Weighted PageRank: cluster-related weights

Danil Nemirovsky[a,b] and Konstantin Avrachenkov[b]

[a] *St. Petersburg State University, Russia*

[b] *INRIA Sophia Antipolis, France*

danil.nemirovsky@gmail.com, K.Avrachenkov@sophia.inria.fr

Abstract

PageRank is a way to rank Web pages taking into account hyper-link structure of the Web. PageRank provides efficient and simple method to find out ranking of Web pages exploiting hyper-link structure of the Web. However, it produces just an approximation of the ranking since the random surfer model uses just uniform distributions for all situation of choice happening during the surf process. In particular, this implies that the random surfer has no preferences. The assumption is limited by its nature. Personalized PageRank was designed to solve the problem but it is still quite restrictive since it assumes non-uniform preferences just at jumping to arbitrary page on the Web and non-preferring behaviour when following outgoing hyper-links. Taking into account these limitations and restrictions of PageRank and Personalized PageRank we propose Weighted PageRank where we are free to weight hyper-links according any possible preferring behaviour of a user. In particular, cluster-related weights are considered.

Key words: PageRank, Weighted PageRank, Web graph

1 Introduction and Methodology

1.1 PageRank and Weighted PageRank

PageRank is a way to rank Web pages taking into account hyper-link structure of the Web [2]. The easiest way to imagine the "physical" meaning of PageRank is to consider random surfer model, or a surfer who explores the Web in a random way. The surfer having found herself at a page jumps with some probability to an arbitrary page on the Web choosing the page uniformly from the set of all pages on the Web or with complementary probability follows one of hyper-links of the page choosing the outgoing link uniformly from the set of all

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2008		**00-00-2008 to 00-00-2008**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Weighted PageRank: cluster-related weights	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **St. Petersburg State University,Russia,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Seventeenth Text REtrieval Conference (TREC 2008) held in Gaithersburg, Maryland, November 18-21, 2008. The conference was co-sponsored bythe National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT
see report

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**9**	

the outgoing hyper-links of the page. The random surfer model forms ergodic Markov chain having a unique stationary probability distribution. An entry of the stationary distribution corresponding to a Web page can be interpreted as probability to find the surfer at the page and, hence, can be considered as a kind of popularity or authority measure of the page. PageRank provides nice and simple method to find out ranking of Web pages exploiting hyper-link structure of the Web but it produces just an approximation of the ranking since the random surfer model uses uniform distributions for all situation of choice happening during surf process implying that the random surfer has no preferences. The assumption is limited by its nature. Personalized PageRank was called upon to solve the problem but it is still quite restricted since assumes preferring behaviour just at jumping arbitrary page on the Web and non-preferring behaviour during regular following outgoing hyper-links. Taking into account these limits and restrictions of PageRank and Personalized PageRank we propose Weighted PageRank where we are free to weight hyper-links according any possible preferring behaviour of an user. The most general definition of Weighted PageRank we can imagine is the following. Let us denote by n the number of pages on the Web. Let $\mathcal{P} = \{1, 2, \ldots, n\}$ be the set of all the Web pages. We use subscripts i, j for enumeration over the set of all the Web pages. Let assume that we have infinite but countable number of types of links. We denote the set of link types as T, where $T = \mathbb{N}$ actually. We choose weights for each link type and denote them by $\{\alpha_k\}_{k=1}^{\infty}$. We use subscript k to enumerate link types. Link type weights meet the following conditions:

