6th International Conference on Applied Informatics
Eger, Hungary, January 27–31, 2004.

Application of the MCMC algorithm in Web mining*

Márton Ispány

Department of Applied Mathematics and Probability, University of Debrecen
e-mail: ispany@inf.unideb.hu

Abstract

In this paper we investigate some popular ranking algorithms used in Web searching and mining. We show that these algorithms can be considered as applications of the general Markov Chain Monte Carlo (MCMC) method. By this observation we suggest some new algorithms and we describe how can we study the properties of the old and new algorithms.

Categories and Subject Descriptors: G.3 [Probability and Statistics]: Probabilistic algorithm; H.2.8 [Database Applications]: Data mining - *Web mining, stochastic search*;

Key Words and Phrases: WWW, MCMC, PageRank, HITS

1. Introduction

Nowadays the Web has become a living and growing storehouse of knowledge built in a decentralized and heterogen manner. Besides many positive implications there are some negative impacts of the complexity of the Web. For many queries there are thousands of responses and this fact makes it hard to find the information we need. To resolve this problem certain techniques of machine learning and data mining have become widely applied. Such techniques detect and exploit statistical dependencies between Web pages and hyperlinks. This new and exciting field is called web mining, see the monograph of Chakrabarty [6].

In this paper, we focus mainly on ranking between the Web pages. It is reasonable to measure the rank or "goodness" of a Web page with the number as many times a typical surfer has visited the page. In general, the motion of a typical surfer on the Web can be modeled by a random walk on the WebGraph. In the standard ranking algorithms (PageRank, HITS) this random walk is as simple as possible, i.e. the surfer chooses between the hyperlinks on a web page uniformly

*Supported by the Hungarian Scientific Research Fund under Grant No. OTKA–F032060/2000.

at random. Thus, the rank of a Web page will be a functional of this random walk, and a random walk defines a Markov chain on the WebGraph. The MCMC type algorithms form a commonly used statistical method to evaluate such kind of functionals on a complex state space, like WebGraph.

The paper is organized as follows. In Section 2 the searching infrastructure of the Web is introduced. Section 3 gives a short overview on the WebGraph. In Section 4 the fundamentals of MCMC are presented. Some old and new ranking algorithms are considered in Section 5 from the point of view of MCMC setup. Finally, Section 6 is devoted to open problems and further works.

2. Searching on the Web

In order to find the information we need on the Web we make clear the following questions

- Which is the "proper" model for the Web?
- Which is the "best" searching and information retrieving strategy?
- How do we use the Web?

These quertions define different areas of the web mining. The first one is the basic question of the Web structure mining. To find the right answer for our query we have to know where can it be on the Web. The second one is the heart of the web content mining. The main part of this subject is to rank different web pages addressed to the same topic. Finally, the third one belongs to the field of web usage mining.

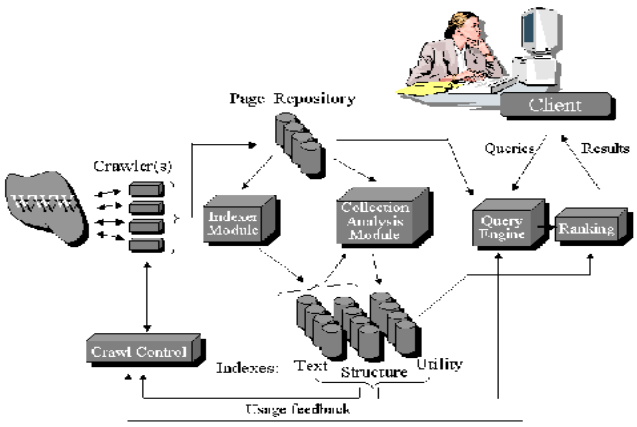

Figure 1: General search engine architecture

Before we describe ranking techniques, it is useful to understand how a Web search engine works. Figure 1 shows such an engine schematically. The basic components of this engine are the followings: crawlers, crawler control module, page repository, indexer module, query engine and ranking module. See Arasu et al. [2] for the detailed role of these components. The ranking module has the task of sorting the results such that results near the top are the most likely to be what we are looking for. This and the algorithms behind him are the main subject of this paper.

