Wyszukiwanie i Przetwarzanie Informacji WWW

Analiza linków (2): Algorytm PageRank

Marcin Sydow

PJWSTK

Plan tego wykładu

- Idea PageRank
- 3 perspektywy: przepływy, losowy internauta i macierze
- Uproszczony i "realny" PageRank
- Matematyczne podstawy
- Obliczanie - metoda Potęgowa
- Rozszerzenia PageRank
- Porównanie HITS z PageRank

Linki a ważność dokumentu: zliczanie linków wchodzących

Skoro każdy link z dokumentu p do dokumentu q może być traktowany jako informacja, że dokument q jest "wartościowy" (w oczach autora dokumentu p) najprościej byłoby oceniać "ważność" lub "jakość" dokumentu docelowego q poprzez **zliczanie linków wchodzących do q** (ang. backlink count).

Im wyższy stopień wchodzący dokumentu q (backlink count) tym dokument może być ważniejszy (skoro wielu autorów wskazuje ten dokument)

Jest to analogiczne do "głosowania" dokumentów na inne dokumenty (każdy link to jeden głos)

To rozwiązanie ma poważną wadę:

Jest **bardzo podatne na celowe** manipulacje (ang. Search Engine Spam)

Ulepszony pomysł

Przy traktowaniu każdego linku jako równoważnego głosu i jednocześnie braku naturalnego mechanizmu w WWW pozwalającego identyfikować "nepotyzm" każdy podmiot może stworzyć **dowolną** ilość dokumentów zawierających linki do wybranego dokumentu będącego pod kontrolą tego samego podmiotu.

Ulepszenie: nie ważna jest **ilość** linków tylko ich **jakość**

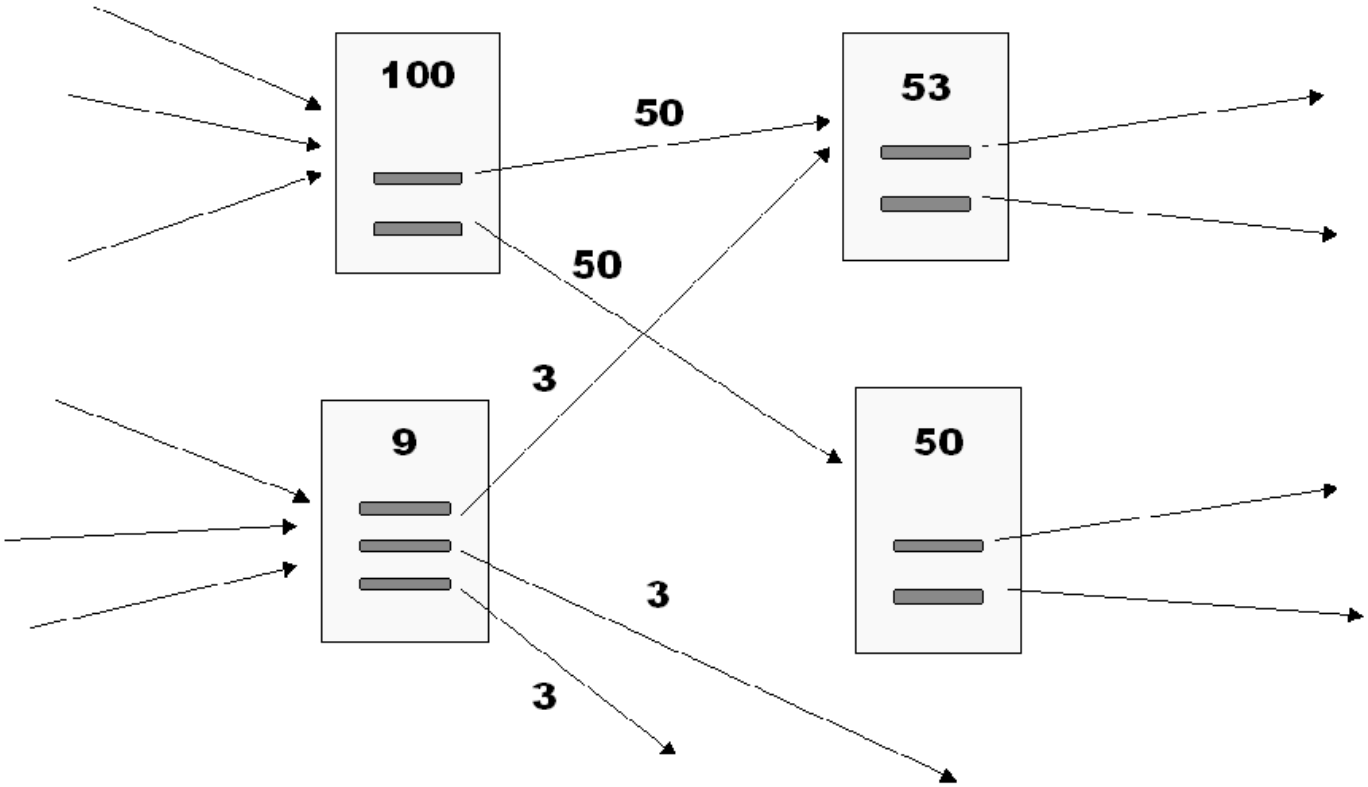
Analogia z głosowaniem: przy zliczaniu głosów uwzględnia się "reputację" głosujących.

