Impact of Search Engines on Page Popularity

Markus Ojala

October 10, 2007

Outline

Introduction

- J. Cho, S. Roy, Impact Of Search Engines On Page Popularity, WWW 2004
- "If your page is not indexed by Google, your page does not exist on the Web"
- PageRank metric ranks currently popular pages at top
- Popular pages get more popular

PageRank and popularity

- PageRang of page p_i is

$$PR(p_i) = d + (1 - d) \sum_{i=1}^{m} \frac{PR(p_i)}{c_i}$$

  with out going links c_i and damping factor d
- Measures the popularity of the page p_i

Outline

Popularity evolution: Experimental study

- We show that the "rich-get-richer" phenomenon exist
- Two snapshots of the Web at different times
- Measure the PageRank and the total number of incoming links for all pages of both snapshots

Experimental setup

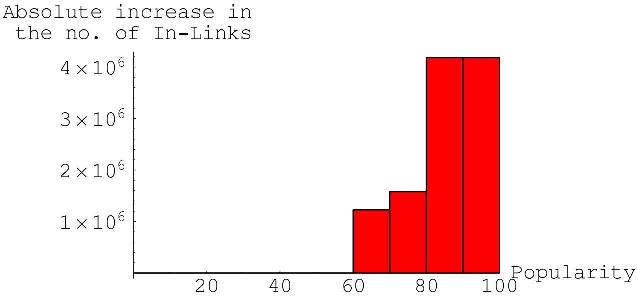
- Complete mirrors of 154 web sites
- Downloaded twice over a period of seven months
- Around 4.6 million pages for first snapshot S_1 and 5 million pages for second snapshot S_2
- Formed a directed graph of the web for each snapshot:
  - Each node corresponds to a unique web page
  - Directed edges corresponds to links
  - S_1 contains 13 million nodes and S_2 15 million nodes
  - Around 7.8 million common nodes
- PageRank and number of incoming links for common nodes
- Damping factor 0.3

Measuring popularity evolution

- We divide the pages into ten groups according to popularity
- Examine how the popularity changes between the groups
- Popularity measure I: Total number of incoming links
  - $IL(G_i, S_j) = \sum_{p \in G_i} IL(p, S_j)$
  - Group G_i, snapshot S_j, incoming links $IL(p, S_j)$ to page p
- Popularity measure II: PageRank
  - $PR(G_i, S_j) = \sum_{p \in G_i} PR(p, S_j)$

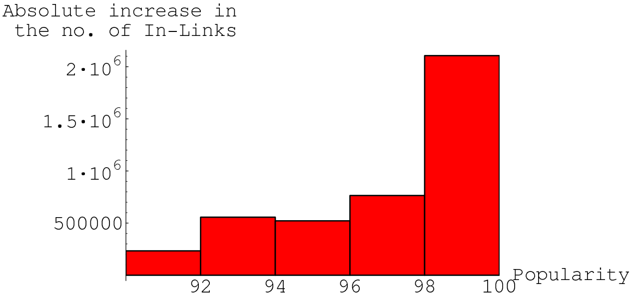
Popularity evolution: Number of incoming links

- Absolute increase in the number of incoming links



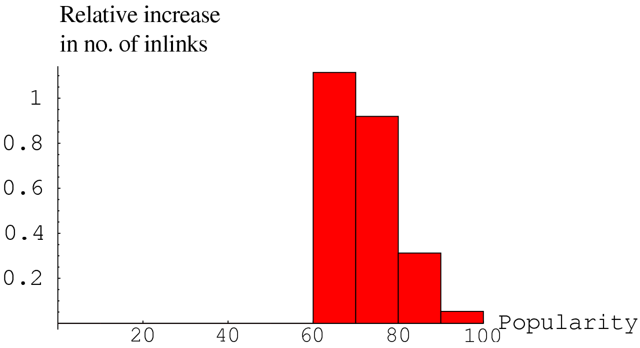
Popularity evolution: Number of incoming links

- Absolute increase in the number of incoming links



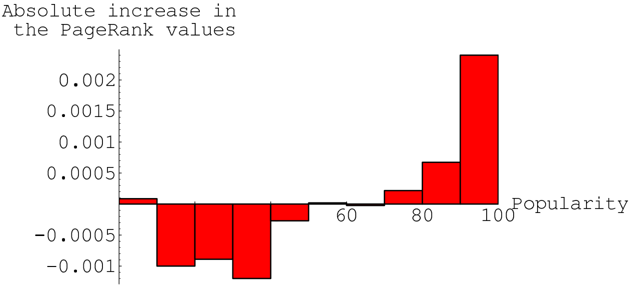
Popularity evolution: Number of incoming links

