Static Ranking of Web Pages, and Related Ideas

Wray Buntine
Complex Systems Computation Group,
Helsinki Institute for Information Technology
P.O. Box 9800, FIN-02015 HUT, Finland
wray.buntine@hiit.fi

Abstract

This working paper reviews some different ideas in link-based analysis for search. First, results about static ranking of web pages based on the so called random-surfer model are reviewed and presented in a unified framework. Second, a topic-based hubs and authorities model using a discrete component method (a variant of ICA and PCA) is developed, and illustrated on the 500,000 page English language Wikipedia collection. Third, a proposal is presented to the community for a Links/Ranking consortium extracted from the Web Intelligence paper Opportunities from Open Source Search.

1 Introduction

PageRank^{TM} used by Google and the Hypertext-Induced Topic Selection (HITS) model developed at IBM [9] are the best known of the ranking models although they represent a very recent part of a much older bibliographic literature (for instance, discussed in [5]). PageRank ranks all pages in a collection and is then used as a static (i.e., query-free) part of a query evaluation. Whereas HITS is intended to be applied to just the subgraph of pages retrieved with a query, and perhaps some of their neighbors. There is nothing, however, to stop HITS being applied like PageRank to a full collection rather than just query results.

PageRank is intended to measure the *authority* of a webpage on the basis that high authority pages have other high authority pages linked to them. HITS is also referred to as the hubs and authority model: a *hub* is a web page that is viewed as a reliable source for links to other web pages, whereas an *authority* is viewed as a reliable content page itself. Generally speaking, good hubs should point to good authorities and *visa verca*. The literature about these methods is substantial [2, 1].

Here I review these two models, and then discuss their use in an Open Source environment.

2 Random Surfers versus Random Seekers

The PageRank model is based on the notion of an idealised *random surfer*. The random surfer starts off by choosing from some selection of pages according to an initial probability vector $\vec{s}$. When at a new page, the surfer can take one of the outgoing links from the current page, or with some smaller probability restart afresh at a completely new page again using the initial probability vector $\vec{s}$. The general start-restart process is depicted in the graph in Figure 1, where the initial state is labelled *start*, and the pages themselves form a subgraph T. Every page in the collection has

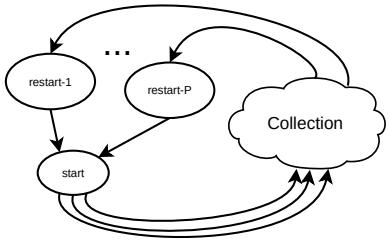

Figure 1. Start-Restart for the Random Surfer

a link to a matching *restart* state leading directly to *start*, and *start* links back to those pages in the collection with a non-zero entry in $\vec{s}$. Note the *restart* states could be eliminated, but are retained for comparison with the later model. This represents a Markov model once we attach probabilities to outgoing arcs, and the usual analysis of Markov chains and linear systems (see for instance [12]) applies [1]. The computed static rank is the long run probability of visiting any page in the collection according to the Markov model.

Extensions to the model include making the initial

probability vector $\vec{s}$ dependent on topic [7, 11], providing a back button so the surfer can reject a new page based on its unsuitability of topic [11, 10], and handling the way in which pages with no outgoing links can be dealt with [6, 1]. These extensions make the idealised surfer more realistic, yet no real analysis of the Markov models on real users has been attempted. A fragment of a graph illustrating the Markov model from the point of view of surfing from one *page*, is given in Figure 2, From

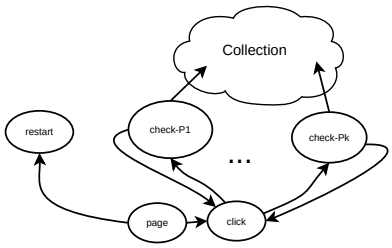

Figure 2. Local view of Primitive States

a *page* j, the surfer decides to either restart with probability r_j, or to *click* on a link to a new page. Once they decide to click, they try different pages k with probability given by matrix $\mathbf{p}$, where

$$p_{j,k} = \begin{cases} 0 & \text{page } j \text{ has no out link to } k \\ 1/L & \text{page } j \text{ has } L \text{ outlinks, one to } k \end{cases}$$

but have a one time opportunity (to *check*) to either accept the new page k, given by a_k, or to try again and go back to the intermediate *click* state. Folding in the various intermediate states (*click* and the *check* states) and just keeping the *pages* and the *start* and *restart* states, yields a transition matrix starting from a page j of

$$p(state \,|\, \text{page } j) = \begin{cases} r_j & state = restart \\ (1-r_j)\frac{p_{j,k}a_k}{\sum_k p_{j,k}a_k} & state = \text{page } k \end{cases} \quad (1)$$

Note in this formulation, if a page j has no outgoing links, then $r_j = 1$ necessarily. This has the parameters summarised in the following table.

