Search Engines Considered Harmful
In Search of an Unbiased Web Ranking

Junghoo "John" Cho
cho@cs.ucla.edu

UCLA

World-Wide Web

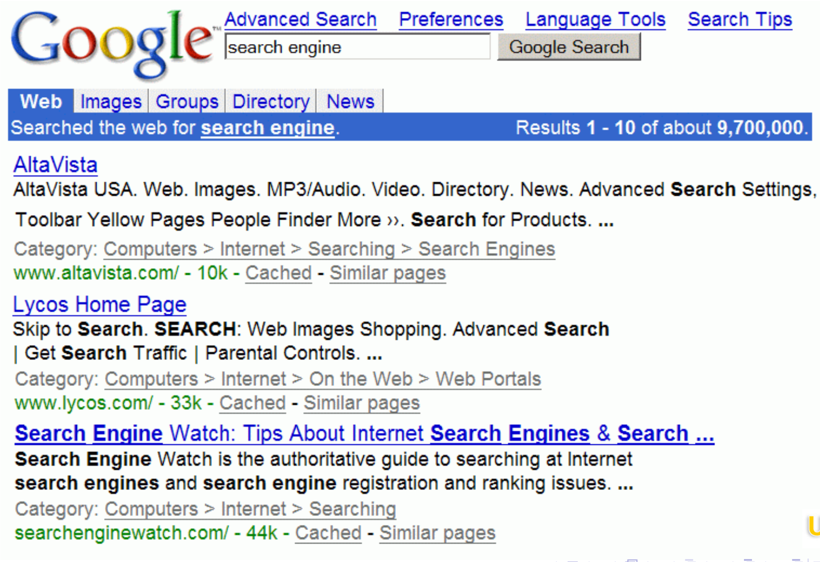
10 years ago

With Web

Information Overload

- Too much information, too much junk
- Too little time

Search Engines: The Savior



Search Engine Success: Flip Side

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

– News.com article, 10/23/2002

- Only a few major players
  - 75% market share by Google alone
- People "discover" pages through search engines
  - Top results: many users
  - Bottom results: no new users

Search Engine Success: Flip Side

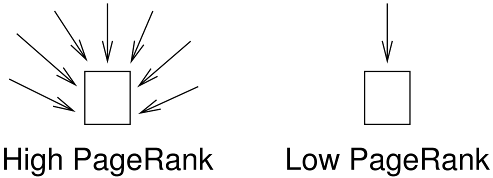
"If you are not indexed by Google, you do not exist on the Web"

– News.com article, 10/23/2002

- Only a few major players
  - 75% market share by Google alone
- People "discover" pages through search engines
  - Top results: many users
  - Bottom results: no new users
- Big question: Are we biased by search engines?

PageRank: "Secret Ranking Recipe"

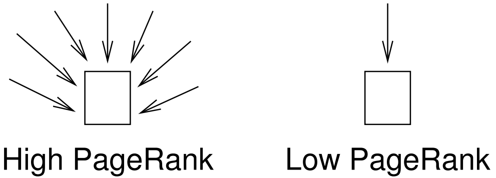
- Intuition: You are "important" if many other pages link to you



  - Popular pages are returned at the top
  - More details later...

PageRank: "Secret Ranking Recipe"

- Intuition: You are "important" if many other pages link to you



  - Popular pages are returned at the top
  - More details later...

- "Rich-get-richer" problem?

Outline

- Web popularity-evolution experiment
  - Is “rich-get-richer” happening?
- Impact of search engines
  - How much bias do search engines introduce?
- New ranking metric
  - Can we avoid search-engine bias?

