Theoretical Computer Science 518 (2014) 96–116

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

On the approximability of the link building problem

Martin Olsen [a], Anastasios Viglas [b,*]

[a] AU Herning, Aarhus University, Birk Centerpark 15, DK-7400 Herning, Denmark
[b] School of Information Technologies, The University of Sydney, 1 Cleveland St, NSW 2006, Australia

ARTICLE INFO

Article history:
Received 30 April 2012
Received in revised form 8 March 2013
Accepted 6 August 2013
Communicated by P. Widmayer

Keywords:
Link building
PageRank
Optimization

ABSTRACT

We consider the LINK BUILDING problem, which involves maximizing the PageRank value of a given target vertex in a directed graph by adding k new links that point to the target (backlinks). We present a theorem describing how the topology of the graph affects the choice of potential new backlinks. Based on this theorem we show that no fully polynomial-time approximation scheme (FPTAS) exists for LINK BUILDING unless P = NP and we also show that LINK BUILDING is $W[1]$-hard.
Furthermore, we show that this problem is in the class APX by presenting the polynomial time algorithm r-Greedy, which selects new backlinks in a greedy fashion and results in a new PageRank value for the target vertex that is within a constant factor from the best possible. We also consider the naive algorithm π-Naive, where we choose backlinks from vertices with high PageRank values compared to the out-degree and show that this algorithm performs much worse on certain graphs compared to our constant factor approximation. Finally, we provide a lower bound for the approximation ratio of our r-Greedy algorithm.

1. Introduction

Ranking problems are found in several areas, including classification, user behavior analysis, social networks, as well as advertising. Ranking on graphs and social networks has a history in social sciences. The problem of determining popularity within a network of friends for example, can be recursively defined; "popular people are friends with other popular people". This intuitive formulation led to many variants of ranking algorithms, including Kleinberg's HITS algorithm, and the PageRank algorithm, for determining popularity (or relative importance) of webpages by considering only the network of the web. The links between websites are considered directed links in the web graph.

Search engine optimization (SEO) is a fast growing industry that deals with optimizing the ranking of webpages in search engine results. SEO is a complex task, especially since the specific details of search and ranking algorithms are often not publicly released, and also can change frequently. One of the key elements of optimizing for search engine visibility is the "external link popularity", which is based on the structure of the web graph. The problem of obtaining optimal new backlinks in order to achieve good search engine rankings is known as Link Building and Link Building is widely considered to be an important aspect of SEO [3,4].

The PageRank algorithm is one of the most well-known methods of defining a ranking among vertices according to the link structure of a graph. The definition of PageRank [5] is based on the "random surfer" walk, which is defined as follows: the walk starts at a random vertex in the graph, given by a starting probability distribution (usually taken to be uniform). The surfer then proceeds to choose a link to continue the walk, uniformly at random from the out-links available at the

* Corresponding author.
E-mail addresses: martino@hih.au.dk (M. Olsen), taso.viglas@sydney.edu.au (A. Viglas).

0304-3975/$ – see front matter
http://dx.doi.org/10.1016/j.tcs.2013.08.003

vertex he is currently on. So if there is only one available out-link, this link will be chosen with probability one. If there are two links, each will be chosen with probability half, and so on. A sink is considered to link to all other vertices in the graph. Therefore when the random walk reaches a sink, it will then pick as a next step any vertex (including itself) uniformly at random. Additionally, at each step in the graph, the surfer has a small probability $1-\alpha$ of *zapping* instead of choosing a link, in which case the walk restarts at a vertex chosen according to the starting probability distribution. This walk is supposed to represent a surfer of the web, clicking links at random and surfing from page to page, but occasionally using the address bar of the browser for example, and going to any part of the web, not necessarily connected to the current page. This random walk, modelled as a Markov chain, has a unique stationary distribution that assigns a probability π_v to each vertex v in the graph. This is defined to be the PageRank value of vertex v. The PageRank value π_v corresponding to vertex v has the pleasing interpretation as the probability for the random surfer of being at vertex v at any given point in time during the walk. It is also the "fraction of time" that a random surfer spends at vertex v relative to the rest of the vertices, as the length of the walk tends to infinity. The parameter that controls the zapping frequency (random restarts) is the probability of continuing the random walk at each step, $\alpha > 0$. In Brin and Page's original implementation, this parameter was chosen to be $\alpha = 0.85$. A small value for α would make the link structure of the web less relevant for the PageRank values computation.