- Relative increase in the number of incoming links



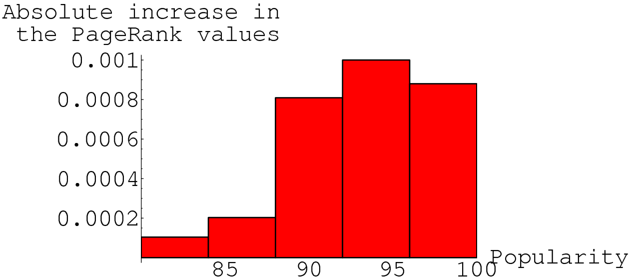
Popularity evolution: PageRank

- Absolute increase in the PageRank values



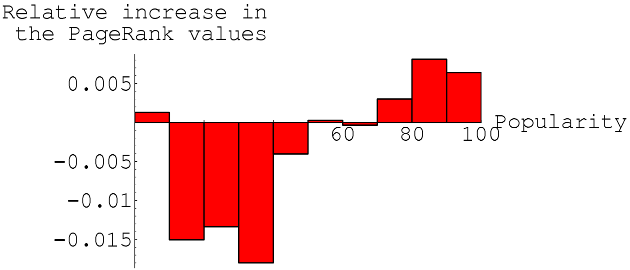
Popularity evolution: PageRank

- Absolute increase in the PageRank values



Popularity evolution: PageRank

- Relative increase in the PageRank values



Outline

Random-surfer model

- In random-surfer model users never use a search engine to discover pages
- New pages are discovered simply by following links
- *Popularity* $\mathcal{P}(p, t)$ of page p at time t is the fraction of web users who like the page, we assume $\mathcal{P}(p, t) = PR(p, t)$
- *Visit popularity* $\mathcal{V}(p, t)$ of page p at time t is the number of visits in the page p within a unit time interval at time t
- Proposition 1:
$$\mathcal{V}(p, t) = r_1 \mathcal{P}(p, t)$$
- Proposition 2: Any visit to a page can be done by any Web user with equal probability

Popularity evolution

- Quality $Q(p)$ of page p is the probability that an average user likes the page p when he visits p
- The total number of web users is n

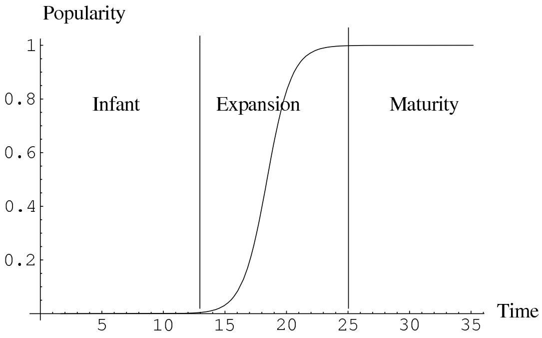
Theorem

The popularity of page p evolves over time as

$$\mathcal{P}(p,t) = \frac{Q(p)}{1 + \left[\frac{Q(p)}{\mathcal{P}(p,0)} - 1\right] e^{-\left[\frac{r_1}{n} Q(p)\right] t}}$$

Popularity evolution: example

- Assume $Q(p) = 1$, $r_1/n = 1$ and $\mathcal{P}(p, 0) = 10^{-8}$

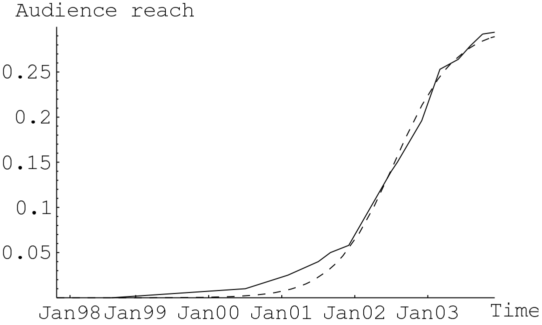


Case study: Google's popularity evolution

- The company Nielsen-NetRatings tracks how many web users visit some of the well-known web sites
- They report the *audience reach*: the fraction of web users visiting the particular site at least once in a week
- Google's popularity evolution is studied:
  - Statistics from the beginning of Google
  - The least affected by popularity-based ranking methods

Case study: Google's popularity evolution

- Google's popularity evolution: observed and random-surfer model ($Q(p) = 0.3$, $\mathcal{P}(p,0) = 5 \times 10^{-6}$, $r_1/n = 8$)