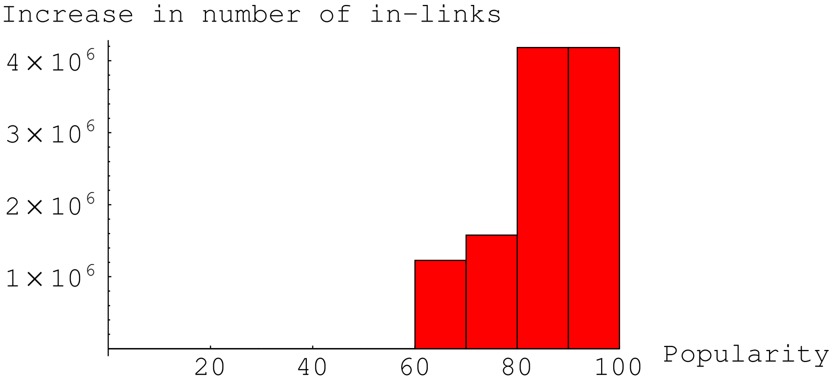
Web Evolution Experiment

- Collect Web history data
  - Is "rich-get-richer" happening?
- From Oct. 2002 until Oct. 2003
- 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

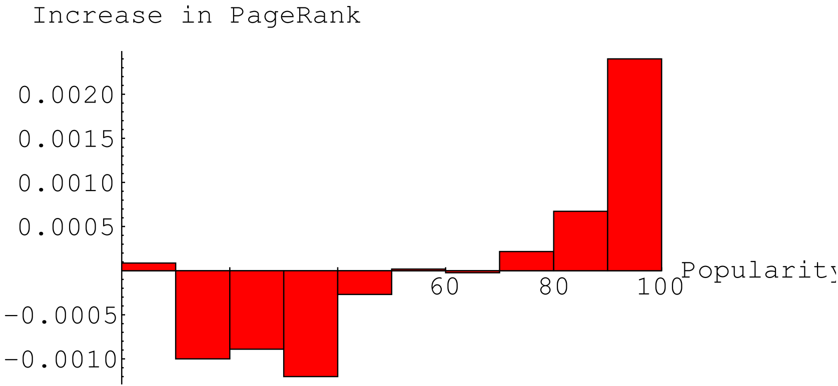
- Construct weekly Web-link graph
  - From the downloaded data
- Partition pages into 10 groups
  - Based on initial link popularity
  - Top 10% group, 10%-20% group, etc.
- How many new links to each group after a month?
  - Rich-get-richer → More new links to top groups

Result: Simple Link Count



- After 7 months
  - 70% of new links to top 20% pages
  - No new links to bottom 60% pages

Result: PageRank



- 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?
- New ranking metric
  - Page quality

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?
- New ranking metric
  - Page quality

Search Engine Impact

- How much bias do search engines introduce?

Search Engine Impact

- How much bias do search engines introduce?
- What we mean by bias?

Search Engine Impact

- How much bias do search engines introduce?
- What we mean by bias?
- What is the ideal ranking?
  How do search engines rank pages?

What is the Ideal Ranking?

What do we mean by page quality?

What is the Ideal Ranking?

What do we mean by page quality?

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

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

- Idea: More people will like a higher quality page
- Democratic measure of quality
  - p_1: 10,000 people, 8,000 liked it, $Q(p_1) = 0.8$
  - p_2: 10,000 people, 2,000 liked it, $Q(p_2) = 0.2$

  $\rightarrow Q(p_1) > Q(p_2)$

Page Quality $Q(p)$ Cont.

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

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

PageRank: Detail

- PageRank of p_i, $PR(p_i)$:

$$PR(p_i) = [PR(p_1)/c_1 + \cdots + PR(p_m)/c_m]^{\dagger}$$

  - $p_1, \ldots, p_m$: pages with links to p_i
  - c_j: number of outgoing links from p_j
  - Links from high PageRank pages have high "weights"

† "Damping factor" is ignored for simplicity

PageRank: Random-Surfer Model

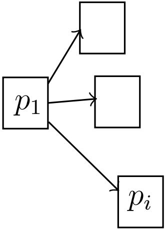
Random-Surfer Model

When users follow links randomly, $PR(p_i)$ is the probability to reach p_i

PageRank: Random-Surfer Model

Random-Surfer Model

When users follow links randomly, $PR(p_i)$ is the probability to reach p_i

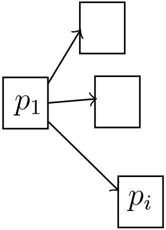


- $PR(p_1)$: probability to be at p_1

PageRank: Random-Surfer Model

Random-Surfer Model

When users follow links randomly, $PR(p_i)$ is the probability to reach p_i

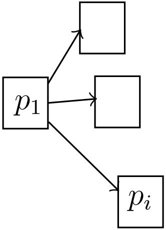


- $PR(p_1)$: probability to be at p_1
- Q: Probability to go from p_1 to p_i?

