Impact of Search Engines on Page Popularity

Junghoo “John” Cho (cho@cs.ucla.edu)
Sourashis Roy (roys@cs.ucla.edu)

University of California, Los Angeles
Motivation

“If you are not indexed by Google, you do not exist on the Web”

– News.com article, 10/23/2002

• People “discover” pages through search engines
  • Top results: many users
  • Bottom results: no new users

• Are we biased by search engines?
Outline

- Are the rich getting richer?
  - Web popularity-evolution experiment
- How much bias do search engines introduce?
  - Impact of search engines
Web Evolution Experiment

- Collect Web history data
  - Is “rich-get-richer” happening?
- 154 sites monitored
  - Top sites from each category of Open Directory
- Pages downloaded every week
  - All pages in each site
  - A total of average 4M pages every week (65GB)
“Rich-Get-Richer” Problem

- Construct weekly Web-link graph
  - From the downloaded data
- Partition pages into 10 groups
  - Based on initial link popularity
  - Top 10% group, 10%-20% group, etc.
- How many new links to each group after a month?
  - Rich-get-richer $\rightarrow$ More new links to top groups
After 7 months
- 70% of new links to top 20% pages
- No new links to bottom 60% pages
After 7 months
- Decrease in PageRank for bottom 50% pages
- Due to normalization of PageRank
Outline

- Web popularity-evolution experiment
  - “Rich-get-richer” is indeed happening
  - Unpopular pages get no attention
- Impact of search engines
  - How much bias do search engines introduce?
Outline

- Web popularity-evolution experiment
  - “Rich-get-richer” is indeed happening
  - Unpopular pages get no attention

- Impact of search engines
  - How much bias do search engines introduce?
What we mean by bias?
Search Engine Bias

What we mean by bias?

What is the ideal ranking?
How do search engines rank pages?
What is the Ideal Ranking?

Rank by intrinsic “quality” of a page?
What is the Ideal Ranking?

Rank by intrinsic “quality” of a page?

- Very subjective notion
- Different quality judgment on the same page
- Can there be an “objective” definition?
Definition

The probability that an average Web user will like page \( p \) enough to create a link to it if he looks at it

In principle, we can measure \( Q(p) \) by

1. showing \( p \) to all Web users and
2. counting how many people like it

\( p_1 \): 10,000 people, 8,000 liked it, \( Q(p_1) = 0.8 \)
\( p_2 \): 10,000 people, 2,000 liked it, \( Q(p_2) = 0.2 \)
The probability that an average Web user will like page \( p \) enough to create a link to it if he looks at it.

**Definition**

In principle, we can measure \( Q(p) \) by:

1. showing \( p \) to all Web users and
2. counting how many people like it

- \( p_1 \): 10,000 people, 8,000 liked it, \( Q(p_1) = 0.8 \)
- \( p_2 \): 10,000 people, 2,000 liked it, \( Q(p_2) = 0.2 \)
Page Quality $Q(p)$

**Definition**

The probability that an average Web user will like page $p$ enough to create a link to it if he looks at it.

- In principle, we can measure $Q(p)$ by:
  1. showing $p$ to all Web users and
  2. counting how many people like it

  - $p_1$: 10,000 people, 8,000 liked it, $Q(p_1) = 0.8$
  - $p_2$: 10,000 people, 2,000 liked it, $Q(p_2) = 0.2$

- Democratic measure of quality
  - When consensus is hard to reach, pick the one that more people like
PageRank: Intuition

- A page is “important” if many pages link to it
- Not every link is equal
  - A link from an “important” page matters more than others
    e.g. Link from Yahoo vs Link from a random home page

\[ PR(p_i) = (1 - d) + d \left[ \frac{PR(p_1)}{c_1} + \cdots + \frac{PR(p_m)}{c_m} \right] \]

Random-Surfer Model

When users follow links randomly, \( PR(p_i) \) is the probability to reach \( p_i \)
Page Quality vs PageRank

- PageRank \( \approx \) Page quality if everyone is given equal chance
- High PageRank \( \rightarrow \) high quality
  - To obtain high PageRank, many people should look at the page and like it.
- Low PageRank \( \rightarrow \) low quality?
  - PageRank is biased against new pages
- How much bias for low PageRank pages?
Ideal experiment:
- Divide the world into two groups
  - The users who do not use search engines
  - The users who use search engines very heavily
- Compare popularity evolution
Ideal experiment:
- Divide the world into two groups
  - The users who do not use search engines
  - The users who use search engines very heavily
- Compare popularity evolution

Problem: Difficult to conduct in practice
Let us do theoretical experiments!

- **Random-surfer model**
  - Users follow links randomly
  - Never use search engines

- **Search-dominant model**
  - Users always start with a search engine
  - Only visit pages returned by the search engine

→ Compare popularity evolution
Basic Definitions for the Models

(Simple) Popularity $P(p, t)$
- Fraction of Web users that like $p$ at time $t$
- E.g., 100,000 users, 10,000 like $p$, $P(p, t) = 0.1$

Visit Popularity $V(p, t)$
- Number of users that visit $p$ in a unit time

Awareness $A(p, t)$
- Fraction of Web users who are aware of $p$
- E.g., 100,000 users, 30,000 aware of $p$, $A(p, t) = 0.3$
Basic Definitions for the Models

(Simple) Popularity $\mathcal{P}(p, t)$
- Fraction of Web users that like $p$ at time $t$
- E.g., 100,000 users, 10,000 like $p$, $\mathcal{P}(p, t) = 0.1$