Jeden link z bardzo ważnej strony może znaczyć dużo więcej niż 1000 linków z mało ważnych stron.

Za tą ideą (wziętą z m.in. analizy cytowań bibliograficznych) poszli twórcy algorytmu PageRank (wprowadzonego w Google ok 1998 roku)

Idea w uproszczeniu - przepływ "wartości" stron

- każda strona ma pewną wartość
- każda strona "głosuje" (poprzez linki) na inne strony
- o wartości strony decyduje wartość stron na nią głosujących



PageRank - uproszczone sformułowanie (perspektywa 1)

Interesuje nas przepływ przez graf WWW taki, że:

- Wartość przepływu sumuje się do 1
- to co wpływa = temu co wypływa (a'la prawo Kirchoffa 1)
- przepływ rozdziela się po równo

Daje to następujące równania:

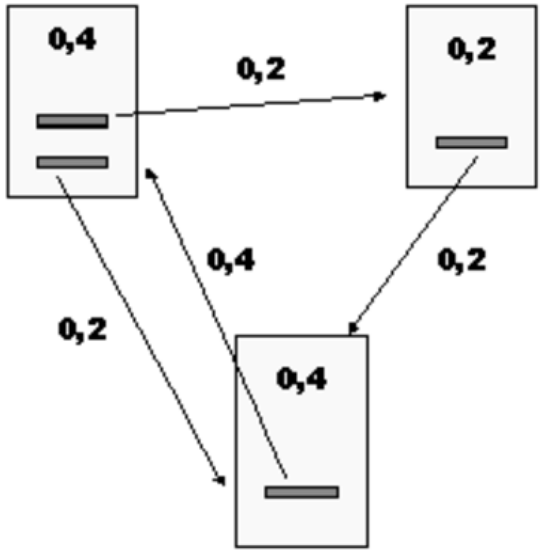
$$\sum_{d \in V} R(d) = 1, \tag{1}$$

$$\sum_{p \in IN(d)} R(p,d) = \sum_{q \in OUT(d)} R(d,q), \tag{2}$$

$$R(q) = \sum_{p \in IN(q)} R(p)/outDeg(p), \tag{3}$$

PageRank to wartość tego przepływu R(d) dla każdego dokumentu d

Przykład dla bardzo prostego grafu



Rysunek: (Jedyny) spełniający warunki przepływ przez przykładowy graf

Perspektywa 2: metafora "losowego internauty" (ang.Random Surfer)

Równoważne zdefiniowanie uproszczonego PageRanku:
Wyobraźmy sobie nieśmiertelnego internautę, który w każdej jednostce czasu przebywa na jakiejś stronie WWW i powtarza następującą akcję:

- wybiera (jednorodnie) losowo wychodzący link i podąża nim na następną stronę

Definition

Uproszczony PageRank dla strony d to graniczna średnia część jednostek czasu spędzonych na stronie d, dla wyżej opisanego procesu, przy czasie dążącym do nieskończoności.

Matematyk powie: "o ile granica istnieje..." I słusznie.

Perspektywa 3 - w języku macierzy

- G(V,E) - rozważany graf
- P - macierz sąsiedztwa G(V, E) zmodyfikowana w ten sposób, że każdy wiersz i jest podzielony przez $outDeg(d_i)$.