PageRank: Random-Surfer Model

Random-Surfer Model

When users follow links randomly, $PR(p_i)$ is the probability to reach p_i



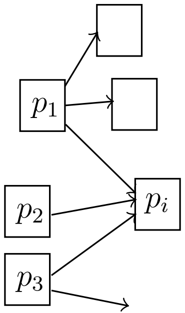
- $PR(p_1)$: probability to be at p_1
- Q: Probability to go from p_1 to p_i?

  A: $PR(p_1)/3$

PageRank: Random-Surfer Model

Random-Surfer Model

When users follow links randomly, $PR(p_i)$ is the probability to reach p_i



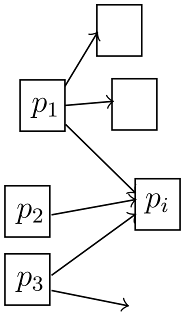
- $PR(p_1)$: probability to be at p_1
- Q: Probability to go from p_1 to p_i?

  A: $PR(p_1)/3$
- Q: Probability to be at p_i, $PR(p_i)$?

PageRank: Random-Surfer Model

Random-Surfer Model

When users follow links randomly, $PR(p_i)$ is the probability to reach p_i



- $PR(p_1)$: probability to be at p_1
- Q: Probability to go from p_1 to p_i?

  A: $PR(p_1)/3$
- Q: Probability to be at p_i, $PR(p_i)$?

  A: $PR(p_1)/3 + PR(p_2) + PR(p_3)/2$

Page Quality vs PageRank

- High PageRank
  $\rightarrow$ The page is currently “popular”
- PageRank $\approx$ Page quality if everyone is given equal chance
  - Before Google, PageRank may have been fair
- What about now?
  - High PageRank $\rightarrow$ High Quality?
    Low PageRank $\rightarrow$ Low Quality?

Page Quality vs PageRank

- High PageRank
  $\rightarrow$ The page is currently "popular"
- PageRank $\approx$ Page quality if everyone is given equal chance
  - Before Google, PageRank may have been fair
- What about now?
  - High PageRank $\rightarrow$ High Quality?
    Low PageRank $\rightarrow$ Low Quality?
- PageRank is biased against new pages
  - How to measure the PageRank bias?

Measuring Search-Engine Bias

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

Measuring Search-Engine Bias

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

Theoretical Web-User Models

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

$\longrightarrow$ Compare popularity evolution

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, $\mathcal{A}(p, t) = 0.3$

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

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

- PageRank is visit probability under random-surfer model
- Higher popularity $\rightarrow$ More visitors

Random-Visit Hypothesis

A visit is done by any user with equal probability

Random-Surfer Model: Analysis

Current popularity $\mathcal{P}(p,t)$

$\longrightarrow$ Number of visitors from $\mathcal{V}(p,t) = r \cdot \mathcal{P}(p,t)$

$\longrightarrow$ Awareness increase $\Delta\mathcal{A}(p,t)$

$\longrightarrow$ Popularity increase $\Delta\mathcal{P}(p,t)$

$\longrightarrow$ New popularity $\mathcal{P}(p,t+1)$

Random-Surfer Model: Analysis

Current popularity $\mathcal{P}(p,t)$

$\rightarrow$ Number of visitors from $\mathcal{V}(p,t) = r \cdot \mathcal{P}(p,t)$

