Applied Numerical Mathematics 75 (2014) 77–99

Contents lists available at ScienceDirect

Applied Numerical Mathematics

www.elsevier.com/locate/apnum

Perron vector optimization applied to search engines

Olivier Fercoq [1]

The University of Edinburgh, School of Mathematics, United Kingdom

ARTICLE INFO

Article history:
Available online 1 February 2013

Keywords:
Optimization
Perron–Frobenius theory
Search engines
Scalable algorithms

ABSTRACT

In the last years, Google's PageRank optimization problems have been extensively studied. In that case, the ranking is given by the invariant measure of a stochastic matrix. In this paper, we consider the more general situation in which the ranking is determined by the Perron eigenvector of a nonnegative, but not necessarily stochastic, matrix, in order to cover Kleinberg's HITS algorithm. We also give some results for Tomlin's HOTS algorithm. The problem consists then in finding an optimal outlink strategy subject to design constraints and for a given search engine.

We study the relaxed versions of these problems, which means that we should accept weighted hyperlinks. We provide an efficient algorithm for the computation of the matrix of partial derivatives of the criterion, that uses the low rank property of this matrix. We give a scalable algorithm that couples gradient and power iterations and gives a local minimum of the Perron vector optimization problem. We prove convergence by considering it as an approximate gradient method.

We then show that optimal linkage strategies of HITS and HOTS optimization problems satisfy a threshold property. We report numerical results on fragments of the real web graph for these search engine optimization problems.

1. Introduction

1.1. Motivation

Internet search engines use a variety of algorithms to sort web pages based on their text content or on the hyperlink structure of the web. In this paper, we focus on algorithms that use the latter hyperlink structure, called link-based algorithms. The basic notion for all these algorithms is the web graph, which is a digraph with a node for each web page and an arc between pages i and j if there is a hyperlink from page i to page j. Famous link-based algorithms are PageRank [12], HITS [30], SALSA [33] and HOTS [56]. See also [31,32] for surveys on these algorithms. The main problem of this paper is the optimization of the ranking of a given web site. It consists in finding an optimal outlink strategy maximizing a given ranking subject to design constraints.

One of the main ranking methods relies on the PageRank introduced by Brin and Page [12]. It is defined as the invariant measure of a walk made by a random surfer on the web graph. When reading a given page, the surfer either selects a link from the current page (with a uniform probability), and moves to the page pointed by that link, or interrupts his current search, and then moves to an arbitrary page, which is selected according to given "zapping" probabilities. The rank of a page is defined as its frequency of visit by the random surfer. It is interpreted as the "popularity" of the page. The

E-mail address: olivier.fercoq@ed.ac.uk.

[1] The present work was completed as O. Fercoq was Ph.D. student at INRIA Saclay – Île-de-France and CMAP Ecole Polytechnique. It was supported by Orange Labs through the research contract CRE 3795 with INRIA.

0168-9274/$36.00 © 2013 IMACS. Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.apnum.2012.12.006

PageRank optimization problem has been studied in several works: [1,36,15,26,13,20]. The last two papers showed that PageRank optimization problems have a Markov decision process structure and both papers provided efficient algorithm that converge to a global optimum. Csáji, Jungers and Blondel in [13] showed that optimizing the PageRank score of a single web page is a polynomial problem. Fercoq, Akian, Bouhtou and Gaubert in [20] gave an alternative Markov decision process model and an efficient algorithm for the PageRank optimization problem with linear utility functions and more general design constraints, showing in particular that any linear function of the PageRank vector can be optimized in polynomial time.

In this paper, we consider the more general situation in which the ranking is determined by the Perron eigenvector of a nonnegative, but not necessarily stochastic, matrix. The Perron–Frobenius theorem (see [4] for instance) states that any nonnegative matrix A has a nonnegative principal eigenvalue called the Perron root and nonnegative principal eigenvectors. If, in addition, A is irreducible, then the Perron root is simple and the (unique up to a multiplicative constant) nonnegative eigenvector, called the Perron vector, has only positive entries. This property makes it a good candidate to sort web pages. The ranking algorithms considered differ in the way of constructing from the web graph a nonnegative irreducible matrix from which we determine the Perron vector. Then, the bigger is the Perron vector's coordinate corresponding to a web page, the higher this web page is in the ranking. In [28], such a ranking is proposed for football teams. The paper [52] uses the Perron vector to rank teachers from pairwise comparisons. See also [57] for a survey on the subject. When it comes to web page ranking, the PageRank is the Perron eigenvector of the transition matrix described above but HITS [30], SALSA [33], CenterRank [9], EntropyRank [16] and other ranking algorithms also rank pages according to a Perron vector.