$$\alpha_k \geqslant 0, \forall k \in T, \tag{1}$$

$$\sum_{k=1}^{\infty} \alpha_k = 1. \tag{2}$$

Let us denote by $\mathcal{P}_i$ the set of the outgoing links from page i, $i \in \mathcal{P}$. It is evident that $\mathcal{P}_i \subset \mathcal{P}$. Let us denote by n_i the number of the outgoing link from page i, $|\mathcal{P}_i| = n_i$. Let us denote by $\mathcal{P}_i^k$ the set of the outgoing links from page i of type $k \in T$. We note that a link can be polytypic and belong to several P_i^k. We assume here that a link is at least monotypic. We note that $\bigcup_{k=1}^{\infty} \mathcal{P}_i^k = \mathcal{P}_i$, and the most of $\mathcal{P}_i^k$ are empty. Let us denote by n_i^k the number of the outgoing link from page i of type $k \in T$, $|P_i^k| = n_i^k$. Let us define a map $t_i(j)$ of outgoing links of page i to their types.

$$t_i(j) : \mathcal{P}_i \to 2^T \backslash \{\emptyset\}.$$

$t_i(j)$ is a list of types of link (i, j). Let us denote by T_i the set of link types which outgoing links of page i have:

$$T_i = \cup_{j \in \mathcal{P}_i} t_i(j).$$

We note that $\sum_{k\in T_i} n_i^k \geqslant n_i$.

Now that we have defined a number of auxiliary notion, let us define weighted matrix W which plays for Weighted PageRank the same role as Google matrix for PageRank.

$$w_{ij} = \sum_{m\in t_i(j)} \frac{\alpha_m}{n_i^m} + \frac{1}{n_i} \sum_{k\notin T_i} \alpha_k, \quad j \in \mathcal{P}_i, \tag{3}$$
$$w_{ij} = 0, \quad j \notin \mathcal{P}_i.$$

A link of a page receives a weight proportional to the weight of its link type with a portion of weights of link types which are not presented at the page.

Let us define one more notion: a set of links marked by their types.

$$M_i = \{(j, m_i(j)) | j \in \mathcal{P}_i, m_i(j) \in t_i(j)\}. \tag{4}$$

Proposition 1 *If an original graph is strongly connected, which corresponds to ergodic Markov chain, then W is a stochastic matrix.*

Proof We shall proof that

$$\sum_{j=1}^{n} w_{ij} = 1, \quad \forall i \in \mathcal{P}.$$

Since the original graph is strongly connected there are not dangling pages.

Let us proof the statement for non-dangling page i, $\mathcal{P}_i \neq \emptyset$.

$$\sum_{j=1}^{n} w_{ij} = \sum_{j\in\mathcal{P}_i} w_{ij} + \sum_{j\notin\mathcal{P}_i} w_{ij}. \tag{5}$$

The second term is zero. Let us analyze the first term.

$$\sum_{j\in\mathcal{P}_i} w_{ij} = \sum_{j\in\mathcal{P}_i}\left(\sum_{m\in t_i(j)}\frac{\alpha_m}{n_i^m} + \frac{1}{n_i}\sum_{k\notin T_i}\alpha_k\right) = \tag{6a}$$

$$= \sum_{j\in\mathcal{P}_i}\sum_{m\in t_i(j)}\frac{\alpha_m}{n_i^m} + \sum_{k\notin T_i}\alpha_k = \tag{6b}$$

$$= \sum_{(j,m)\in M_i}\frac{\alpha_m}{n_i^m} + \sum_{k\notin T_i}\alpha_k = \tag{6c}$$

$$= \sum_{m\in T_i}\sum_{j\in\mathcal{P}_i^m}\frac{\alpha_m}{n_i^m} + \sum_{k\notin T_i}\alpha_k = \tag{6d}$$

$$= \sum_{m\in T_i}\alpha_m + \sum_{k\notin T_i}\alpha_k = 1 \tag{6e}$$

We substitute definition of weight matrix entry in (6a). We open brackets and sum the second term in (6b). Double summation is equivalent to summation over M_i set (6c). Now we can account elements of M_i by their type (6d). Summation all the terms in (6e) applying (2) finishes the proof. □

Weighted PageRank in its general definition can produce almost any ranking of the Web pages and actually possess too high liberty, therefore, it would be useful to consider special instantiations of Weighted PageRank with particular choice of link type weights.