$\rightarrow$ Awareness increase $\Delta\mathcal{A}(p,t)$

$\rightarrow$ Popularity increase $\Delta\mathcal{P}(p,t)$

$\rightarrow$ 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

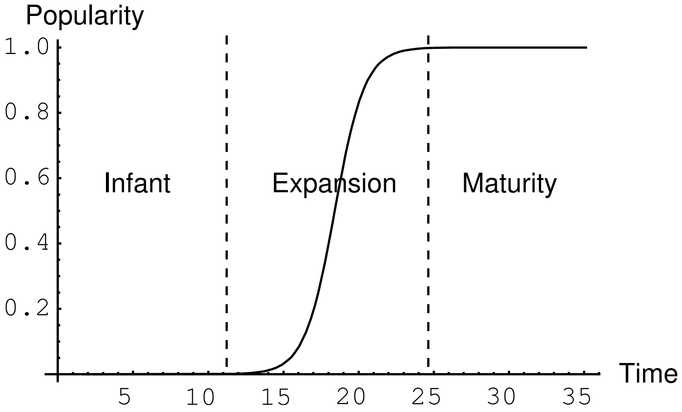
Theorem

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

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

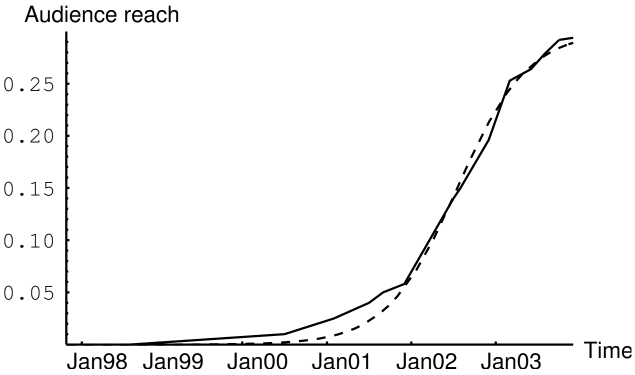
- $Q(p)$: quality of p
- $\mathcal{P}(p,0)$: initial popularity of p at time zero
- n: total number of Web users.
- r: normalization constant in $\mathcal{V}(p,t) = r \cdot \mathcal{P}(p,t)$

Random-Surfer Model: Popularity Graph



$$Q(p) = 1,\ \mathcal{P}(p, 0) = 10^{-8},\ \frac{r}{n} = 1$$

Comparison with Google Evolution



Data from Nielsen//NetRatings

$$Q(p) = 0.3,\ \mathcal{P}(p,0) = 5 \times 10^{-6},\ \frac{r}{n} = 8$$

Search-Dominant Model

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

Search-Dominant Model

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

- For ith result, how many clicks?
- For PageRank $\mathcal{P}(p,t)$, what ranking?

Search-Dominant Model

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

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

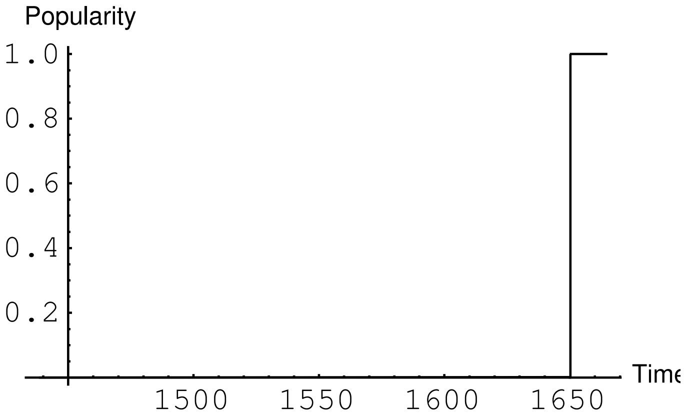
New Visit-Popularity Hypothesis

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

Random-Visit Hypothesis

A visit is done by any user with equal probability

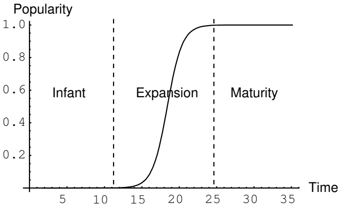
Search-Dominant Model: Result

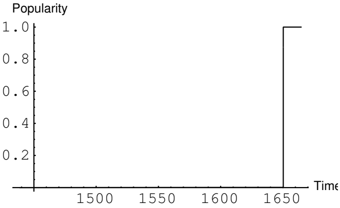


$$\sum_{i=1}^{\infty} \frac{[\mathcal{P}(p,t)]^{(i-\frac{9}{4})} - [\mathcal{P}(p,0)]^{(i-\frac{9}{4})}}{\left(i-\frac{9}{4}\right)\, Q(p)^i} = \frac{r}{n}\, t$$ (same parameters as before)

Comparison of Two Models

- Time to final popularity
  - Random surfer: 25 time units
  - Search dominant: 1650 time units
    $\longrightarrow$ 66 times increases!
- Expansion stage
  - Random surfer: 12 time units
  - Search dominant: non existent





Outline

- Web popularity-evolution experiment
  - Is “rich-get-richer” happening?
- Impact of search engines
  - Random-surfer model
  - Search-dominant model
- New ranking metric
  - How to measure page quality?