3. The WebGraph

We consider the HTML "points–to" relation on Web pages as a directed graph $G = (V, E)$, where the set V of vertices consists of the Web pages, and the set E of directed edges (i, j), which exist iff page i has a hyperlink to page j. In practice this graph is usually pruned to remove self–loops and other forms of spam. Then the graph G is called the WebGraph. Here we summarize some basic properties of the WebGraph:

- Many nodes, 1 billion pages, 15 terabytes;
- Very sparse, average links per page is between 5 and 10;
- Highly dynamic, 1 million new pages per day, over 600 Gbyte of pages change per month.

Due to a comprehensive project organized by Bröder et al. [5] some old conjectures have been rejected concerning the graph structure of the Web. They have been investigated 1.5 billion links on 200 million web pages. Bröder et al. realized that "the old picture—where no matter where you start, very soon you'll reach the entire Web—is not quite right". They have been suggested the so–called bow tie model, see Figure 2. The core of this model is the strongly connected component (SCC) of the WebGraph G. It is the subset of vertices such that for all pairs of pages i, j there exists a directed path from i to j, i.e. $(i, j) \in E$. The left bow consists of the new or obscure web pages, i.e. such pages where there exists a hyperlink to a page belonging to SCC. The right bow mainly consists of commercial sites. Finally, there is a large number of disconnected pages.

Different theoretical models are developed to describe the structure and evolution of the WebGraph. We have been collected the most important ones:

- Random graphs (Erdős, Rényi [7]);
- ACL model for massive graph (Aiello, Chung, Lu [1]);
- Evolving networks (Albert, Barabási, Jeong [4]);
- Copying models (Kumar, Rhagavan [10]);

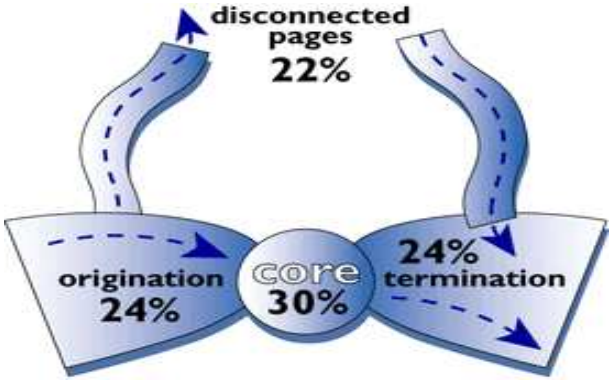

Figure 2: Bow tie model of the Web

- Small Worlds (Watts, Strogats [17]);
- Multi–Layer models (Caldarelli, De Los Rios, Leonardi [11]).

Some of these models, e.g. ACL and Barabási's one, generate a random graph possessing bow tie structure but others, e.g. model introduced by Erdős and Rényi, do not. To find a proper model for the Webgraph is crucial to develope an appropriate ranking algorithm because of the movement of a surfer can be described in some sense as a random walk on the (random) WebGraph.

4. Markov Chain Monte Carlo

Consider a Markov chain $X_0, X_1, X_2, \ldots$ on a state space $\mathcal{X}$, with transition probabilities $P(x, \cdot)$, $x \in \mathcal{X}$, and stationary distribution π. Our aim is to estimate the mean, in other words the average behaviour, of various functionals of this random motion on the state space $\mathcal{X}$. For example, let $\pi(h) := \int_{\mathcal{X}} h \, \mathrm{d}\pi$, where $h : \mathcal{X} \to \mathbb{R}$ is a function. One of the possible estimates for $\pi(h)$ is

$$\widehat{\pi}(h) := \frac{1}{n} \sum_{i=1}^{n} h(X_i). \tag{1}$$

Specific examples and applications of such "MCMC algorithms" include the followings:

- Counting in large combinatoric structures;
- Integrating in high dimension;
- Sampling for complex distributions (e.g. Gibbs sampler);

- Simulating complex stochastic systems (Metropolis–Hastings algorithm).