	Description
$\vec{s}$	initial probabilities for pages, normalised
$\vec{r}$	restart probabilities for pages
$\vec{a}$	acceptance probabilities for pages

With appropriate choice of these, all of the common models in the literature can be handled [7, 11, 1].

A new model proposed by Amati, Ounis, and Plachouras [13] is the static absorbing model for the web. The absorbing model is instead based on the notion of a random seeker. The random seeker again surfs the web, but instead of continuously surfing, can "find" a page and thus stop. The general model comparable to Figure 1 is now given by Figure 3, In the random seeker

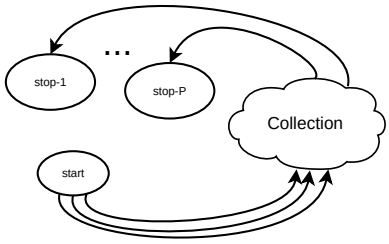

Figure 3. Start-Stop for the Random Seeker

model, the computed static rank is the long run probability of stopping at ("finding") any given page. It is thus given by the probabilities for the absorbing states in the Markov model, and again the usual analysis applies. The page to page transition probabilities, however, can otherwise be modelled in various ways using Equation (1).

The structure of the graphs suggests that these two models (random surfer versus random seeker) should have a strong similarity in their results. We can work out the exact probabilities by folding the transition matrices. The following lemma do this.

Lemma 1. *Given the random seeker model with parameters $\vec{s}$, $\vec{r}$, $\vec{a}$ and $\mathbf{p}$, where $r_j = 1$ for any page j without outgoing links. Let $\mathbf{P}$ denote the transition matrix*

$$P_{j,k} = \begin{cases} 0 & j \text{ not linked to } k \\ p_{j,k}a_k/(\sum_k p_{j,k}a_k) & \text{page } j \text{ linked to } k \end{cases}$$

Let $\mathbf{D_r}$ denote the diagonal matrix formed using entries $\vec{r}$. The total probability of the stop states for paths of length less than or equal to $n+2$ is given by

$$\mathbf{D_r}\left(\mathbf{I} + \sum_{i=1}^{n}((\mathbf{I}-\mathbf{D_r})\mathbf{P})^i\right)\vec{s} \quad (2)$$

This can be proven by straight forward enumeration of states. Equation (2) is evaluated in practice using a recurrence relation such as $\vec{q}_0 = \vec{s}$, $\vec{p}_{i+1} = \vec{p}_i + \mathbf{D_r}\vec{q}_i$ and $\vec{q}_{i+1} = (\mathbf{I}-\mathbf{D_r})\mathbf{P}\vec{q}_i$.

Lemma 2. *Given the random seeker model with parameters $\vec{s}$, $\vec{r}$, $\vec{a}$ and $\mathbf{p}$, where $r_j = 1$ for any page j without outgoing links,, and let $\mathbf{P}$ and $\mathbf{D_r}$ be defined as above. Assume $r_j > 0$ for all pages j. The total absorbing probability of the stop states is given by*

$$\mathbf{D_r}\left(\mathbf{I} - (\mathbf{I}-\mathbf{D_r})\mathbf{P}\right)^{-1}\vec{s} \quad (3)$$

The matrix inverse exists. Moreover, the L_1 norm of this minus Equation (2) is less than $(1-r_0)^{n+1}/r_0$ where $r_0 = \min_j r_j > 0$.

Note in the standard PageRank interpretation, $r_0 = 1-\alpha$, so the remainder is $\alpha^{n+1}/(1-\alpha)$, the same order as for the PageRank calculation [1].

Proof. Consider $\vec{q}_i = ((\mathbf{I}-\mathbf{D_r})\mathbf{P})^i \vec{s}$, and prove the recursion $||\vec{q}_i||_1 \leq (1-r_0)^i$. Since $\mathbf{P}$ is a probability matrix with some rows zero, $||\mathbf{P}\vec{q}_i||_1 \leq ||\vec{q}_i||_1$ and hence $||\vec{q}_{i+1}||_1 \leq (1-r_0)^{i+1}$. Consider $\vec{q}_{n,m} = \left(\sum_{i=n+1}^{m} ((\mathbf{I}-\mathbf{D_r})\mathbf{P})^i\right)\vec{s}$. Hence $||\vec{q}_{n,m}||_1 \leq \sum_{i=n+1}^{m}(1-r_0)^i$ which is $((1-r_0)^{n+1}-(1-r_0)^m)/r_0$. Thus $\vec{q}_{n,\infty}$ is well defined, and has an upper bound of $(1-r_0)^{n+1}/r_0$. Thus the total absorbing probability is given by Equation (2) as $n \to \infty$, with L_1 error after n steps bounded by $(1-r_0)^{n+1}/r_0$. Since the sum is well defined and converges, it follows that $(\mathbf{I}-(\mathbf{I}-\mathbf{D_r})\mathbf{P})^{-1}$ exists. □