Outline

Search-dominant model

- In search-dominant model users discover pages solely based on search results
- Assumption 1: users use only one search engine
- Assumption 2: search engine always returns the *same* set of pages in the *same* order, ranked purely by their popularity
- The proposition 1 of random-surfer model not valid:

$$\mathcal{V}(p,t) = r_1 \mathcal{P}(p,t)$$

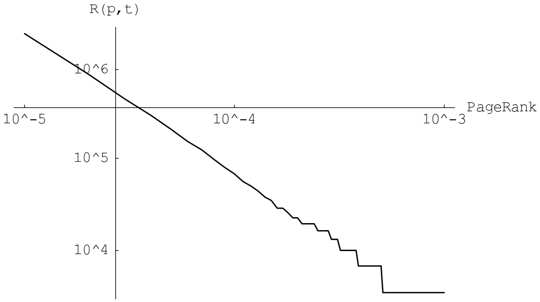
Visit popularity under the search-dominant model

- Derivation of new relationship between $\mathcal{V}(p,t)$ and $\mathcal{P}(p,t)$:
  1. Page returned as i^{th} entry, how likely is user to click it?
  2. Given the PageRank of a page, what is its ranking in the search result?
- $R(p,t)$ the rank of page p in the search result
- Lempel and Moran empirical measurements:

$$\mathcal{V}(p,t) = c_1 R(p,t)^{-\frac{3}{2}}$$

Visit popularity under the search-dominant model

- Probabilistic cumulative distribution of PageRank values:



- $R(p,t) = c_2 \mathcal{P}(p,t)^{-\frac{3}{2}}$

Visit popularity under the search-dominant model

- We get the relationship:

$$\begin{aligned}\mathcal{V}(p,t) &= c_1 R(p,t)^{-\frac{3}{2}} \\ &= c_1\left(c_2\mathcal{P}(p,t)^{-\frac{3}{2}}\right)^{-\frac{3}{2}} \\ &= r_2\mathcal{P}(p,t)^{\frac{9}{4}}\end{aligned}$$

- Example: pages p_1 and p_2, with popularity values 0.9 and 0.1.
  - Random-surfer: $\frac{\mathcal{V}(p_1,t)}{\mathcal{V}(p_2,t)} = \frac{0.9}{0.1} = 9$
  - Search-dominant: $\frac{\mathcal{V}(p_1,t)}{\mathcal{V}(p_2,t)} = \left(\frac{0.9}{0.1}\right)^{\frac{9}{4}} = 140$

Popularity evolution

- We get the following result:

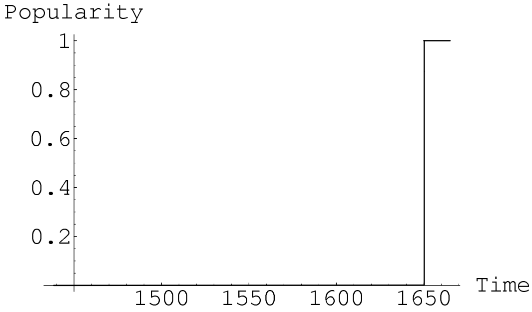
Theorem

Under the search-dominant model, the popularity of page p evolves through the equation

$$\sum_{i=1}^{\infty} \frac{[\mathcal{P}(p,t)]^{(i-\frac{9}{4})} - [\mathcal{P}(p,0)]^{(i-\frac{9}{4})}}{\left(i-\frac{9}{4}\right) Q(p)^{i}} = \frac{r_2}{n} t$$

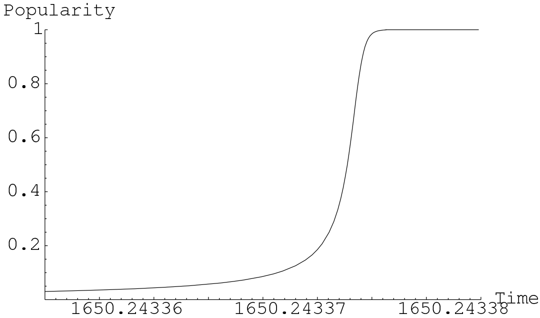
Popularity evolution

- Popularity evolution under the search dominant model with the same parameters as earlier ($Q(p) = 1$, $r_1/n = 1$ and $\mathcal{P}(p,0) = 10^{-8}$)



Popularity evolution

- Closer look:



Summary and conclusion

- We showed that "rich-get-richer" phenomenon exists
- We analyzed two theoretical models: Random-surfer model and search-dominant model
- It took 66 times longer to become popular with search-dominant model than with random-surfer model
- New ranking mechanism needed which can identify high-quality pages early