The PageRank Citation Ranking: Bringing Order to the Web
Introduction

Paper: The PageRank Citation Ranking: Bringing Order to the Web, 1999

Authors: Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd

Page and Brin were MS students at Stanford
They founded Google in September, 98.

Most of this presentation is based on the original paper (link)
The Initiative's focus is to dramatically advance the means to collect, store, and organize information in digital forms, and make it available for searching, retrieval, and processing via communication networks -- all in user-friendly ways.

Stanford WebBase project
Pagerank
Motivation

- Web is vastly large and heterogeneous
  - Original paper's estimation were over 150 M pages and 1.7 billion of links
- Pages are extremely diverse
  - Ranging from "What does the fox say?" to journals about IR
- Web Page present some "structure"
  - Pagerank takes advantage of links structure
Pagerank
Motivation

- Inspiration: Academic citation
- Papers
  - are well defined units of work
  - are roughly similar in quality
  - are used to extend the body of knowledge
  - can have their "quality" measured in number of citations
Pagerank
Motivation

- Web pages, on the other hand
  - proliferate free of quality control or publishing costs
  - huge numbers of pages can be created easily
    ▶ artificially inflating citation counts
  - They vary on much wider scale than academic papers in quality, usage, citations and length
Pagerank

Motivation

A random archived message posting asking an obscure question about an IBM computer is very different from the IBM home page.

A research article about the effects of cellphone use on driver attention is very different from an advertisement for a particular cellular provider.
The average web page quality experienced by a user is higher than the quality of the average web page. This is because the simplicity of creating and publishing web pages results in a large fraction of low quality web pages that users are unlikely to read.
Pagerank

Idea

- Creates a graph based on link structure
  - Pages are nodes
  - Links are edges
  - Forward links are outedges
  - Backlinks are inedges
- A and B are backlinks of C
Pagerank
Assumptions

- Link from page A to page B is a vote from A to B
- Highly linked pages are more “important” than pages with few links
- Backlinks from high PR-pages count more than links from low PR-pages
- Combination of PR and text-matching techniques result in highly relevant search results
Pagerank
Pagerank

Definition

Simplification of Pagerank

A simple ranking function

- $u$ is a web page
- $F_u$ is a set of pages that $u$ points
- $B_u$ is the set of pages pointing to $u$
- $c$ is a normalization factor
- $N_u = |F_u|$

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$
Pagerank
Definition

- Rank of a page is divided among its forward links
- Equation is recursive
- May be computed by starting with any set of ranks
  - it iterates until it converges.
Problem with previous equation:

- Consider two web pages that point to each other
  - but to no other page.
- Suppose there is some web page which points to one of them.
- During iteration, this loop will accumulate rank but never distribute any rank
  - Since there are no outedges.
Pagerank

Definition

To overcome the problem:

\[ R'(u) = c \sum_{v \in B_u} \frac{R'(v)}{N_v} + cE(u) \]

where \( E(u) \) is some vector over the web pages.

Based on a random surfer model.
Finally, Pagerank is usually defined as:

\[
PR(u) = 1 - d + d \sum_{v \in B_v} \frac{PR(v)}{N(v)}
\]
Finally, Pagerank is usually defined as:

\[ PR(u) = 1 - d + d \sum_{v \in B_v} \frac{PR(v)}{N(v)} \]

represents the change to get to page \( u \) from any other page (random walk)
Pagerank
Definition

- Finally, Pagerank is usually defined as:

\[ PR(u) = 1 - d + d \sum_{v \in B_u} \frac{PR(v)}{N(v)} \]

represents the change to get to \( u \) from pages that points to \( u \), \( B_u \).
Pagerank Example

Complete iteration process can be found here (made by Alberto Ueda).
Pagerank
Convergence

Pretty quick and robust!
Pagerank
Paper Implementation

- Repository size: 24M web pages (over 75M unique URLs)
- computing PR of entire repository takes ~5h
- Issues:
  - Volume
  - incorrect HTML
  - dynamics of the web, page exclusion (robots.txt)
Pagerank

Usage

- Search
  - combination of retrieve models and pagerank for ranking
- Commercial Interests
  - It is not easily manipulated
- Estimation of Web Traffic
  - Corresponds to a random web surfer
- Backlink predictor for crawling
Pagerank
Unwanted usage

- Bmw.de banned from Google in early 2016
  - due to doorway page (link)
- Google bomb
  - President article (2007)
  - Repub & Dem article (2017)