation: to understand knowledge flow within and across firms, to describe how knowledge flow influences the success of innovation, and to identify antecedents and outcomes of product innovation. An example of this line of research is [30] where the author shows that the number of patents owned by a firm (its patent count) correlates with R&D expenditure and represent a specific type of resource (intellectual property) the firm can use in various market processes.

Other examples of multi-parameter networks are: *social or twitter graphs*, where we have information about status, geographical location, education, *etc.* of users, and *healthcare data networks* where we have information on patients, doctors, treatments, diseases, *etc*.

With a little abuse of notation in the following we use the term "multigraph" to denote this kind of relationships between items and features, while other authors identify this kind of graph with as heterogeneous information networks [28]. As shown by the above examples, the multigraph framework encompasses many different applications in which one has to compare different entities on the basis of their attributes and relationships. For this reason our results should be of interest for researchers in the information retrieval community as well as economists and people interested in the analysis of social networks.

In this paper we describe different models for representing the multigraph structure of a network. We analyze different techniques for assigning weights to features and to use these weights in the ranking process. These weights capture the importance that each link confers to the linked object. We then build a fast and stable numerical method for computing the ranking score according to our models. The proposed algorithm is obtained by combining two non-stationary methods (BCGStab [23] and TFQMR [23]) and a final phase of iterative refinement.

We perform many tests on two large datasets of patent data extracted from the US patent office: the first dataset consists of all the patents granted in the period 1976–1990 (roughly 2.5 million patents), and the second of those issued between 1976 and 2012 (almost eight million patents). The experiments aim at understanding the differences between the various models and the role of the parameters involved in the algorithm. We also compare the results with those returned by PageRank and the ordering induced by the simple citation count.

We briefly investigate also the robustness of our models when data are incomplete and unrecoverable. In this setting our goal is to use all the information available without advantaging players (items or features) with more complete data respect to those where some information is missing. We treat unknown values as zeroes, since often we cannot distinguish between missing (not available) or absent (not existent) features. This is the only viable choice when the missing data are unrecoverable and is the strategy implemented in patent repositories and in citation databases such as *Scopus, Mathscinet* where, for example, a citation is not attributed to anyone when the name of an author has been misspelled.

Following an established approach [16,36], we evaluate the robustness of our ranking schemas on incomplete data by randomly removing features from items with an assigned probability. Our experiments show that, even removing up to half of the features, the ranks provided by our algorithm highly correlate to the ranks computed on the complete data. As expected, as more and more features are removed, the ranks converge to the rank obtained using only the linkage structure.

Finally, we tested the robustness of our models with respect to the granularity of the features. For example if we are dealing with bibliographic data we can group papers into subject classes where the granularity can be subject macro areas (math, computer science, *etc.*) or finer classifications (algebra, number theory, calculus, algorithms, data bases, *etc*). In this context it is desirable that, when using a finer classification, the sum of the ranks of topic A subtopics is close to A's rank computed using the coarser classification. Experiments with the US patent dataset show that most of our models have such desirable feature.

The paper is organized as follows. In Section 1.1 we formally introduce the problem considered in the paper. In Section 1.2 we motivate our study and connect the techniques and the algorithm we propose with the existing literature. In Section 2 we briefly present some models discussing how extra information and features can be added to the citation structure to improve ranking and possible weighting criteria for such features. In our models the ranking is obtained approximating the Perron vector of a suitable stochastic matrix.

In Section 3 we discuss different ways for approximating the Perron vector showing that it can be obtained computing the solution of a linear system. In Section 4 we discuss different methods for the numerical solution of such linear system and we describe the databases used for the experiments. In Section 5 we report an extensive numerical testing to compare the different models in terms of convergence for missing data and consistence for class aggregation. Section 6 contains the conclusion and some discussions about possible improvements of the models.

1.1. Preliminaries and notations on multigraphs

In this paper we consider a multigraph as described by a directed graph $G = (V, E)$ and two mapping functions, one for the nodes $\tau : V \to \mathcal{A}$ and one for the edges $\phi : E \to \mathcal{R}$. Each node $v \in V$ belongs to a particular type $\tau(v) \in \mathcal{A}$ and each edge $e \in E$ belongs to a particular type of relation $\phi(e) \in \mathcal{R}$. Functions ϕ and τ are such that if e_1 and e_2 are two edges, $e_1 = (v_1, v_2)$ and $e_2 = (w_1, w_2)$, with $\phi(e_1) = \phi(e_2)$, then $\tau(v_1) = \tau(w_1)$ and $\tau(v_2) = \tau(w_2)$. When $|\mathcal{A}| > 1$ and $|\mathcal{R}| > 1$ we say that the graph is a multigraph.