O randze stron

albo

Kogo lubi wyszukiwarka?

Piotr Kasprzyk

2008-05-27

Tytułem wstępu
albo
O czym będzie ten referat?

- Najpopularniejsza wyszukiwarka świata – Google
  - Najważniejsze narzędzie Google'a – PageRank
    - Najistotniejsze fakty o pierwszym i drugim

"The heart of our software is PageRank™, a system for ranking web pages developed by our founders Larry Page and Sergey Brin at Stanford University. And while we have dozens of engineers working to improve every aspect of Google on a daily basis, PageRank continues to play a central role in many of our web search tools."

http://www.google.com/technology

Spis rzeczy
albo
Długo jeszcze?

- Krótki wstęp
- Odrobina historii
- Opis działania algorytmu
- Obiektywizm rangi
- Varia

Odrobina historii
albo
Jak to się zaczęło?

- Eugene Garfield,

  University of Pennsylvania, 1955 (ISI)

- Larry Page, Sergey Brin,

  Stanford University, 1995

- Page + Brin = PageRank + Google,

  Google Inc., 1998

Krótki opis
albo
Jak to działa?

- U.S. Patent 6,285,999 (2001); 7,058,628 (2006)
- Wzór podstawowy:

$$PR(T_0) = \sum_{i=1}^{N} \frac{PR(T_i)}{C(T_i)}$$

  - PR(A) – wartość PageRank dla strony A
  - C(A) – liczba linków wychodzących ze strony A
  - N – łączna liczba rozważanych stron

- Intuicja: rozkład prawdopodobieństwa

$$\sum_{i=1}^{N} PR(T_i) = 1$$

- Rola czynnika tłumiącego (damping factor)

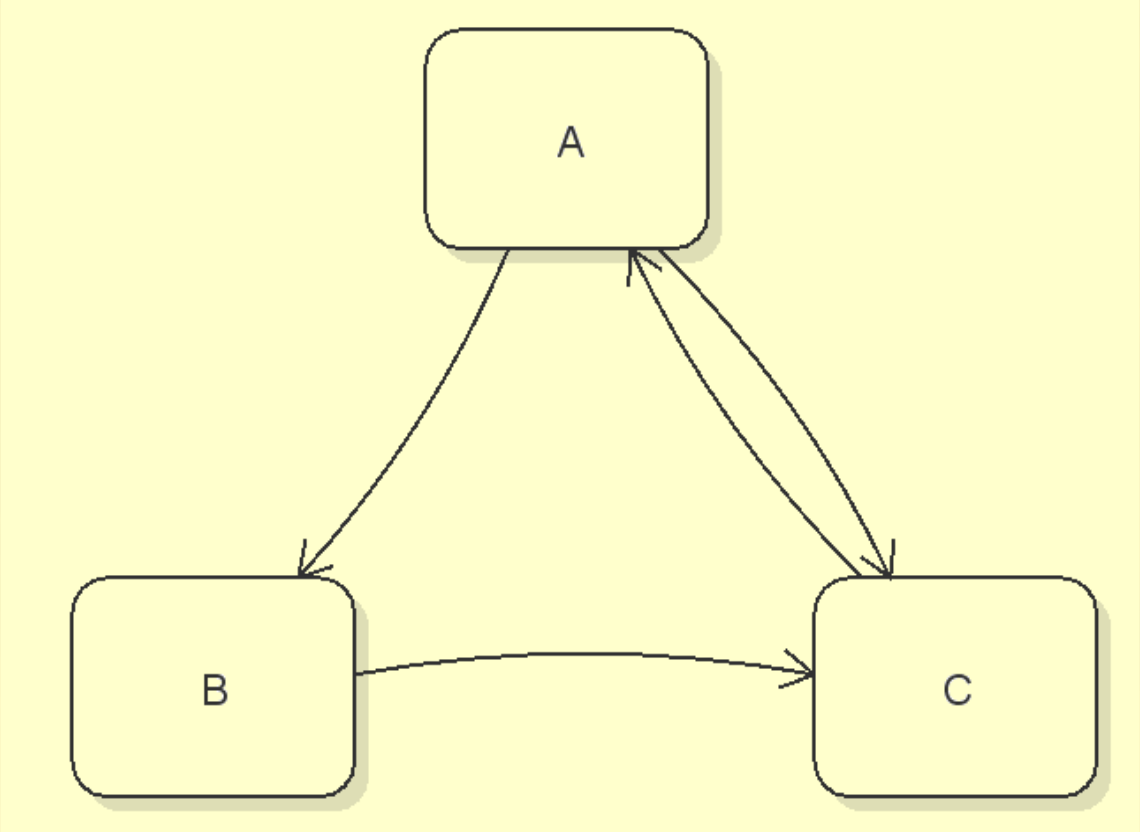
Krótki opis
albo
Jak to działa?