Visit Popularity $\mathcal{V}(p, t)$
- Number of users that visit $p$ in a unit time

Awareness $\mathcal{A}(p, t)$
- Fraction of Web users who are aware of $p$
- E.g., 100,000 users, 30,000 aware of $p$, $\mathcal{A}(p, t) = 0.3$

$\mathcal{P}(p, t) = Q(p) \cdot \mathcal{A}(p, t)$
Random-Surfer Model

Popularity-Equivalence Hypothesis

\[ V(p, t) = r \cdot P(p, t) \quad \text{(or} \quad V(p, t) \propto P(p, t)) \]

- PageRank is visit probability under random-surfer model
- Higher popularity → More visitors

Random-Visit Hypothesis

A visit is done by any user with equal probability
Random-Surfer Model: Analysis

Current popularity $\mathcal{P}(p, t)$
→ Number of visitors from $\mathcal{V}(p, t) = r \cdot \mathcal{P}(p, t)$
→ Awareness increase $\Delta A(p, t)$
→ Popularity increase $\Delta \mathcal{P}(p, t)$
→ New popularity $\mathcal{P}(p, t + 1)$
Random-Surfer Model: Analysis

Current popularity $\mathcal{P}(p, t)$
→ Number of visitors from $\mathcal{V}(p, t) = r \cdot \mathcal{P}(p, t)$
→ Awareness increase $\Delta A(p, t)$
→ Popularity increase $\Delta \mathcal{P}(p, t)$
→ New popularity $\mathcal{P}(p, t + 1)$

Formal Analysis: Differential Equation

$$\mathcal{P}(p, t) = \left[1 - e^{-\frac{r}{n} \int_0^t \mathcal{P}(p,t) dt}\right] Q(p)$$
Random-Surfer Model: Result

Theorem

The popularity of page \( p \) evolves over time through the following formula:

\[
P(p, t) = \frac{Q(p)}{1 + \left[ \frac{Q(p)}{P(p, 0)} - 1 \right] e^{-\left[ \frac{r}{n} Q(p) \right] t}}
\]

- \( Q(p) \): quality of \( p \)
- \( P(p, 0) \): initial popularity of \( p \) at time zero
- \( n \): total number of Web users.
- \( r \): normalization constant in \( V(p, t) = r \cdot P(p, t) \)
Impact of Search Engines on Page Popularity

Random-Surfer Model: Popularity Graph

\[ Q(p) = 1, \quad P(p, 0) = 10^{-8}, \quad \frac{r}{n} = 1 \]
Search-Dominant Model

\[ V(p, t) \sim P(p, t) \]
Search-Dominant Model

\( V(p, t) \sim P(p, t) \)?

- For \( i \)th result, how many clicks?
- For PageRank \( P(p, t) \), what ranking?

Empirical measurement by Lempel et al. and us

New Visit-Popularity Hypothesis

\( V(p, t) = r \cdot P(p, t) \)

Random-Visit Hypothesis

A visit is done by any user with equal probability

Impact of Search Engines on Page Popularity

J. Cho, S. Roy
\( \mathcal{V}(p, t) \sim \mathcal{P}(p, t) ? \)

- For \( i \)th result, how many clicks?
- For PageRank \( \mathcal{P}(p, t) \), what ranking?
- Empirical measurement by Lempel et al. and us
Search-Dominant Model

\( \mathcal{V}(p, t) \sim \mathcal{P}(p, t) \)?

- For \( i \)th result, how many clicks?
- For PageRank \( \mathcal{P}(p, t) \), what ranking?
- Empirical measurement by Lempel et al. and us

New Visit-Popularity Hypothesis

\[
\mathcal{V}(p, t) = r \cdot \mathcal{P}(p, t)^{\frac{9}{4}}
\]
Search-Dominant Model

\[ \mathcal{V}(p, t) \sim \mathcal{P}(p, t)? \]
- For \( i \)th result, how many clicks?
- For PageRank \( \mathcal{P}(p, t) \), what ranking?
- Empirical measurement by Lempel et al. and us

New Visit-Popularity Hypothesis

\[ \mathcal{V}(p, t) = r \cdot \mathcal{P}(p, t)^{\frac{9}{4}} \]

Random-Visit Hypothesis

A visit is done by any user with equal probability
Impact of Search Engines on Page Popularity

J. Cho, S. Roy

\[
\sum_{i=1}^{\infty} \frac{[\mathcal{P}(p, t)](i - \frac{9}{4}) - [\mathcal{P}(p, 0)](i - \frac{9}{4})}{(i - \frac{9}{4}) Q(p)^i} = \frac{r}{n} t \quad \text{(same parameters as before)}
\]
Comparison of Two Models

- **Time to final popularity**
  - Random surfer: 25 time units
  - Search dominant: 1650 time units
  - → 66 times increases!

- **Expansion stage**
  - Random surfer: 12 time units
  - Search dominant: non existent
Web popularity-evolution experiment
  “Rich-get-richer” is indeed happening

Impact of search engines
  Search engines have worrisome impact but can also be very helpful

New ranking metric avoiding the bias?
  Page Quality: In Search of an Unbiased Web Ranking
  UCLA CS Department, Nov. 2003.
Thank You

Any Questions?
Popularity Increase: Relative Link Count

Relative increase in number of in-inks

Impact of Search Engines on Page Popularity

J. Cho, S. Roy
Popularity Increase: Relative PageRank

Relative increase in PageRank

Impact of Search Engines on Page Popularity

J. Cho, S. Roy
Search-Dominant Model: Result

Impact of Search Engines on Page Popularity

J. Cho, S. Roy