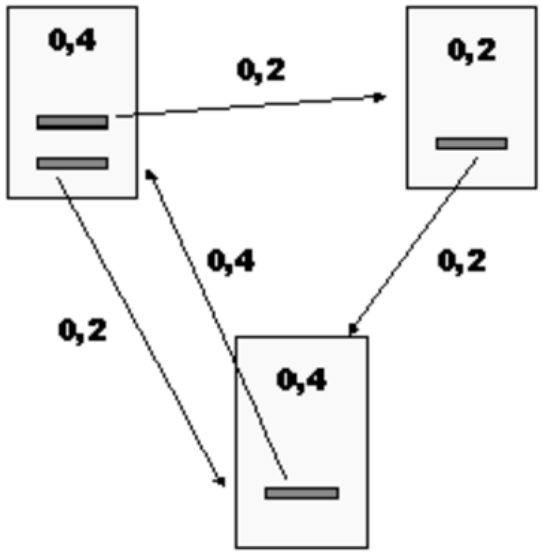
Oba poprzednie sformułowania PageRanku można wyrazić następująco:

Definition

PageRank
to wektor R będący punktem stałym przekształcenia liniowego P^T:

$$R = P^T R \tag{4}$$

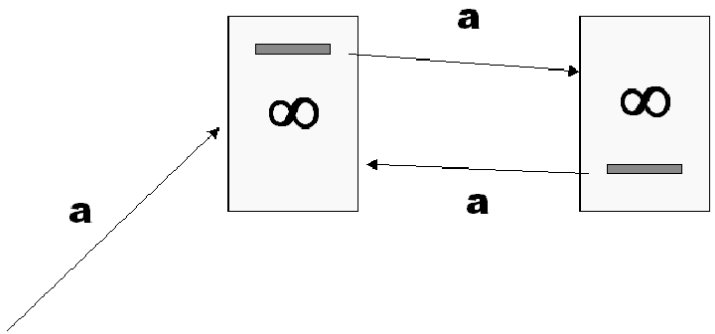
Przykład grafu i (jedyne) rozwiązanie



$$R = P^T R \tag{5}$$

$$\begin{pmatrix} 0.4 \\ 0.2 \\ 0.4 \end{pmatrix} = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}^T \begin{pmatrix} 0.4 \\ 0.2 \\ 0.4 \end{pmatrix} \tag{6}$$

Problemy z uproszczonym PageRankiem



Rysunek: "czarne dziury" (ang. rank sinks)

Problemy:

- Każdy maksymalny podgraf właściwy nie posiadający linków wychodzących pochłania cały PageRank w grafie.
- dokumenty nielinkowane otrzymają zerową wartość.

Jak poprawić uproszczony PageRank?

- łączymy każdy dokument bez wychodzących linków z każdym dokumentem
- dodajemy sztuczne linki pomiędzy wszystkimi pozostałymi parami dokumentów. Są one ważone ułamkowym współczynnikiem $0 < d < 1$ zwanym *decay factor*
- "prawdziwe" linki ważymy wartością (1 - d)

Powyższe sprawi, że w macierzy przejść P każdy wiersz będzie się sumował do 1. (przedtem niektóre wiersze były zerowe)
Macierz taka nazywa się *stochastyczna* i istnieje dla niej jednoznaczne rozwiązanie równania

$$R = P^T R \tag{7}$$

Rozwiązanie to jest głównym wektorem własnym tej macierzy.

Przykład na macierzach: (decay factor: 0.1)

$$P_0 = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$$P_1 = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$$P_2 = \begin{pmatrix} 1/60 & 28/60 & 28/60 & 1/60 & 1/60 & 1/60 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 19/60 & 19/60 & 1/60 & 1/60 & 19/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 28/60 & 28/60 \\ 1/60 & 1/60 & 1/60 & 28/60 & 1/60 & 28/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 55/60 & 1/60 \end{pmatrix}$$

Poprawiony PageRank w języku losowego internauty...

W każdej jednostce czasu losowy internauta przebywający na stronie s dokonuje następującej akcji:

- jeśli s zawiera linki wyjściowe:
  - z prawdopodobieństwem (1 - d) wybiera (jednorodnie) losowo link wychodzący z danej strony i nim podąża.
  - z prawdopodobieństwem d skacze do dowolnej losowo wybranej strony
- jeśli strona s nie ma linków wychodzących - z prawdopodobieństwem 1 skacze do losowo wybranej strony.