The HITS algorithm [30] is not purely a link-based algorithm. It is composed of two steps and the output depends on the query of the user. Given a query, we first select a seed of pages that are relevant to the query according to their text content. This seed is then extended with pages linking to them, pages to which they link and all the hyperlinks between the pages selected. We thus obtain a subgraph of the web graph focused on the query. Then, the second step assigns each page two scores: a hub score v and an authority score u such that good hubs should point to good authorities and good authorities should be pointed to by good hubs. Introducing the adjacency matrix A of the focused graph, this can be written as $v = \rho A u$ and $u = \rho A^T v$ with $\rho \in \mathbb{R}_+$, which means that the vector of hub scores is the Perron eigenvector of the matrix $A^T A$ and that the vector of authority scores is the Perron eigenvector of AA^T. The construction of HITS' focused subgraph is a combination of text content relevancy with the query and of hyperlink considerations. Maximizing the probability of appearance of a web page on this subgraph is thus a composite problem out of the range of this paper. We shall however study the optimization of HITS authority, for a given focused subgraph.

We also studied the optimization of Tomlin's HOTS scores [56]. In this case, the ranking is the vector of dual variables of an optimal flow problem. The flow represents an optimal distribution of web surfers on the web graph in the sense of entropy minimization. The dual variable, one by page, is interpreted as the "temperature" of the page, the hotter a page the better. Tomlin showed that this vector is the solution of a nonlinear fixed point equation: it may be seen as a nonlinear eigenvector. Indeed, we show that most of the arguments available in the case of Perron vector optimization can be adapted to HOTS optimization. We think that this supports Tomlin's remark that "malicious manipulation of the dual values of a large scale nonlinear network optimization model [...] would be an interesting topic".

1.2. Contribution

In this paper, we study the problem of optimizing the Perron eigenvector of a controlled matrix and apply it to PageRank, HITS and HOTS optimization. Our first main result is the development of a scalable algorithm for the local optimization of a scalar function of the Perron eigenvector over a set of nonnegative irreducible matrices. Indeed, the global Perron vector optimization over a convex set of nonnegative matrices is NP-hard, so we focus on the searching of local optima. We give in Theorem 1 a power-type algorithm for the computation of the matrix of the partial derivatives of the objective, based on the fact that it is a rank 1 matrix. This theorem shows that computing the partial derivatives of the objective has the same order of complexity as computing the Perron vector by the power method, which is the usual method when dealing with the large and sparse matrices built from the web graph. Then we give an optimization algorithm that couples power and gradient iterations (Algorithms 2 and 3). Each step of the optimization algorithm involves a suitable number of power iterations and a descent step. By considering this algorithm to be an approximate projected gradient algorithm [49,48], we prove that the algorithm converges to a stationary point (Theorem 2). Compared with the case when the number of power iterations is not adapted dynamically, we got a speedup between 3 and 20 in our numerical experiments (Section 7) together with a more precise convergence result.

Our second main result is the application of Perron vector optimization to the optimization of scalar functions of HITS authority or HOTS scores. We derive optimization algorithms and, thanks to the low rank of the matrix of partial derivatives, we show that the optimal linkage strategies of both problems satisfy a threshold property (Propositions 9 and 12). This property was already proved for PageRank optimization in [15,20]. As in [26,13,20] we partition the set of potential links (i, j) into three subsets, consisting respectively of the set of *obligatory links*, the set of *prohibited links* and the set of *facultative links*. When choosing a subset of the facultative links, we get a graph from which we get any of the three ranking vectors. We are then looking for the subset of facultative links that maximizes a given utility function. We also study the associated relaxed problems, where we accept weighted adjacency matrices. This assumes that the webmaster can influence the importance of the hyperlinks of the pages she controls, for instance by choosing the size of the font, the color or the