For references see Gilks et al. [8] and Tierney [16].

5. Ranking algorithms

The first observation of making an efficient ranking module is that the Web is an example of a social network. Networks of social interaction are formed between academics by co–authoring (e.g. Erdős number), movie stars in Hollywood, football stars who played in the same team anytime, contries via trading relation and so on. The basic concepts of this theory like status, prestige and centrality play important role in the theory of Web ranking. The graph theoretical modelling is also common. For example, Seeley [15] early realized the recursive nature of prestige in a social network:

> ...we are involved in an "infinite regress": an actor's status is a function of the status of those who choose him; and their status is a function of those who choose them, and so ad infinitum.

Thus any measurement, which describes the importance or goodness of a web page, can be considered as an invariant functional of a (stochastic) dynamical system.

The general scheme of defining and calculating a ranking measure on the whole or a part of the Web is the following:

- Define a Markov chain on the WebGraph to model a typical surfer with a unique invariant probability measure π. Then π mesures the average residence time of a random surfer at a web page.
- Define a functional $h : G \to \mathbb{R}$ on the WebGraph. By h we want to measure a property of web pages, e.g. prestige or centrality with respect to a topic.
- Estimate $\pi(h)$ by Markov Chain Monte Carlo method.

In the next examples we demonstrate that the two standard ranking algorithm, PageRank and stochastic HITS, work along the above scheme.

The algorithm PageRank has been introduced by Page and Brin [13], the founders of the popular Web search engine Google. We define $d(i)$ as the out–degree of page i, i.e. the number of hyperlinks on page i. The Markov chain behind the PageRank algorithm is defined by the random surfer model. This model supposes that the surfer chooses an out–neighbour of the current page uniformly at random. Denote by $p(i)$ the prestige of the web page i. In PageRank it is supposed that the prestige of a page depends on the prestige of its in–neighbours. This means the following recursion:

$$p(i) = \sum_{(j,i)\in E} d^{-1}(j)p(j), \qquad i \in V. \tag{2}$$

In vector form we have

$$\pi = \pi A, \qquad \text{where} \qquad A_{ij} = \begin{cases} 1/d(i) & \text{if } (i,j) \in V, \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

and $\pi := (p(i),\, i \in V)$. Thus, in case of PageRank the Markov chain defined by the transition matrix A, the prestige is the stationary distribution π of the chain, and the functionals are the page–functionals h_i, $i \in V$, where $h_i(j) = 1$ if $i = j$, and $h_i(j) = 0$ otherwise. In Section 3 we mentioned that the whole WebGraph is not strongly connected. This implies that the Markov chain behind the PageRank is not aperiodic and irreducible. Hence, in general, its stationary distribution is not unique.

To avoid this kind of anomalities Page et al. have been introduced the modified PageRank algorithm. In this case the recursion (2) is replaced by the following:

$$p(i) = \alpha + (1-\alpha) \sum_{(j,i)\in E} d^{-1}(j)p(j), \tag{4}$$

where $0 < \alpha < 1$. This modification already yields an irreducible, aperiodic Markov chain. In the Google $\alpha = 0.15$. In this case the Markov chain behind the model is defined by the jumping surfer model in the following way:

- with probability α, the surfer jumps to a random page on the Web;
- with probability $1 - \alpha$, the surfer chooses an out–neighbour page of the current page uniformly at random.

For the unique stationary distribution we have the linear equation:

$$\pi = \pi \left(\frac{\alpha}{n} E + (1-\alpha) A \right), \tag{5}$$

where $E_{ij} = 1$ for all $i, j \in V$. Numerical methods to solve this equation, like power iteration and Gauss–Seidel method, can be found in Arasu et al. [3].

