Ranking Links on the Web: Search and Surf Engines

Jean-Louis Lassez[1], Ryan Rossi[1], and Kumar Jeev[2]

[1] Coastal Carolina University, USA
{jlassez, raross}@coastal.edu
[2] Johns Hopkins University, USA
kjeev@cs.jhu.edu

Abstract. The main algorithms at the heart of search engines have focused on ranking and classifying sites. This is appropriate when we know what we are looking for and want it directly. Alternatively, we surf, in which case ranking and classifying links becomes the focus. We address this problem using a latent semantic analysis of the web. This technique allows us to rate, suppress or create links giving us a version of the web suitable for surfing. Furthermore, we show on benchmark examples that the performance of search algorithms such as PageRank is substantially improved as they work on an appropriately weighted graph.

Keywords: Search Engines, Surf Engines, Singular Value Decomposition, Heuristic Search, Intelligent Systems.

1 Introduction

The ergodic theorem and/or its associated iterative construction of principal eigenvectors forms the backbone of the main search algorithms on the web. (PageRank [1], HITS [2], SALSA [3]). The standard proofs of the ergodic theorem rely on the Perron Frobenius theorem, which implies the use of a certain amount of mathematical machinery and restrictive hypotheses. A new fundamentally simpler proof of the ergodic theorem was derived in [4]. In a second section we will show how this proof can be used to clarify the role played by Markov models, the Perron Frobenius theorem and Kirchhoff's Matrix Tree theorem in the design of search engines. In a short third section we make a case that the ranking of links should play a major role in the design of surf engines. In the fourth section we first recall how singular value decomposition is used to extract latent semantic features [5, 6]. In the next three subsections we apply this technique to automatically rate and update links, leading to improved efficiency for search algorithms, we then generate surf sessions and extract meta sites and target sites. This construction of meta sites and targets can be used to generate hubs and authorities [2] and the bipartite graphs defined in SALSA [3] and Trawling [7].

2 A Symbolic View: From Kirchhoff to Google

In this section we review the results from [4] and see how they relate to PageRank, SALSA, HITS and other algorithms that form the core of search engines.

N.T. Nguyen et al. (Eds.): IEA/AIE 2008, LNAI 5027, pp. 199–208, 2008.

In the patent application for PageRank we find the statements: "the rank of a page can be interpreted as the probability that a surfer will be at a particular page after following a large number of forward links. The iteration circulates the probability through the linked nodes like energy flows through a circuit and accumulates in important places." The first sentence shows that the web is considered as a Markov chain and that the ranking of sites is given as an application of the ergodic theorem [8], which indeed computes how frequently each site is visited by a surfer. The second sentence is related to Kirchhoff's [9] current law.

The proof of the ergodic theorem is most frequently given as an application of the Perron Frobenius theorem, which essentially states that the probabilities of being at a particular site are given as the coefficients of the principal eigenvector of the stochastic matrix associated to the Markov chain, which is computed as

$$\lim_{n\to\infty} M^n e, \quad \textit{where e is the unit vector}$$

The implementation of the PageRank algorithm uses this construction (as a foundation, there is more to the PageRank algorithm and to the Google search engine), as well as SALSA. So we have two separate problems to consider. One is the use of the Markov Chain model for the web, and the other is the use of the Perron Frobenius theorem as a basis for an implementation. Indeed alternative constructions for computing the most frequently visited sites have been proposed for instance based on Gauss Seidel [10]. And if Kleinberg's HITS algorithm is not based on the Markov Chain model or the ergodic theorem, it nevertheless makes systematic use of the Perron Frobenius theorem.

The ergodic theorem now plays a major role in computer science, but its complete proof is a challenge at least for undergraduate computer scientists. Indeed we have issues of convergence involving further theorems from analysis, computing eigenvalues which involves considerations of complex numbers, issues of uniqueness of solution which creates serious problems leading to restrictive hypotheses and further mathematical machinery [10].

In [11,12] it was shown that elimination theory, based on Tarski's meta theorem could be used to derive strikingly simple proofs of important theorems whose known proofs were very involved. This technique was applied in [4] to the ergodic theorem. We informally present here the essential result that allows us to present the ergodic theorem with minimal mathematical machinery and no overly restrictive hypotheses.

Let G be a graph representing a Markov chain where the nodes s_i are called states (or sites in our application) and the edges represent links between states.

Consider the system of equations below. The x_i are the probabilities of being in state i, while $p_{i,j}$ is the probability of moving from state i to state j. So if we are in state 2 with probability x_2, it is because we were previously in state 1 with probability x_1 and we transitioned with probability p_{12}, or we were in state 4 with probability x_4 and we transitioned to state 2 with probability p_{42}.