Lemma 3. *Given the random surfer model with parameters $\vec{s}$, $\vec{r}$, $\vec{a}$ and $\mathbf{p}$, where $r_j = 1$ for any page j without outgoing links,, and let $\mathbf{P}$ and $\mathbf{D_r}$ be defined as above. Assume $r_j > 0$ and $s_j > 0$ for all pages j. Then the long run probability over pages exists independently of the initial probability over pages and is proportional to*

$$(\mathbf{I}-(\mathbf{I}-\mathbf{D_r})\mathbf{P})^{-1}\vec{s} \tag{4}$$

Proof. Eliminate the *start* and *restart* states, and then the transition matrix becomes as follows: given a probability over pages of $\vec{p}_i$, then at the next cycle

$$\vec{p}_{i+1} = \vec{s}(\vec{r}^\dagger\vec{p}_i) + (\mathbf{I}-\mathbf{D_r})\mathbf{P}\vec{p}_i$$

Since $\vec{r}$ and $\vec{s}$ are strictly positive, the Markov chain is ergodic and irreducible [12], and thus the long run probability over pages exists independently of the initial probability over pages. Consider the fixed point for these equations. Make a change of variables to $\vec{p} = \mathbf{D_r}\vec{p}/(\vec{r}^\dagger\vec{p})$. This is always well defined since the positivity constraints on $\vec{r}$ ensure $\vec{r}^\dagger\vec{p} > 0$. Then

$$\vec{p}' = \mathbf{D_r}\vec{s} + \mathbf{D_r}(\mathbf{I}-\mathbf{D_r})\mathbf{P}{\mathbf{D_r}}^{-1}\vec{p}'$$

Rewriting,

$$\mathbf{D_r}\left(\mathbf{I}-(\mathbf{I}-\mathbf{D_r})\mathbf{P}\right){\mathbf{D_r}}^{-1}\vec{p}' = \mathbf{D_r}\vec{s}$$

We know from above that the inverse of the middle matrix expression exists. Thus

$$\vec{p}' = \mathbf{D_r}\left(\mathbf{I}-(\mathbf{I}-\mathbf{D_r})\mathbf{P}\right)^{-1}\vec{s}$$

Substituting back for $\vec{p}$ yields the result. □

Note the usual recurrence relation for computing this is

$$\vec{p}_{i+1} += \vec{s}(\vec{r}^\dagger\vec{p}_i) + (\mathbf{I}-\mathbf{D_r})\mathbf{P}\vec{p}_i \ ,$$

and due to the correspondence between Equations (3) and Equations (4), the alternative occurrence for the absorbing model could be adapted as well. The recurrence relation holds: $\vec{q}_0 = \vec{s}$, $\vec{p}_{i+1} = \vec{p}_i + \vec{q}_i$ and $\vec{q}_{i+1} = (\mathbf{I}-\mathbf{D_r})\mathbf{P}\vec{q}_i$, noting that the final estimate $\vec{p}_{i+1}$ so obtained needs to be normalised. This can, in fact, be supported on a graphical basis as well.

This correspondence gives us insight into how to improve these models. How might we make the Markov models more realistic? Could the various parameters be learned from click stream data? While in the surfing model $\vec{r}$ corresponds to the probability of restarting, in the seeking model it is the probability of accepting a page and stopping. One is more likely to use the back button on such pages, and thus perhaps the acceptance probabilities $\vec{a}$ should be modified. Some versions are suggested in [6].

3 Probabilistic Hubs and Authorities

A probabilistic authority model for web pages, based on PLSI [8], was presented by [5]. By using the Gamma-Poisson version of Discrete PCA [4, 3], a generalisation of PLSI using independent Gamma components, this can be extended to a probabilistic version of the hubs and authorities model. The method is topic based in that hubs and authorities are produced for K different topics. An authority matrix $\mathbf{\Theta}$ gives the authority score for each page j for the k-th topic, $\theta_{j,k}$, normalised for each topic. Each page j is a potential hub, with hub scores $l_{j,k}$ for topic k taken from the hub matrix $\mathbf{l}$. The links in a page are modelled independently using the Gamma$(1/K, 1)$ distribution. The occurrences of link j in page i are then Poisson distributed with a mean given by authority scores for the link weighted by the hub scores for the page, Poisson$(\sum_k l_{i,k}\theta_{j,k})$. More details of the model, and the estimation of the authority matrix and hub matrix are at [3],

To investigate this model, the link structure of the English language Wikipedia from May 2005 was used as data. The output of this analysis is given at `http://cosco.hiit.fi/search/MPCA/HA100.html`. This is about 500,000 documents and $K = 100$ hub and authority topics are given. The authority scores are the highest values for a topic k from the authority matrix $\mathbf{\Theta}$, and the hub scores are the highest component estimates for topic k for $l_{j,k}$ for a page j.

