Predictive Ranking

A Novel Page Ranking Approach by Estimating the Web Structure

Dept. of Computer Science and Eng.
The Chinese University of Hong Kong

Haixuan Yang, Irwin King, Michael R. Lyu
{hxyang, king, lyu}@cse.cuhk.edu.hk

MOTIVATIONS

The page ranks computed by PageRank are inaccurate because of its incomplete information about the Web structure. To avoid such inaccuracy, we

☐ Distinguish different kinds of nodes:

- **Type 1: nodes that are found, but not visited or not visited successfully.**
- **Type 2: those visited but without outlink.**
- **Type 3: those visited and with at least one outlink.**

☐ Predict the web structure according to assumptions:

- **The number of found links to a page is proportional to the real number of links to the page.**
- **The estimated links are shared by all Type 1 pages.**

NOTATION

- **V: all the pages that are found (|V|=n);**
- **S: all the pages of Type 3 (|S|=m);**
- **D^1 : all the pages of Type 2 ($|D^1|=m_1$);**
- **D^2: all the pages of Type 1 ($|D^2|=n-m-m_1$) ;**
- **$d^-(i)$: the true indegree of node i;**
- **$fd^-(i)$: the number of the found links to node i;**
- **A: the matrix that models the users' behavior; the element a(i,j) means the probability of jumping from node j to node i.**
- **$l_i = (d^-(i) - fd^-(i))/m_2 Z$, Z is a normalized factor.**

PREDITIVE RANKING MODEL

(1) Predict $d^-(i)$ by the formula: $d^-(i) = n/(m+m_1)\, fd^-(i)$

(2) Divide A by

$$A = \begin{pmatrix} C & P & M \\ D & Q & N \end{pmatrix}$$

where C and D are used to model the known link structure from S to D^1 and D^2 , c(i,j) in C and d(i,j) are defined as $1/d_j$ if there is a link from j to i and 0 otherwise. M and N are used to model the link structure from D^2 to D^1 and D^2 , and are defined according to the estimated link information:

$$\begin{pmatrix} M & N \end{pmatrix}^T = diag\{l_1, l_2, \ldots, l_n\} 1_{n\times(n-m-m_1)}$$

P and Q are used to model the link structure from D^1 to D^1 and D^2 , and are defined in (4).

(3) Assume that the users will jump to page i with a probability of f_i. The matrix modeling the teleportaion is fe^T .

(4) Assume that the user follows the same kind of teleportation as in (3) when there is no outlink that the users can follow. So

$$\begin{pmatrix} P & Q \end{pmatrix}^T = diag\{f_1, f_2, \ldots, f_n\} 1_{n\times m_1}$$

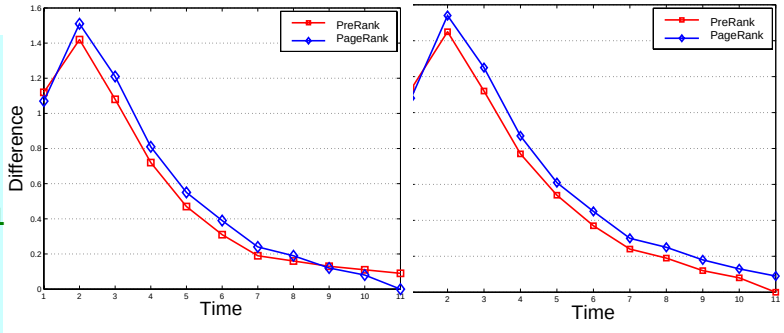
(5) Rank x_i should satisfy

$$x = [(1-\alpha) fe^T + \alpha A]x$$

EXPERIMENTS

Time	1	2	3	4	5	6	7	8	9	10	11
Pages viited	7,712	78,662	109,383	160,019	252,522	301,701	373,579	411,724	444,974	471,684	502,610
Pages found	18,542	120,970	157,196	234,701	355,720	404,728	476,961	515,534	549,162	576,139	607,170

- **Both PreR and PageRank are applied to these 11 datasets collected at different time.**
- **Compute difference between the early results with the PreR and PageRank results at time 11.**
- **The left figure is based on PageRank result at time 11. The right is based on PreR at time 11.**



CONCLUSION

■ **The results of PreR are more accurate than those of PageRank even when we consider PageRank as the reference (bias against ours).**

■ **Our approach to predicting the Web structure is simple and efficient.**

REFERENCES

[1] N. Eiron, K. S. McCurley, and J. A. Tomlin. Ranking the Web frontier. In Proc. of the 13th World Wide Web Conference, pages 309–318, 2004.
[2] S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H.Golub. Exploiting the block structure of the Web for computing pagerank. Technical report, Stanford University, 2003.
[3] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report, Stanford University, 1999.

ACKNOWLEDGEMENTS

We thank Mr. Patrick Lau for his contributions to the experiments. This work is supported by grants from the Research Grants Councils of the HKSAR, China (CUHK4205/04E and CUHK4351/02E).