- Wzór właściwy:

$$PR(T_0) = \frac{1-d}{N} + d \cdot \sum_{i=1}^{N} \frac{PR(T_i)}{C(T_i)}$$

  - d – czynnik tłumiący

- Mały przykład:



Krótki opis?
albo
Jak to działa?

- ... a tak naprawdę?
- Algorytm iteracyjny
  - “In the case of millions of documents, sufficient convergence typically takes on the order of 100 iterations. It is not always necessary or even desirable, however, to calculate the rank of every page with high precision. Even approximate rank values, using two or more iterations, can provide very valuable, or even superior, information.”

    (z opisu patentu nr 7,058,628)
- Rangi wszystkich stron tworzą wektor własny odpowiednio zdefiniowanej macierzy sąsiedztwa

Koniec krótkiego opisu
albo
Nie ma żadnych pytań?

Wektor R postaci:

$$\mathbf{R} = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ \vdots \\ PR(p_N) \end{bmatrix}$$

Jest rozwiązaniem równania:

$$\mathbf{R} = \begin{bmatrix} (1-d)/N \\ (1-d)/N \\ \vdots \\ (1-d)/N \end{bmatrix} + d \begin{bmatrix} \ell(p_1,p_1) & \ell(p_1,p_2) & \cdots & \ell(p_1,p_N) \\ \ell(p_2,p_1) & \ddots & & \vdots \\ \vdots & & \ell(p_i,p_j) & \\ \ell(p_N,p_1) & \cdots & & \ell(p_N,p_N) \end{bmatrix} \mathbf{R}$$

przy definicji

$$\ell(p_i,p_j) = \begin{cases} 0 & na\ p_j\ nie\ ma\ linka\ do\ p_i \\ \frac{1}{C(p_j)} & wpp \end{cases}$$

cześć rysunków za: http://en.wikipedia.org/wiki/PageRank

Narzędzia pomocnicze
albo
Gdzie to mogę kupić?

- Google Toolbar, skala 0..10

  http://toolbar.google.com/

- Google Directory, skala ośmiopunktowa

  http://www.google.com/dirhp

- Strony “niegoogle'owe”

  np. http://www.selfseo.com/check_google_pagerank.php

Obiektywizm rangi
albo
A może by tak, wie pan...?

- Google chwali się swoim obiektywizmem – nie manipuluje PageRankiem, a reklamy są czytelnie (prawie) oddzielone od zwykłych wyników

- **Skoro jednak PageRank jest “głosem ludu”, to czy “lud” może coś zepsuć?**

Oczywiście!

Obiektywizm rangi
albo
A może by tak, wie pan...?

- Google Bomb
  - niesławne "miserable failure"
  - zasadniczo nie działa od początku 2007
- Exploit 302
  - przekierowywanie http 302 GoogleBotów
  - “nie wydaje się być już dużym problemem”
- Kupno i sprzedaż linków
  - ban lub “dewaluacja”
  - rel='nofollow' dla uczciwych
- Farmy linków (link farms)

- ... i wiele innych

Varia
albo
Mamy jeszcze chwilkę?

- Słynny i szeroko reklamowany obiektywizm Google'a działa, ale nie do końca
- Pewne rzeczy po prostu znikają z indeksów czy innych zasobów
- Słynny casus “chińskiego Google'a”
- “Dziury” w GoogleMaps (głownie w kształcie baz wojskowych) 1, 2, 3, 4...

Koniec
albo
Ileż można?

Linki jak zwykle przydatne:

- http://www.google.com/
- http://en.wikipedia.org/

Dziękuję serdecznie za uwagę