Definition

PageRank jest to rozkład stacjonarny zdefiniowanego powyżej nieredukowalnego i acyklicznego łańcucha Markowa

(rozkład ten określa graniczne prawdopodobieństwo bycia internauty na poszczególnych stronach)

...i w języku przepływów

- Uproszczony PageRank:

$$R(p) = \sum_{i \in IN(p)} R(i)/outDeg(i), \tag{8}$$

...i w języku przepływów

- Uproszczony PageRank:

$$R(p) = \sum_{i \in IN(p)} R(i)/outDeg(i), \tag{8}$$

- Dodanie sztucznych linków (uspójnienie):

$$R(p) = (1-d) \sum_{i \in IN(p)} \frac{R(i)}{outDeg(i)} + d \cdot v(p) \tag{9}$$

...i w języku przepływów

- Uproszczony PageRank:

$$R(p) = \sum_{i \in IN(p)} R(i)/outDeg(i), \tag{8}$$

- Dodanie sztucznych linków (uspójnienie):

$$R(p) = (1-d) \sum_{i \in IN(p)} \frac{R(i)}{outDeg(i)} + d \cdot v(p) \tag{9}$$

- Uwzględnienie "przymusowego" skoku z dokumentów bez linków wychodzących:

$$R(p) = (1-d) \sum_{i \in IN(p)} \frac{R(i)}{outDeg(i)} + d \cdot v(p) + (1-d)v(p) \sum_{i \in ZEROS} R(i), \tag{10}$$

Obliczanie PageRank z matematycznego punktu widzenia

$$R = P^T R \tag{11}$$

Z punktu widzenia matematyki, znalezienie wektora R jest łatwe.
Znajdowanie głównego wektora własnego jest równoważne rozwiązaniu układu równań liniowych.

Obliczanie PageRank w praktyce...

Czy można obliczyć PageRank rozwiązując układ równań?
Problemem jest rozmiar zadania.
Dla przykładu: załóżmy, że ilość dokumentów w grafie to 85M.

- Czas obliczeń: rozwiązywanie układu n równań ma złożoność $\Omega(n^2)$
- Rozmiar macierzy: $7,2P \times 4B = 28PB$ (!)

Co najmniej z tych powodów należy szukać specjalnych metod.

Obejście problemu czasu obliczeń

Metoda Potęgowa: Pozwala szybko obliczyć główny wektor własny macierzy w iteracjach, z teoretycznie dowolną precyzją:

1. $R_0 = v(p)$
2. $i = 0$
3. $R_{i+1} = P^T \cdot R_i$
4. i++
5. if $((|R_{i+1} - R_i| < threshold)$ OR $(i > max))$: stop
6. else: goto 3

Stawiamy pytanie: dla jakich macierzy P metoda potęgowa zbiega i daje jednoznaczny wektor R?

Warunki stosowalności metody potęgowej

Theorem

Metoda potęgowa zbiega do jednoznacznego rozwiązania R równania:

$$R = P^T R \tag{12}$$

jeśli stochastyczna macierz P jest nieredukowalna (odpowiada grafowi silnie spójnemu) i acykliczna. Wtedy, R to główny wektor własny tej macierzy.

- Graf silnie spójny: istnieje ścieżka między każdymi dwoma wierzchołkami
- Macierz acykliczna - odpowiada grafowi, w którym największy wspólny dzielnik długości wszystkich nietrywialnych cykli wynosi 1

Zauważmy, że dodanie sztucznych linków uczyniło graf silnie spójnym i acyklicznym.
Metoda potęgowa jest więc matematycznie poprawna.

Obejście problemu rozmiaru macierzy

- Macierz P jest bardzo duża.
- Oryginalna macierz P_0 (odpowiadająca uproszczonemu PageRankowi) jest jednak rzadka - zawiera "prawie same zera". Zmodyfikowane macierze P_1 i P_2 wprawdzie nie są już rzadkie, ale zmiany w stosunku do P_0 dadzą się wyrazić poprzez pojedyncze wektory
- W praktyce oznacza to, że informacje o strukturze grafu przechowuje się w postaci list sąsiedztwa.
- Rozmiar list sąsiedztwa dla grafu G(V,E) to O(|E|+|V|). (albo O(|E|) par (źródło,cel))
- Pojedyncza iteracja metody potęgowej jest zdominowana przez jednokrotny przegląd list sąsiedztwa