Example 1 *PageRank is a particular case of Weighted PageRank.*

We have a graph. Let us assume that we have two types of links. Let us assume that all the links of the graph are of the first type and of the second type. Let us consider the complement the of graph to a fully connected graph. The links belonging to the complement we assume to be of the second type. Denote the weight of the first type by c, and the weight of the second type by $(1-c)$. Then transition matrix written according to definition is the following:

$$w_{ij} = \frac{c}{n_i^1} + \frac{c}{n_i^2}, \quad \mathcal{P}_i^1 \neq \emptyset, \mathcal{P}_i^2 \neq \emptyset, j \in \mathcal{P}_i^1, j \in \mathcal{P}_i^2,$$
$$w_{ij} = \frac{1-c}{n_i^2}, \quad \mathcal{P}_i^1 \neq \emptyset, \mathcal{P}_i^2 \neq \emptyset, j \in \mathcal{P}_i^2,$$
$$w_{ij} = \frac{1-c}{n_i^2} + \frac{c}{n_i}, \quad \mathcal{P}_i^1 = \emptyset, \mathcal{P}_i^2 \neq \emptyset, j \in \mathcal{P}_i^2.$$

After some simplification one can get:

$$
\begin{aligned}
w_{ij} &= \frac{c}{n_i^1} + \frac{c}{n_i}, \quad \mathcal{P}_i^1 \neq \emptyset, \mathcal{P}_i^2 \neq \emptyset, j \in \mathcal{P}_i^1, j \in \mathcal{P}_i^2, \\
w_{ij} &= \frac{1-c}{n_i}, \quad \mathcal{P}_i^1 \neq \emptyset, \mathcal{P}_i^2 \neq \emptyset, j \in \mathcal{P}_i^2, \\
w_{ij} &= \frac{1}{n}, \quad \mathcal{P}_i^1 = \emptyset, \mathcal{P}_i^2 \neq \emptyset, j \in \mathcal{P}_i^2.
\end{aligned}
$$

1.2 Weights according to a clustering

Let assume that we have a clustering $\mathcal{C}$ of all the pages on the Web. It does not matter how we get the clustering.

$$
\begin{aligned}
&\mathcal{C} = \{C_1, C_2, \ldots, C_N\}, \\
&\cup_{i=1}^{N} C_i = \mathcal{P}, \\
&C_i \cap C_j = \emptyset, \quad i \neq j, \ i, j = \overline{1, N}.
\end{aligned}
$$

Having the clustering we can select two types of links: links lying inside clusters, e.g. links between pages belonging to same cluster, such links are also called intra-links, and links crossing cluster borders, e.g. links between pages belonging to different clusters, such links are also called inter-links. Let α_k be weights for these two types of links:

$$
\begin{aligned}
&\alpha_k > 0, \quad k = 1, 2, \\
&\alpha_1 + \alpha_2 = 1.
\end{aligned}
$$

Let us define subindeces in the following way: $k, m = 1, 2, k \neq m$.

The Weighted transition matrix can be simplified in this case:

$$
\begin{aligned}
w_{ij} &= \frac{1}{n}, \quad \mathcal{P}_i = \emptyset, \\
w_{ij} &= \frac{\alpha_k}{n_i^k}, \quad j \in \mathcal{P}_i^k, \mathcal{P}_i^m \neq \emptyset, \\
w_{ij} &= \frac{\alpha_k}{n_i^k} + \frac{\alpha_m}{n_i}, \quad j \in \mathcal{P}_i^k, \mathcal{P}_i^m = \emptyset, \\
w_{ij} &= 0, \quad j \notin \mathcal{P}_i.
\end{aligned}
$$

2 Experiments

2.1 TOmUW

We have calculated usual cos measure between tf*idf vectors of documents and tf*idf vectors of topics. Both query and narration fields were used. For each query 1000 documents receiving highest values were reported. Non-zero value is obtained if a document and a topic has at least one common term.