Measuring Quality: Basic Idea

- Quality: probability of link creation by a new visitor

Measuring Quality: Basic Idea

- Quality: probability of link creation by a new visitor
- Assuming the same number of visitors

$$Q(p) \propto \text{Number of new links (or popularity increase)}$$

Measuring Quality: Basic Idea

- Quality: probability of link creation by a new visitor
- Assuming the same number of visitors

$$Q(p) \propto \text{Number of new links (or popularity increase)}$$

Quality Estimator

$$Q(p) = \Delta\mathcal{P}(p)$$

Measuring Quality: Problem 1

- Different number of visitors to each page
  - More visitors to more popular page
- How to account for number of visitors?

Quality Estimator

$$Q(p) = \qquad \Delta\mathcal{P}(p)$$

Measuring Quality: Problem 1

- Different number of visitors to each page
  - More visitors to more popular page
- How to account for number of visitors?
- Idea: PageRank = visit probability

Quality Estimator

$$Q(p) = \quad \Delta\mathcal{P}(p)$$

Measuring Quality: Problem 1

- Different number of visitors to each page
  - More visitors to more popular page
- How to account for number of visitors?
- Idea: PageRank = visit probability

Quality Estimator

$$Q(p) = \quad \Delta\mathcal{P}(p)/\mathcal{P}(p)$$

Measuring Quality: Problem 2

- No more new links to very popular pages
  - Everyone already knows them
  - $\Delta\mathcal{P}(p)/\mathcal{P}(p) \approx 0$ for well-known pages
- How to account for well-known pages?

Quality Estimator

$$Q(p) = \quad \Delta\mathcal{P}(p)/\mathcal{P}(p)$$

Measuring Quality: Problem 2

- No more new links to very popular pages
  - Everyone already knows them
  - $\Delta\mathcal{P}(p)/\mathcal{P}(p) \approx 0$ for well-known pages
- How to account for well-known pages?
- Idea: $\mathcal{P}(p) = Q(p)$ when everyone knows p
  - Use $\mathcal{P}(p)$ to measure $Q(p)$ for well-known pages

Quality Estimator

$$Q(p) = \quad \Delta\mathcal{P}(p)/\mathcal{P}(p)$$

Measuring Quality: Problem 2

- No more new links to very popular pages
  - Everyone already knows them
  - $\Delta\mathcal{P}(p)/\mathcal{P}(p) \approx 0$ for well-known pages
- How to account for well-known pages?
- Idea: $\mathcal{P}(p) = Q(p)$ when everyone knows p
  - Use $\mathcal{P}(p)$ to measure $Q(p)$ for well-known pages

Quality Estimator

$$Q(p) = C \cdot \Delta\mathcal{P}(p)/\mathcal{P}(p) + \mathcal{P}(p)$$

C: weight given to popularity increase

Measuring Quality: Theoretical Proof

Theorem

Under the random-surfer model, the quality of page p, $Q(p)$, always satisfies the following equation:

$$Q(p) = \left(\frac{n}{r}\right)\left(\frac{d\mathcal{P}(p,t)/dt}{\mathcal{P}(p,t)}\right) + \mathcal{P}(p,t)$$

Compare it with $Q(p) = C \cdot \dfrac{\Delta\mathcal{P}(p)}{\mathcal{P}(p)} + \mathcal{P}(p)$

Is Page Quality Effective?

- How to measure its effectiveness?
  - Implement it to a major search engine?
  - Any other alternatives?

Is Page Quality Effective?