The high level idea is that the PageRank algorithm will assign high PageRank values to vertices that would appear more often in a random surfer type of walk. In other words, the vertices with high PageRank are hot-spots that will see more random surfer traffic, resulting directly from the link structure of the graph. If we add a small number of new links to the graph, the PageRank values of certain vertices can be affected very significantly.

The *LINK BUILDING problem* [1,2] arises as a natural question: given a specific target vertex in the graph, what is the best set of k links that will achieve the maximum increase for the PageRank of the target vertex? We consider the problem of choosing the optimal set of *backlinks* for maximizing π_x, the PageRank value of some target vertex x. A backlink (with respect to a target vertex x) is a link from any vertex towards x. Given a graph $G(V, E)$ and an integer $k \geqslant 1$, we want to identify the k links to add, pointing towards vertex x in G in order to maximize the resulting PageRank of x, π_x. Intuitively, the new links added should redirect the random surfer walk towards the target vertex as often as possible. For example, adding a new link from a vertex of very high PageRank would usually be a good choice. We show that there is no polynomial time algorithm to find the optimal set of links to add unless $\mathrm{P} = \mathrm{NP}$. Therefore we look at polynomial time approximation algorithms that will find a set of links that result in an increase in PageRank for the target vertex that is always close to the best possible. The best type of approximation algorithm is a fully polynomial-time approximation scheme (FPTAS) which allows us to get a solution that is arbitrarily close to the optimal, while the running time increases polynomially with the added accuracy. We prove that LINK BUILDING does not have fully polynomial-time approximation schemes unless $\mathrm{P} = \mathrm{NP}$. On the positive side, we show that there exists a polynomial time algorithm that achieves a constant approximation ratio: the algorithm provides a solution that achieves a PageRank value for the target vertex that is always within a constant factor of the optimal. We also provide a lower bound for the approximation ratio that our algorithm can achieve and compare it with a corresponding bound of a naive link selection algorithm that is based on PageRank values and out-degrees only.

The link building problem is similar to the problem where a target vertex aims at maximizing its PageRank by adding new out-links. Note that in this case, new out-links can actually decrease the PageRank of the target vertex. This is different to the case of the LINK BUILDING problem (which uses only backlinks) where the PageRank of the target can only increase [6]. For the problem of maximizing PageRank with out-links there are known results [6,7] that show optimal linking structures for a single vertex, or a set of related vertices. A different approach to PageRank optimization using out-links only considers a variation of the PageRank optimization problem where the link weights can be decided or fixed [8]. In this case the problem reduces to a continuous optimization problem and a polynomial time algorithm exists based on linear programming techniques. A discrete version of this problem is also considered in the same work, where different types of out going links can be added, but there are no associated continuous weights. The discrete problem also has a polynomial time algorithm under some additional assumptions. Link prediction [9], is not directly related to the link building problem but it is also concerned with identifying new links that are likely to form in a network. There are several results for many applications such as link recommendations, community detection, inferring missing or unobserved links, and building models for network evolution [10–13]. There are many results related to the computation of PageRank values [14,15] and re-calculating PageRank values after adding a set of new links in a graph [6].

1.1. Outline

In Section 2 we discuss the PageRank algorithm in more detail. We then formally define the LINK BUILDING problem. In Section 3 we develop a theorem expressing among other things how the topology of the graph determines the PageRank potential for a set of backlinks to a given target vertex x. Based on this theorem we show in Section 4 that $\mathrm{NP} \neq \mathrm{P}$ implies that no FPTAS exists for LINK BUILDING and we also show that LINK BUILDING is $W[1]$-hard. In Section 5, we present r-Greedy, a polynomial time algorithm yielding a PageRank value within a constant factor from the optimal and therefore show that the LINK BUILDING problem is in the class APX. The approximation ratio is a function of the parameter α and is upper bounded by 5.7 for $\alpha = 0.85$. In Section 6.1 we consider π-Naive, a naive and intuitively clear algorithm for LINK BUILDING where we choose backlinks from the vertices with the highest PageRank values compared to their out-degree. We show how to construct graphs where we force a high approximation ratio of at least 13.8 for $\alpha = 0.85$. Note that we