Another way to resolve the problem of periodicity and reducibility is of using time–sampled chains introduced by Rosenthal [14]. The time–sampled Markov chain is defined by the transition matrix $A_\mu = \sum_n \mu(n) A^n$, where μ is probability distribution on the non–negative integers and A^n is the n–step transition matrix. For example, if $\mu(1) = \mu(2) = 1/2$, then $A_\mu = (A + A^2)/2$. We can interpret this new chain as the average of the motion of two random surfers, where the first one takes one step and the second one takes two steps at same time according to the original random surfing model. One can realize easily time–sampled Markov chain based ranking algorithm by using several crawlers, each crawler browses the Web following the same random surfer model with different step number.

The other well–known ranking algorithm is the HITS (hyperlink induced topic search) introduced by Kleinberg [9]. In HITS and its stochastic variant SALSA (see Lempel and Moran [12]), instead of the whole WebGraph, a root set R with its

neighbours $\overline{R}$ are involved. Here the root set, which is a query dependent graph, is a collection of web pages given by a query of a standard IR system. The two basic concepts of HITS are the authority and centrality. Namely, Klenberg observed that there are two different kinds of web pages, like in the academic literature, popular pages or authorities, which contain definitive high–quality information and link collections, which are comprehensive lists of links to authorities. Denote the authority of a page i by $a(i)$ and the centrality by $c(i)$. Moreover, denote $f(i)$ the in–degree of the page i. Then the stochastic HITS is defined by the recursions

$$a(i) = \sum_{(j,i)\in E} d^{-1}(j)c(j), \qquad c(i) = \sum_{(i,j)\in E} f^{-1}(j)a(j), \tag{6}$$

for all $i \in \overline{R}$. The vector form of these equations is

$$\begin{pmatrix} \boldsymbol{a} & \boldsymbol{b} \end{pmatrix} = \begin{pmatrix} \boldsymbol{a} & \boldsymbol{b} \end{pmatrix} \begin{pmatrix} O & A \\ C & O \end{pmatrix}, \tag{7}$$

where

$$A_{ij} = \begin{cases} 1/d(i) & \text{if } (i,j) \in V, \\ 0 & \text{otherwise} \end{cases}, \qquad C_{ij} = \begin{cases} 1/f(i) & \text{if } (j,i) \in V, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

One can see that the stochastic HITS can be considered as a direct product of a random walk and a reversed random walk on $\overline{R} \times \overline{R}$. For the authority–to–authority transition matrix we have

$$a(i) = \sum_k a(k)p_{ki}, \qquad p_{ki} := f^{-1}(k) \sum_{(j,i),(j,k)\in E} d^{-1}(j). \tag{9}$$

For the centrality–to–centrality matrix we have

$$c(i) = \sum_k c(k)q_{ki}, \qquad q_{ki} := d^{-1}(k) \sum_{(i,j),(k,j)\in E} f^{-1}(j). \tag{10}$$

There are several proposals for functional h to measure different properties of the web pages. For example

- environment functional of the page i, i.e. $h = \sum_{(i,j)\in E} I_j$;
- portal functional defined by $h = I_P$, where P is a collection of web pages that form a portal;
- topic functional, i.e. $h : V \to [0,1]$, where $h(i)$ is the goodness of the page i with respect to a topic.

Finally, consider the following more complicated example for ranking. Suppose that we would like to measure the goodness of web pages from several different point of view. Let K be the number of topics, and denote by $p(i,k)$ the prestige

or rank of the web page i with respect to the kth topic. Define the following modification of the standard PageRank:

$$p(i,k) = \sum_{(j,i)\in E} \sum_{\ell=1}^{K} \frac{w(k,\ell)}{d(j)} p(j,\ell), \tag{11}$$

where $W = \{w(k,\ell)\}$ is a weight matrix, which measures the strength of the interactions between the topics. By this way we have a Markov chain, and the general MCMC scheme can be applied.