- How to measure its effectiveness?
  - Implement it to a major search engine?
  - Any other alternatives?
- Idea: Pages eventually obtain deserved popularity (however long it may take...)
  - "Future" PageRank $\approx Q(p)$

Page Quality: Evaluation (1)

$Q(p)$ as a predictor of future PageRank

- Compare the correlations of
  - "current" $Q(p)$ with "future" PageRank
  - "current" PageRank with "future" PageRank

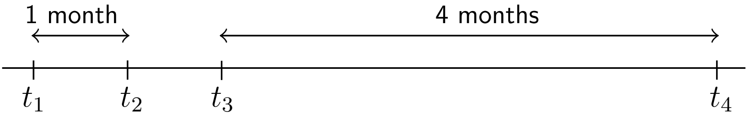
$\rightarrow$ $Q(p)$ predicts "future" PageRank better?

Page Quality: Evaluation (1)

$Q(p)$ as a predictor of future PageRank

- Compare the correlations of
  - "current" $Q(p)$ with "future" PageRank
  - "current" PageRank with "future" PageRank

  $\rightarrow$ $Q(p)$ predicts "future" PageRank better?
- Download the Web multiple times with long intervals

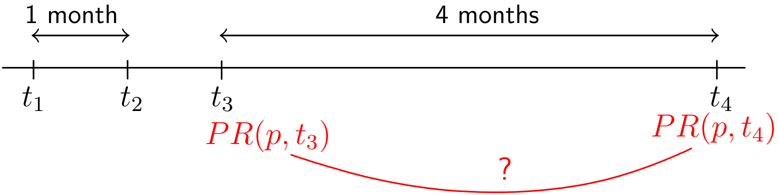


Page Quality: Evaluation (1)

$Q(p)$ as a predictor of future PageRank

- Compare the correlations of
  - “current” $Q(p)$ with “future” PageRank
  - “current” PageRank with “future” PageRank

  $\longrightarrow$ $Q(p)$ predicts “future” PageRank better?
- Download the Web multiple times with long intervals

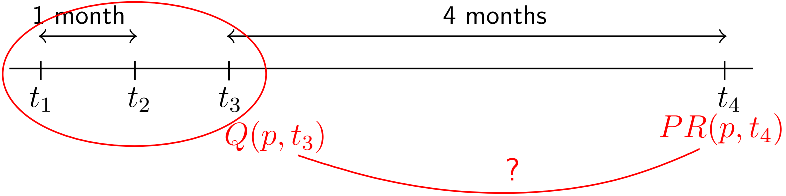


Page Quality: Evaluation (1)

$Q(p)$ as a predictor of future PageRank

- Compare the correlations of
  - "current" $Q(p)$ with "future" PageRank
  - "current" PageRank with "future" PageRank

  $\rightarrow Q(p)$ predicts "future" PageRank better?
- Download the Web multiple times with long intervals



Page Quality: Evaluation (2)

- Compare the average relative error

$$err(p) = \begin{cases} \left| \frac{PR(p,t_4) - Q(p,t_3)}{PR(p,t_4)} \right| \\ \left| \frac{PR(p,t_4) - PR(p,t_3)}{PR(p,t_4)} \right| \end{cases}$$

*For the pages whose PageRank consistently increased/decreased from t_1 through t_3.

Page Quality: Evaluation (2)

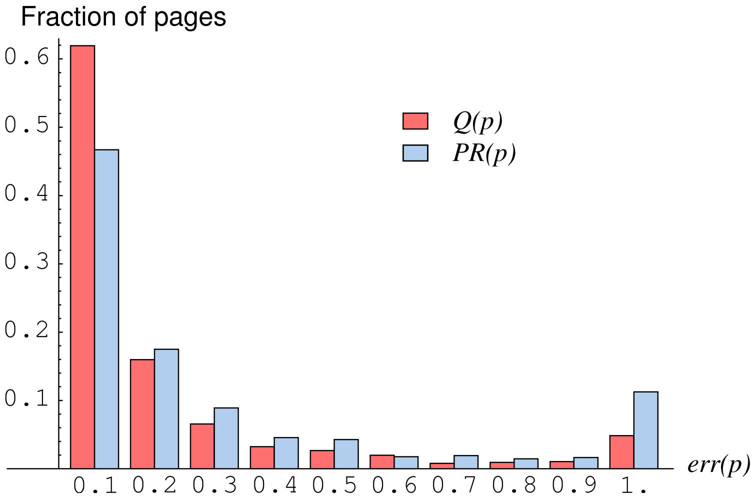
- Compare the average relative error

$$err(p) = \begin{cases} \left| \frac{PR(p,t_4) - Q(p,t_3)}{PR(p,t_4)} \right| \\ \left| \frac{PR(p,t_4) - PR(p,t_3)}{PR(p,t_4)} \right| \end{cases}$$

