Outlink Estimation For Pagerank Computation Under Missing Data

Sreangsu Acharyya
Department of E.C.E
University of Texas, Austin.
sreangsu@ece.utexas.edu

Joydeep Ghosh
Department of E.C.E
University of Texas, Austin.
ghosh@ece.utexas.edu

ABSTRACT

The enormity and rapid growth of the web-graph forces quantities such as its pagerank to be computed under missing information consisting of outlinks of pages that have not yet been crawled. This paper examines the role played by the size and distribution of this missing data in determining the accuracy of the computed pagerank, focusing on questions such as (i) the accuracy of pageranks under missing information, (ii) the size at which a crawl process may be aborted while still ensuring reasonable accuracy of pageranks, and (iii) algorithms to estimate pageranks under such missing information. The first couple of questions are addressed on the basis of certain simple bounds relating the expected distance between the true and computed pageranks and the size of the missing data. The third question is explored by devising algorithms to predict the pageranks when full information is not available. A key feature of the "dangling link estimation" and "clustered link estimation" algorithms proposed is that, they do not need to run the pagerank iteration afresh once the outlinks have been estimated.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering

General Terms

Algorithms

1. INTRODUCTION

Selecting and ordering query results, from over 3 billion hyperlinked pages that now constitutes the web graph $G(V, E)$ is a difficult web mining problem of extreme importance and one in which link analysis plays a key role. The enormity and dynamic nature of the web-graph, and especially its rapid rate of growth, forces link analysis based ranking schemes like Pagerank to operate under a significant amount of outdated and missing data, present in the form of unknown outlinks from uncrawled pages. This naturally raises questions on the accuracy of the pageranks under such severe conditions, for example, how much of the web graph need to be traversed so that enough faith can be ascribed to the computed pagerank values, or how may one estimate the unknown outlinks and incorporate them in the pagerank calculation without re-running the iteration from scratch again. These are examined below.

WWW2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

2. PAGERANK ITERATION AND INCOMPLETE DATA

The lack of information about the outlinks of the uncrawled pages gets expressed in the Pagerank iterations as incomplete rows of the transition matrix whose stationary distribution is the Pagerank vector. Thus one either has to remove the uncrawled but known vertices for calculation or substitute a predicted distribution (normalized outlink vectors) in its place. Here we show this lack of information may seriously affect the accuracy of the Pagerank vector. But first we define what we mean by accurate Pageranks on a subgraph of the web.

Definition 1. *Given a subset V_k of the vertices of the web-graph $G(V, E)$, the* true *pageranks of V_k are defined to be those that are calculated on the subgraph $G'(V_k, E_k)$ induced by the vertices V_k, i.e. G' contains all and only those edges $xy \in E$, s.t. $x, y \in V_k$.*

At any stage of operation of a pageranking engine, the entire set of web pages V can be divided into the crawled set C and its complement, the uncrawled set C'. We define *forward* set of C as $F = \{p : \exists(q \in C)|(q, p) \in E\}$. Henceforth the shorthand $q \to p$ is used to denote $(q, p) \in E$. For pages in the set of known but uncrawled pages $F_{C'} = \{F \cap C'\}$ the pageranking engine will have no knowledge of their outward links. This constitutes missing information under which the ranking scheme has to operate. We call the set $\{C \cup F_{C'}\}$ the known set V_k and represent its cardinality by $N_k = |V_k|$.

Definition 2. **Robustness:** *Given an incomplete transition matrix of size N, and a distribution $p(\cdot)$ from which the unspecified rows (normalized outlink vectors) have been drawn, the resulting pageranks are called robust if the expected distance between the calculated and true pageranks is of order $O(1)$.*

Proposition 1. *For outlinks drawn with $p(.)$ as uniform, the pagerank computation is robust if the size of the set of unknown vertices of the order $O(\sqrt{N_k})$.*

The proof is omitted for lack of space and can be found in an expanded version [4]. From practical standpoint, however, the assumption of uniform distribution of outlinks is debatable. Web statistics suggest that outlinks are a lot sparser than those generated from uniform sampling over the entire regular unit N simplex. A closer approximation would be that they are sampled from lower dimensional simplices that constitute the boundary of our original N-dimensional simplex. This would increase the expected distance, for a distribution uniform over the lower dimensional simplex n.

This points to the fact that careful imputations of the unknown values in the transition matrix may be required. One adhoc method suggested in the original Pagerank paper [2] is to ignore the unknown rows and work with fully know submatrix corresponding to pages in C. Once the stationary distribution of the induced transition has been computed, the final ranks are computed by running *multiple*, but fixed number of iterations of the Pagerank algorithm applied to vertices in $F_{C'}$ to estimate their pageranks. We show that a *single* iteration suffices under certain assumptions. This paper examines schemes with which one may apportion a pagerank value to the pages in $F_{C'}$ that incorporates the unknown outlinks.