6. Further work

In the last section we summarize some further works we have to do in order to make a really efficient ranking algorithm.

The first one is to find better and better model for the WebGraph. We have to emphasize that the Web evolves like a living creature. It is changing for time to time, and its structure can also change. We mention, for example, that in the beginning the Web was connected. So it happens that the bow tie model will be no longer valid.

The second one is to find an appropriate model for descibing the motion of a typical surfer on the Web. We think that the random and jumping surfer models are too simple.

The third one is to find the most efficient MCMC methods to evaluate the goodness of web pages from different point of view. A proposal, which can work well, is the multiple MCMC. We observed that an MCMC algorithm for Web ranking can be realized by a crawler. However, it is known that the largest crawler covers less than 16% of the WebGraph. Thus, if we want to develope a ranking algorithm for the whole Web, then we need to use as much crawler as possible.

References

[1] Aiello, W., Chung, F., Lu, L.: Random Evolution in Massive Graph, in *Handbook on Massive Data Sets*, Eds. Abello et al., Kluwer, 2002, 97–122.

[2] Arasu, A., Cho, J., Garcia–Molina, H., Paepcke, A., Raghavan, S.: Searching the Web. *ACM Trans. Internet Tech.*, Vol. 1, No. 1 (2001), 2–43.

[3] Arasu, A., Novak, J., Tomkins, A., Tomlin, J.: PageRank computation and the structure of the Web: Experiments and algorithms. *Technical Report*, IBM Almaden Research Center, 2001.

[4] Barabási A.L., Albert, R., Jeong, H.: Scale-free characteristics of random networks: the topology of the world-wide web, *Physica A*, Vol. 281 (2000), 69–77.

[5] Bröder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., Wiener, J.: Graph structure in the Web: Experiments and models. In WWW9, pp. 309–320, Amsterdam, May 2000. Elsevier Science.

[6] Chakrabarti, S.: *Mining the Web, Discovering Knowledge from Hypertext Data*, Morgan Kaufmann, 2003.

[7] Erdős, P., Rényi, A.: On random graphs. I. *Publ. Math. Debrecen*, Vol. 6 (1959) 290–297.

[8] Gilks, W.R., Richardson, S., Spiegelhalter, D.J.(eds.): *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London, 1996.

[9] Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. *Proceeding of ACM–SIAM Symposium on Discrete Algorithms*, 1998.

[10] Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., Upfal, E.: Stochastic models for the web-graph. *Proc. 41st Annual Symp on Foundations of Computer Science*, 2000.

[11] Laura, L., Leonardi, S., Millozzi, S., Caldarelli, G., De Los Rios, P.: A study of the properties of the Webgraph. *2nd workshop on Algorithms and Models for the Webgraph*, May 20, in conjunction with WWW03.

[12] Lempel, R., Moran, S.: SALSA: The stochastic approach for link–structure analysis. *ACM Transactions on Information Systems*, Vol. 19., No. 2 (2001), 131–160.

[13] Page, L., Brin, S., Motwani, R., Winograd, T.: The PageRank citation ranking: Bringing order to the Web. Manuscript. *google.stanford.edu/∼backrub/pageranksub.ps.*

[14] Rosenthal, J.S.: Geometric convergence rates for time–sampled Markov chains. *J. Theor. Probab.*, Vol. 16, No. 3 (2003), 671–688.

[15] Seeley, J.R.: The net of reciprocal influence: A problem in treating sociometric data. *Canadian Journal of Psychology*, Vol. 3 (1949), 234–240.

[16] Tierney, L.: Markov chains for exploring posterior distribution (with discussion). *Ann. Stat.*, Vol. 22 (1994), 1701–1762.

[17] Watts, D.J., Strogatz, S.H.: Collective dynamics of "small-world" networks, *Nature*, Vol. 393 (1998), 440-442.

Postal address

Márton Ispány
Department of Applied Mathematics and Probability
University of Debrecen
Egyetem tér 1. Pf. 12, 4010 Debrecen
Hungary