- Result *
  - For $Q(p, t_3)$: average $err = 0.32$
  - For $PR(p, t_3)$: average $err = 0.78$
  - $Q(p, t_3)$ twice as accurate.

*For the pages whose PageRank consistently increased/decreased from t_1 through t_3.

Quality Evaluation: More Detail



Summary

- Web popularity-evolution experiment
  - “Rich-get-richer” is indeed happening
- Impact of search engines
  - Random-surfer model
  - Search-dominant model

    $\longrightarrow$ Search engines have worrisome impact
- New ranking metric
  - Page quality: Based on popularity evolution
  - Identify high-quality pages early on

Thank You

For more details, see

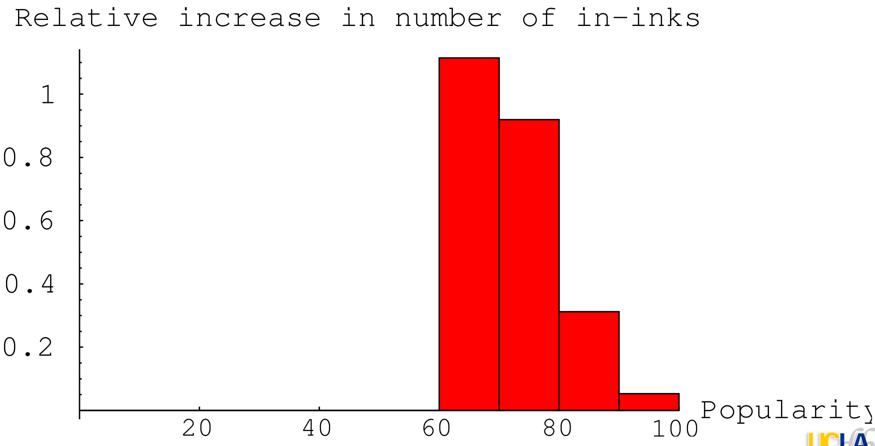
A. Ntoulas, J. Cho and C. Olston.
What's New on the Web?
In *WWW Conference*, 2004.

J. Cho and S. Roy
Impact of Web Search Engines on Page Popularity
In *WWW Conference*, 2004.

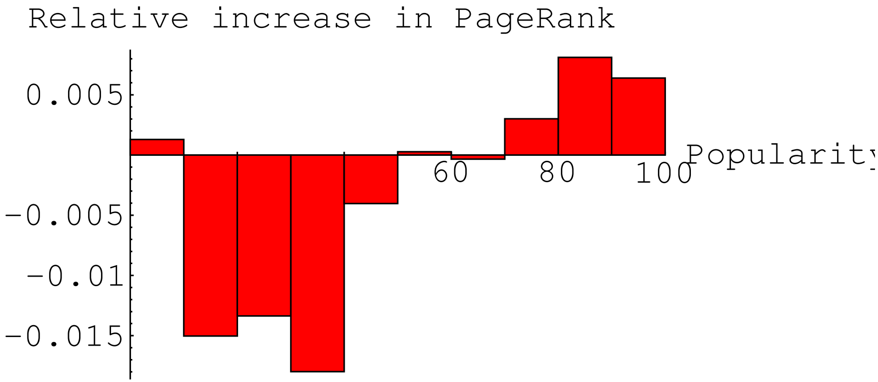
J. Cho and R. Adams.
Page Quality: In Search of an Unbiased Web Ranking
UCLA CS Department, Nov. 2003.

Any Questions?

Popularity Increase: Relative Link Count



Popularity Increase: Relative PageRank



Search-Dominant Model: Result