Note a variety of hub and authority models have been investigated in the context of query evaluation [2]. It is not clear if this is the right approach for using these models. Nevertheless, these represent another family of link-based systems than can be used in a search engine, and an alternative definition of authority to the previous section.

4 A Trust/Reputation Consortium for Open Source Ranking

Having reviewed some methods for link analysis, let use now consider their use. Opportunities for their use abound once the right infrastructure is in place for open source search. Here I describe one general kind of system that could exist in the framework, intended either as an academic or commercial project.

On Google the ranking of pages is influenced by the PageRank of websites. Sites appearing in the first page of results for popular and commercially relevant queries get a significant boost in viewership, and thus PageRank has become critical for marketing purposes. This method for computing authority for a web page borrows from early citation analysis, and the broader fields of trust, reputation, and social networks (which blog links could be interpreted to represent) provide new opportunities for this kind of input to search. Analysis of large and complex networks such as the Internet is readily done on todays grid computing networks.

What are some scenarios for the use of new kinds of data about authority, trust and reputation, standards set up by a consortium perhaps. A related example is the new OpenID[1], a distributed identity system.

ACM could develop a "computer science site rank" that gives web sites an authority ranking according to "computer science" relevance and reputation. In this ranking the BBC Sports website would be low, Donald Knuth's home page high, and Amazon's Computer Science pages perhaps medium. Our search engines can then incorporate this authority ranking into their own scores when asked to do so. ACM might pay for the development and maintenance of this ranking as a service to its members, possibly incorporating its rich information about citations as well, thus using a sophisticated reputation model well beyond simple PageRank. In an open source search network, consumers of these kinds of organisational or professional ranks could be found.

To take advantage of such a system, a user could choose to search Australian university web sites via a P2P universities search engine and then enrol with the ACM ranking in order to help rank their results.

Yahoo could develop a vendor web site classification that records all websites according to whether they primarily or secondarily perform retail or wholesale services, product information, or product service, extending its current Mindset demonstration[2]. This could be coupled with a vendor login service so that venders can manage their entries, and trust capabilities so that some measure of authority exists about the classifications. Using this, search engines then have a trustworthy way of placing web pages into different product genres, and thus commercial and product search could be far more predictable.

To take advantage of this, a user could search for product details, but enrol with the Yahoo service classification to restrict their search to relevant pages.

Network methods for trust, reputation, community groups, and so forth, could all be invaluable to small local search engines, that cannot otherwise gain a global perspective on their content. They would also serve as a rich area for business potential.

[1] http://www.openid.net/

[2] Search for Mindset at Yahoo Research

References

[1] L. A.N., , and M. C.D. Deeper inside pagerank. *Internet Mathematics*, 1(3):335–400, 2004.

[2] A. Borodin, G. Roberts, J. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *ACM Trans. Inter. Tech.*, 5(1):231–297, 2005.

[3] W. Buntine. Discrete principal component analysis. submitted, 2005.

[4] J. Canny. GaP: a factor model for discrete data. In *SIGIR 2004*, pages 122–129, 2004.

[5] D. Cohn and H. Chang. Learning to probabilistically identify authoritative documents. In *Proc. 17th International Conf. on Machine Learning*, pages 167–174. Morgan Kaufmann, San Francisco, CA, 2000.

[6] N. Eiron, K. McCurley, and J. Tomlin. Ranking the web frontier. In *WWW '04: Proceedings of the 13th international conference on World Wide Web*, pages 309–318, 2004.

[7] T. Haveliwala. Topic-specific pagerank. In *11th World Wide Web*, 2002.

[8] T. Hofmann. Probabilistic latent semantic indexing. In *Research and Development in Information Retrieval*, pages 50–57, 1999.

[9] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[10] F. Mathieu and M. Bouklit. The effect of the back button in a random walk: application for pagerank. In *WWW Alt. '04: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pages 370–371, 2004.

[11] M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in pagerank. In *NIPS*14*, 2002.

[12] S. Ross. *Introduction to Probability Models*. Academic Press, fourth edition, 1989.

[13] I. O. V. Plachouras and G. Amati. The static absorbing model for hyperlink analysis on the web. *Journal of Web Engineering*, 4(2):165–186, 2005.