2.2 4FvfI

We calculated PageRank for all the documents in the collection with damping factor equal to 0.85. For each query we took the documents having at least one common term with the query and order them according to their PageRank. We reported the first thousand documents having highest PageRank values and having at least one common term with the query.

2.3 Rkylv

We calculated Weighted PageRank for all the documents in the collection. For each query we took the documents having at least one common term with the query and order them according to their Weighted PageRank. We reported the first thousand documents having highest Weighted PageRank values and having at least one common term with the query.

The weights in the Weighted PageRank have been assigned according to a clustering. We clustered all documents using CDC clustering algorithm [1, 3] which works with text content of documents and intent to cluster documents according their topics. We gave different weight to links between documents belonging to the same cluster (intra-links) and to links between documents belonging to different clusters (inter-links). The weight of intra-links is equal to 0.15 while the weight of inter-links equal to 0.85. The weights were obtained by optimization by the number of relevant documents in the first thousand according to the topics and relevance judgements of TREC2007.

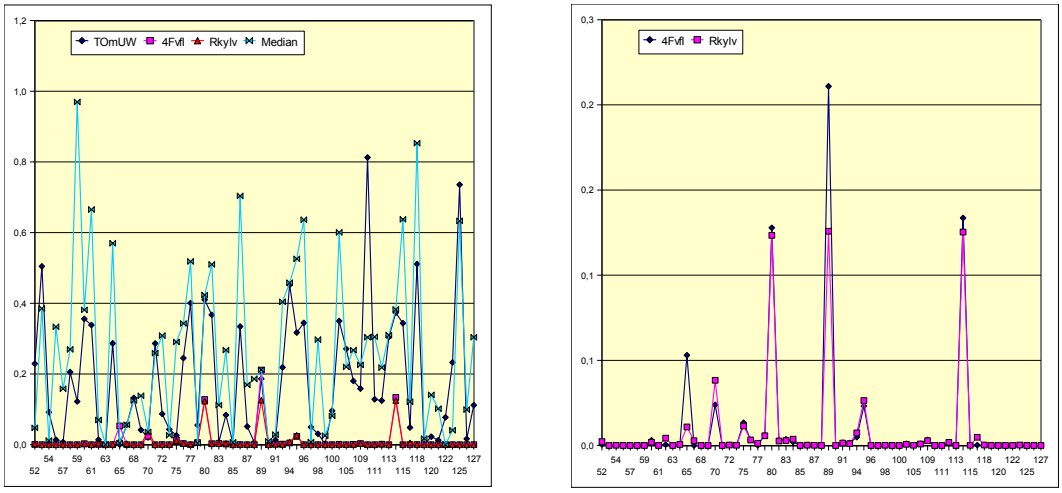

Fig. 1. AP measure on each topic with (left) and without (right) median results over all systems. Topics *id*s are placed at abscissa axis, and AP measure values are places at ordinate axis.

2.4 ycbLS

We have calculated usual cos measure between tf*idf vectors of documents and tf*idf vectors of topics. Only query field was used. For each query 1000 documents receiving highest values were reported. Non-zero value is obtained if a document and a topic has at least one common term.

3 Experimental results

The experimental results are preseted in Table 1. The experiments do not show an improvement of retrieval results by using Weighted PageRank with cluster related weigths, see experiment $Rkylv$, comparing to PageRank, experiment $4FvfI$.

	$TOmUW$	$4FvfI$	$Rkylv$	$ycbLS$	$Median$	$Best$
infAP	0.1803	0.0099	0.0082	0.1879	0.2670	0.5541
infNDCG	0.3852	0.0535	0.0584	0.3785	0.4679	0.7803

Table 1
Average measures over all topics. $Median$ is the average over of all the topics of the median measures of all the participated systems, and $Best$ is the average over of all the topics of the best achieved result among all the participated systems.

