PageSim: A Novel Link-based Measure of Web Page Similarity

Zhenjiang Lin, Michael R. Lyu, and Irwin King

Department of Computer Science and Engineering
The Chinese University of Hong Kong, Shatin, NT, Hong Kong
{zjlin, lyu, king }@cse.cuhk.edu.hk

ABSTRACT

To find similar web pages to a query page on the Web, this paper introduces a novel link-based similarity measure, called *PageSim*. Contrast to SimRank, a recursive refinement of cocitation, PageSim can measure similarity between *any* two web pages, whereas SimRank cannot in some cases. We give some intuitions to the PageSim model, and outline the model with mathematical definitions. Finally, we give an example to illustrate its effectiveness.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval − *clustering, search process, information filtering*; G2.2 [Discrete Mathematics]: Graph Theory − *graph algorithms*

General Terms: Algorithms, Measurement

Keywords: similarity measure, link analysis, search engine, PageRank, SimRank

1. INTRODUCTION

Finding similar web pages to a query page is a crucial task for a search engine. Recently, a variety of link-based similarity measures, which use only the hyperlinks in the Web, have been proposed for this task. This includes companion algorithm [3], cocitation algorithm [3], and SimRank [4], etc.

In this paper, we propose a novel link-based similarity measure, called PageSim. Contrast to SimRank, our method can measure similarity between any two web pages, whereas SimRank cannot in some cases.

SimRank is a fixed point of the recursive definition: *two pages are similar if they are linked to by similar pages*. Numerically, for any web page u and v, this is specified by defining $simrank(u,u) = 1$ and

$$simrank(u,v) = \gamma \cdot \frac{\sum_{a \in I(u)} \sum_{b \in I(v)} simrank(a,b)}{|I(u)||I(v)|} \quad (1)$$

for $u \neq v$ and $\gamma \in (0,1)$, where $I(x)$ denotes the set of inlink pages of x, $|I(x)|$ denotes the cardinality of the set. If $I(u)$ or $I(v)$ is empty, then $simrank(u,v)$ is zero by definition. The SimRank iteration starts with $simrank_0(u,v) = 1$ for $u = v$ and $simrank_0(u,v) = 0$ for $u \neq v$. The *SimRank score* between u and v is defined as $lim_{k\to\infty} simrank_k(u,v)$.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-332-9/06/0005.

Unfortunately, the result of SimRank is not convincing in some cases. In one case, if one of two web pages has no inlink, then the SimRank score of them is zero by definition, which means they are not similar. However, this is not always true. For example, in Figure 1 of section 4, v_1 has no inlink, but it is clear that both v_2 and v_3 have some similarity with it for they are linked to by v_1 . In another case (also in Figure 1), SimRank concludes that v_2 and v_4 are not similar. In fact, obviously v_2 and v_4 indeed have some similarity, for they *link to each other*. More detailed illustration is given in the last part of this paper.

2. PAGESIM

PageSim can be considered as an extension of cocitation algorithm, in which the similarity score between two web pages is defined by the number of inlink neighbors that they have in common. Actually, on the Web, not all links are equally important. For example, if the only common neighbor of page a and b is the Yahoo home page [1], whereas page a and c have several common neighbors from obscure places, then which page is more similar to page a, page b or page c? As we know, hyperlink from web page u to v can be considered as a recommendation of page v by page u [2], and the more important a web page is, the more important its recommendation is. Evidently, the reasonable answer should be page b, since the Yahoo home page is much more "important". In another perspective, the action of recommendation can be considered that page u *propagates* some kind of similarity to page v, and the more pages it links to, the less similarity it should propagate to each of these pages. Therefore, it is also reasonable to think that the Yahoo home page has some kind of similarity with both page a and page b.

Since PageRank [5] is one of the most prominent ranking algorithm which assigns global ranking scores to all pages on the Web, we take the PageRank score of a web page as the *importance* (*weight* or *similarity score*) of it in the PageSim method. The intuitions to PageSim model is described as follows, and the mathematical definitions will be given later.

At the beginning, each web page only contains its own similarity score, and then each web page propagates its own similarity score to its outlink neighbors, receiving and propagating the similarity scores of others at the same time. After the propagation, each page contains its own similarity score as well as the similarity scores of others. These scores are stored in a vector called the *similarity vector* of this page. Then we can calculate the **PageSim score** of each pair of pages by *summing their common similarity scores up*.

3. DEFINITIONS

We model the Web as a directed graph $G = (V, E)$ with vertices V representing web pages $v_i (i = 1, 2, \cdots, n = |E|)$ and directed edges E representing hyperlinks between web pages. Let $I(v)$ and $O(v)$ denote the set of *inlink* pages and *outlink* pages of v respectively, for any $v \in V$. Let $path(u_1, u_s)$ denotes a sequence of vertices $u_1, u_2, \ldots, u_s$ such that $(u_i, u_{i+1}) \in E$ $(i = 1, \cdots, s-1)$ and u_i are distinct, it is called a **path** from u to v. Let $length(p)$ denotes the **length** of path p, define $length(p) = |p| - 1$. Let $PATH(u, v)$ denotes the set of all possible paths from page u to v.