3. DANGLING LINK ESTIMATION

Unlike the method where the unknown rows of the transition matrix are filled by a non-committal uniform distribution, one may use the expected distribution i.e. the distribution of the state transition events averaged over an infinite time, which under mild conditions converge to the stationary distribution or the pagerank values. Replacing unknown values by their expectation has been a standard method of imputation, however in this case we can argue for it even more strongly by drawing upon studies conducted on the power law nature of web graphs. It has been shown that preferential attachment of links is crucial for explaining such power laws. A model where web vertices link to other vertices proportional to the pagerank values generate power laws which are similar to those observed in practice [3].

One is obviously tempted to run this scheme iteratively where the next pagerank estimates are computed by replacing the unknown rows by the current pagerank values till convergence. Thus we are looking for a vector (equivalently, a probability distribution) $\mathbf{r}$ such that if we substitute it in the place of the unknown rows we get back $\mathbf{r}$ as our pagerank, i.e. $\mathbf{r}$ is a fixed point. The fixed point can however be computed analytically without the need for such iterative updates.

Proposition 2. *Calculating the pagerank of pages in C followed by a* **single** *pagerank iteration on the incomplete transition matrix provides the valid pagerank under the assumption that entries in the unknown rows have the same outlink distribution as the converged pagerank vector.*

4. CLUSTERED LINK ESTIMATION

Our objective here is to estimate the unspecified rows of the pagerank matrix T, i.e. the conditional distribution table $P(y_2|y_1)$ and its corresponding stationary distribution $\mathbf{r}$ as the new pagerank vector. For that we introduce a latent variable model, which unlike those proposed for co-occurrence data [1] is applicable even when closed world assumption is not made, allowing one to use in our dynamic setup.

The probability of page y_1 linking y_2 is expressed through latent variables Z. We pose the model in terms of a single random variable Z by introducing constraints that the row and column marginals of its joint distribution are identically distributed. This constraint also has another significance that unlike a general joint distribution over discrete random variables this can mapped directly into a Markov chain on

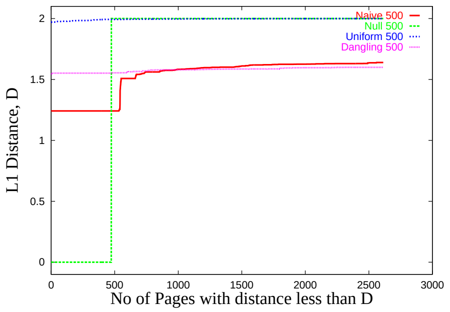

Figure:1 Comparison of outlink Prediction by Dangling Link, Naive Bayes, Null and Uniform distribution for a crawl set of 500 pages.

K states, thereby making it possible to compute the pagerank in the coarser representation and estimating the final pageranks from it. The model is represented as

$$\begin{aligned} P(y_2|y_1) &= \frac{1}{P(y_1)} \sum_{z_1} \sum_{z_2} P(y_1|z_1)P(z_1).P(z_2|z_1)P(y_2|z_2) \\ &= P(y_2). \sum_{z_1} \sum_{z_2} P(z_1|y_1) \frac{P(z_2, z_1)}{P(z_1).P(z_2)}.P(z_2|y_2) \qquad (1) \end{aligned}$$

Now except for $P(y_2)$, which is essentially the pagerank of y_2, no other term assigns probabilities over the set Y and therefore can be modeled by a fixed set of parameters, irrespective of whether the set Y changes or not, this property allows us to use the model even in a dynamic scenario. However we do seem to have a chicken and egg problem because to estimate the pagerank $P(y_2)$ we need the transition probabilities $P(y_2|y_1)$ which in turn requires $P(y_2)$. Let us express the equation above in a more compact matrix notation, let $\Lambda[1,2] = \frac{P(z_2,z_1)}{P(z_1),P(z_2)}$, $U[i,j] = P(Z(y_j) = i|y_j)$, a diagonal matrix $R[i,i] = P(y_i) = r_i$ and $\mathbf{r}[i] = P(y_i)$. Using the equation (1) above and the stationarity property of the pageranks one has

$$T^T\mathbf{r} = R[U^T\Lambda^T U].\mathbf{r} = \mathbf{r} \text{ or, } [U^T\Lambda^T U]\mathbf{r} = R^{-1}\mathbf{r} = \mathbf{1} \qquad (2)$$

Given the matrices Λ and U the linear equation above in $|Y| = N_k$ unknowns can be solved using iterative scaling that chooses the maximum entropy solution under the linear constraints. We have omitted the derivation of the E step and that of the M step under row column marginal constraints. Full details maybe found at [4]. The L1 distance between the true and the predicted rows of the transition matrix is shown for a small subset of the webgraph from utexas domain in figure (1), a naive Bayes classifier taking the inlinks as input was used as a comparison including a predictor which always predicted an uniform distribution and another which predicted that the page did not have any outlinks .

5. REFERENCES

[1] T. Hofmann and J. Puzicha. Unsupervised learning from dyadic data. Technical Report TR-98-042, University of California, Berkeley, Berkeley, CA, 1998.

[2] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web, 1998.

[3] G. Pandurangan, P. Raghavan, and E. Upfal. Using PageRank to Characterize Web Structure. In *8th Annual International Computing and Combinatorics Conference (COCOON)*, 2002.

[4] www.lans.ece.utexas.edu/ srean/wip/missing.pdf.