Experimental resuls at each topic are presented at Fig.1 and Fig.2. One can see from left graphs of Fig.1 and Fig.2 that in some cases using such classical distance measure between document and query as cosine of $tf * idf$ presentations

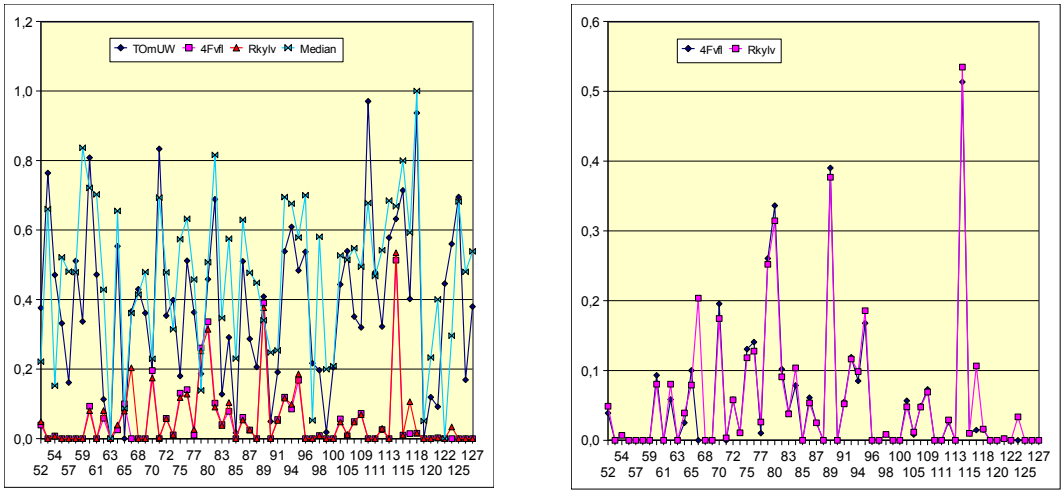

Fig. 2. NDCG measure on each topic with (left) and without (right) median results over all systems. Topics *id*s are placed at abscissa axis, and NDCG measure values are places at ordinate axis.

($TOmUW$) gives better results than achieved by all systems at median.

We cannot conclude from right graphs of Fig.1 and Fig.2 that ordering of relevant documents according to Weighted PageRank gives better performance than ordering by PageRank and values in Table 1 of AP measure and NDCG measure support this, but AP measure for Weighted PageRank is greater than one for PageRank in 48 cases from 63, which gives about 76%, and NDCG measure - in 42 cases from 63, which gives about 67%. We performed Kolmogorov-Smirnov test to check if the distribution of AP measures for PageRank and Weighted PageRank (or NDCG measures for PageRank and Weighted PageRank) follow the same distribution (the null hypothesis) and got that we cannot reject the null hypothesis up to significance level equal to 0.9865 for AP measure (or 0.9293 for NDCG measure), which makes us to conclude that AP measure of Weighted PageRank is greater than AP measure for PageRank in majority of cases is an arbitrary result. One can say the same about NDCG measure.

Failure of link-base ranking as PageRank and Weighted PageRank in the experiments can be explained by lack of recomendational links and prevalence of navigational links in CSIRO dataset since, being centrally ruled organization, CSIRO attends to support the very few number of related research project which makes it very differ from the Web where PageRank is used with success.

References

[1] Vladimir Dobrynin, David W. Patterson, and Niall Rooney. Contextual document clustering. In Sharon McDonald and John Tait, editors, *ECIR*, volume 2997 of *Lecture Notes in Computer Science*, pages 167–180. Springer, 2004.

[2] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia, 1998.

[3] Niall Rooney, David W. Patterson, Mykola Galushka, and Vladimir Dobrynin. A scaleable document clustering approach for large document corpora. *Inf. Process. Manage.*, 42(5):1163–1175, 2006.