DEFINITION 1. *Let $PR(v)$ denotes the PageRank score of page v, for $v \in V$. Let $PG(u, v)$ denotes the PageRank score that page u propagates to page v through $PATH(u, v)$, i.e., $PG(u, v) = \sum_{p \in PATH(u,v)} \frac{PR(u)}{\prod_{w \in p, w \neq v} |O(w)|}$, where $u, v \in V$.*

DEFINITION 2. *Let $\overrightarrow{SV}(v)$ denotes the **similarity vector** of page v, we have $\overrightarrow{SV}(v) = (PG(v_i, v))^T, i = 1, \cdots, n$, where $v, v_i \in V$. Let $PS(u, v)$ denotes the **PageSim score** of page u and v, $PS(u, v) = \sum_{i=1}^{n} min(PG(v_i, u), PG(v_i, v))$, where $u, v \in V$.*

4. PAGESIM *VS* SIMRANK

A good evaluation of PageSim is difficult without performing extensive user studies or having a reliable external measure of similarity to compare against. In this section, we give a simple example in which PageSim is compared with SimRank to illustrate the performance of PageSim.

For a given graph $G(V, E)$, where $V = \{v_i\} (i = 1, \cdots, 6)$ (see Figure 1). Let $\overrightarrow{PR}(V) = (PR(v_i))^T, i = 1, \cdots, 6$. We have $\overrightarrow{PR}(V) = (0.08, 0.23, 0.18, 0.14, 0.14, 0.23)^T$.

The *PageSim score matrix* is

$$\begin{pmatrix} 0.08 & 0.04 & 0.05 & 0.01 & 0.01 & 0.05 \\ 0.04 & 0.41 & 0.16 & 0.23 & 0.14 & 0.16 \\ 0.05 & 0.16 & 0.35 & 0.14 & 0.14 & 0.35 \\ 0.01 & 0.23 & 0.14 & 0.23 & 0.14 & 0.14 \\ 0.01 & 0.14 & 0.14 & 0.14 & 0.28 & 0.14 \\ 0.05 & 0.16 & 0.35 & 0.14 & 0.14 & 0.58 \end{pmatrix}.$$

Let $top(v, t)$ denotes the top t similar pages to page v (excluding v). Let $t = 2$, we have

$$\begin{array}{ll} top(v_1, 2) = \{v_3, v_6\}, & top(v_2, 2) = \{v_4, v_{3,6}\}, \\ top(v_3, 2) = \{v_6, v_2\}, & top(v_4, 2) = \{v_2, v_{3,5,6}\}, \\ top(v_5, 2) = \{v_{2,3,4,6}, v_1\}, & top(v_6, 2) = \{v_3, v_2\}. \end{array}$$

The *SimRank score matrix* of graph G is

$$\begin{pmatrix} 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.25 & 0.00 & 0.00 & 0.25 \\ 0.00 & 0.25 & 1.00 & 0.50 & 0.50 & 0.13 \\ 0.00 & 0.00 & 0.50 & 1.00 & 1.00 & 0.25 \\ 0.00 & 0.00 & 0.50 & 1.00 & 1.00 & 0.25 \\ 0.00 & 0.25 & 0.13 & 0.25 & 0.25 & 1.00 \end{pmatrix}.$$

Thus, we have

$$\begin{array}{ll} top(v_1, 2) = \{\}, & top(v_2, 2) = \{v_3, v_6\}, \\ top(v_3, 2) = \{v_{4,5}, v_2\}, & top(v_4, 2) = \{v_5, v_3\}, \\ top(v_5, 2) = \{v_4, v_3\}, & top(v_6, 2) = \{v_{2,4,5}, v_3\}. \end{array}$$

We can see that the results of PageSim and SimRank are different. First, SimRank shows that there's no page similar to v_1. While PageSim shows that v_3 is most similar to v_1, which is more reasonable. Because the fact that v_1 links to v_3 implies v_1 "considers" v_3 has some level of similarity with it. Secondly, SimRank shows v_4 is not similar to v_2, while PageSim shows that is not true. Obviously, v_2 and v_4 are similar, for they *link to each other*. Moreover, PageSim considers that v_4 is most similar to v_2. SimRank shows v_3 is most similar to v_2, for they have a common inlink page v_1. We believe PageSim is the winner in this situation because the "link to each other" relationship really implies stronger similarity than that of the "common inlink" relationship.

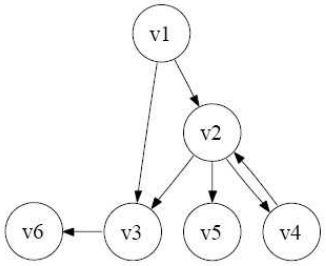

Figure 1: graph G

5. CONCLUSION AND FUTURE WORK

This paper introduces PageSim, a novel link-based similarity measure. Based on the strategy of *PageRank score propagation*, PageSim is capable of measuring similarity between any two web pages.

There are numbers of avenues for future work. Foremost, we must address the efficiency issue. For example, the computing time of PageSim is expected to be greatly reduced by limiting the radius of propagation, i.e., the path length of propagation. Especially, when the radius is reduced to 1, PageSim becomes a "*weighted* cocitation algorithm". Finally, extensive evaluations of PageSim are needed.

6. ACKNOWLEDGMENT

This work is supported by grants from the Research Grants Councils of the HKSAR, China (Project No. CUHK4205/04E and Project No. CUHK4235/04E) and is affiliated with the VIEW Technologies Laboratory and the Microsoft-CUHK Joint Laboratory for Human-centric Computing & Interface Technologies.

7. REFERENCES

[1] http://www.yahoo.com.

[2] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the web. *ACM Transactions on Internet Technology*, 1(1):2–43, 2001.

[3] J. Dean and M. R. Henzinger. Finding related pages in the World Wide Web. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1467–1479, 1999.

[4] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *KDD '02: Proceedings of the eighth ACM SIGKDD*, pages 538–543, New York, NY, USA, 2002. ACM Press.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.