An Improved Computation of the PageRank Algorithm

Sung Jin Kim and Sang Ho Lee
School of Computing, Soongsil University, Korea
lace@nowuri.net, shlee@computing.ssu.ac.kr
http://orion.soongsil.ac.kr/

Abstract. The Google search site (http://www.google.com) exploits the link structure of the Web to measure the relative importance of Web pages. The ranking method implemented in Google is called PageRank [3]. The sum of all PageRank values should be one. However, we notice that the sum becomes less than one in some cases. We present an improved PageRank algorithm that computes the PageRank values of the Web pages correctly. Our algorithm works out well in any situations, and the sum of all PageRank values is always maintained to be one. We also present implementation issues of the improved algorithm. Experimental evaluation is carried out and the results are also discussed.

1. Introduction

Web information retrieval tools typically make use of the text on the Web pages as well as the links of the Web pages that contain valuable information implicitly. The link structure of the Web represents a considerable amount of latent human annotation, and thus offers a starting point for structural studies of the Web. Recent work in the Web search area has recognized that the hyperlink structure of the Web is very valuable for locating information [1, 2, 3, 4, 5, 9, 13, 14, 15].

The Google search site (http://www.google.com), which emerged in 1998, exploits the link structure of the Web to measure the relative importance of Web pages. The ranking method implemented in Google is called PageRank [3]. PageRank is an objective measure of citation importance that corresponds with people’s subjective idea of importance. Pages that are well cited from many places are worth looking at. PageRank postulates that a link from page u to v implies the author of u recommends to take a look at page v. A page has a high PageRank value if there are many pages that point to it, or if there are pages that point to it and have a high PageRank value. It is known that PageRank helps to rank pages effectively in the Google site.

The PageRank algorithm and implementation details are described in [7, 12]. The PageRank algorithm represents the structure of the Web as a matrix, and PageRank values as a vector. The PageRank vector is derived by computing matrix-vector multiplications repeatedly. The sum of all PageRank values should be one during the computation. However, we learned that the sum becomes less than one as the computation process continues in some cases. In those cases, all PageRank values become smaller than they should be.

1 This work was supported by grant No. (R-01-2000-00403) from the Korea Science and Engineering Foundation.

In this paper, we present an improved PageRank algorithm that computes the PageRank values of the Web pages correctly. Our algorithm works out well in any situations, and the sum of all PageRank values is always maintained to be one. We also present implementation issues of the improved algorithm. Experimental evaluation is carried out and the results are also discussed.

This paper is organized as follows. The PageRank algorithm is presented in section 2. Section 3 identifies drawbacks of the original PageRank algorithm [7] and presents an improved PageRank algorithm. In section 4, we discuss experimental evaluation of algorithms and its results. Section 5 contains closing remarks.

2. PageRank Computation

Let \( v \) be a Web page, \( F_v \) be the set of pages \( v \) points to, and \( B_v \) be the set of pages that point to \( v \). Let \( N_v = |F_v| \) be the number of links from \( v \). The PageRank (PR) equation [12] for \( v \) is recursively defined as:

\[
PR(v) = \sum_{u \in B_v} \frac{PR(u)}{N_v}.
\]  

The pages and hyperlinks of the Web can be viewed as nodes and edges in a directed graph [10]. Let \( M \) be a square matrix with the rows and columns corresponding to the directed graph \( G \) of the Web, assuming all nodes in \( G \) have at least one outgoing edge. If there is a link from page \( j \) to page \( i \), then the matrix entry \( m_{ij} \) has a value \( \frac{1}{N_j} \). The values of all other entries are zero. PageRank values of all pages are represented as an \( N \times 1 \) matrix (a vector), \( \text{Rank} \). The \( i \)th entry, \( \text{rank}(i) \), in \( \text{Rank} \) represents the PageRank value of page \( i \).

![Fig. 1. A small Web, its matrix, and its PageRank values](image)

Fig. 1 shows a simple example of \( M \) and \( \text{Rank} \). The rectangular shape like a document denotes a page. A page identifier appears above each page. The small rectangle represents a URL in a page. The directed line denotes a link from one page to another. For an instance, page 5 has two outgoing edges to page 3 and 4 (\( N_v = 2 \)), \( m_{35} \), and \( m_{45} \) of \( M \) are \( \frac{1}{2} \), and \( m_{15} \), \( m_{25} \), and \( m_{55} \) are 0. Page 5 is pointed by page 2 and 3, so its PageRank value is determined by PageRank values of page 2 and 3. Since page 2 and 3 have four links and one link respectively, the PageRank of page 5, \( \text{rank}(5) \), is the sum of a fourth of \( \text{rank}(2) \) and \( \text{rank}(3) \). Such computation corresponds to the matrix-vector multiplication.