Szybkość metody potęgowej

W praktyce więc, pojedyncza iteracja dla grafu G(V,E) ma złożoność liniową (O(|E|+|V|))
Co ciekawe, ilość iteracji nie zależy silnie od |V|.
Ilość iteracji zależy od:

- współczynnika *decay factor*
- progu błędu t

Przy ustalonym progu błędu ilość iteracji metody potęgowej zależy od drugiej głównej wartości własnej macierzy P.
Można pokazać, że druga główna wartość własna P to wlaśnie (1 - d).
Wartość residuum zbiega do zera tak jak $(1-d)^n$
W praktyce ilość iteracji nie przekracza 100 dla zupełnie zadowalającej precyzji.

Usprawnienia obliczeniowe PageRank

Ze względu na rolę algorytmu PageRank i pokrewnych algorytmów w wyszukiwarkach oraz wielkość danych na których one pracują intensywnie badano usprawnienia związane z praktycznym ich obliczaniem:

- efektywne obliczanie w ograniczonej pamięci (podział grafu)
- adaptacyjne obliczanie (wykorzystanie niejednorodnej zbieżności na poszczególnych wierzchołkach grafu)
- wykorzystanie matematycznych własności równania PageRank (druga wartość własna)
- wykorzystanie blokowej struktury grafu WWW do równoległego obliczania PageRank
- przyspieszone obliczanie po niewielkich modyfikacjach grafu WWW

Problem "zwisających linków"

Nie jest możliwe posiadanie grafu całego WWW - ma się jedynie dostęp do jego części uzyskanej w procesie crawlowania.
W związku z tym, problem stanowi "brzeg" crawla - ta część dokumentów, do których odkryto linki, ale których nie zdążono ściągnąć. Linki takie nazywa się "zwisającymi" (ang. dangling).
Niestety, brzeg crawla rośnie w czasie i jego rozmiar zwykle przekracza rozmiar ściągniętego grafu, dla dużych crawli.
Aby to obejść przed liczeniem PageRank można usunąć w i iteracjach (ok. 5) zwisające linki aby dodać je z powrotem do grafu w ostatnich i iteracjach metody potęgowej.

Status algorytmu PageRank

PageRank jest opatentowany w USA:

- Method for node ranking in a linked database
  Inventor: Lawrence Page
  Assignee: The Board of Trustees of the Leland Stanford Junior University
  US Patent 7,058,628
  Granted June 6, 2006
  Filed July 2, 2001
- Filed January 9, 1998 and granted September 4, 2001:
  Method for node ranking in a linked database
- Filed July 6, 2001, and granted September 28, 2004:
  Method for scoring documents in a linked database

Znaczenie PageRank

Nowatorski w 1998 roku algorytm PageRank zrewolucjonizował rynek wyszukiwarek.

Niewielka, dysponująca niewielkim budżetem wyszukiwarka zaczęła skutecznie rywalizować z ówczesnymi gigantami dzięki pomysłowemu algorytmowi, który potrafił efektywnie i trafnie automatycznie porządkować wyniki wyszukiwania.

Obecnie, znaczenie klasycznego algorytmu PageRank w porządkowaniu wyników zmniejszyło się, gdyż wynaleziono techniki "oszukiwania" i jego (mimo, że z założenia należy do bardziej odpornych na manipulacje). Aktualna wersja używana przez wyszukiwarkę, w której powstał nie jest oczywiście publicznie znana i jest zaledwie jednym z wielu czynników uwzględnianych przy obliczaniu rankingu.

Rozszerzenia PageRank

Ze względu na swoje znaczenie historyczne, praktyczne zastosowania i ciekawe własności matematyczne algorytm PageRank doczekał się ogromnej ilości wariantów i rozszerzeń.

Do ważnych rozszerzeń należą m.in.:

- wersje personalizowane
- Topic-sensitive PageRank (czyli zależny od kontekstu zapytania)
- Trust-Rank, i Anti-TrustRank, (zastosowania w zwalczaniu spamu)
- rozmaite wersje rozszerzające model losowego internauty

Personalizacja

Klasyczna wersja PageRank pozwala na prostą i efektywną obliczeniowo "personalizację" za pomocą odpowiedniej modyfikacji "wektora ucieczki". W klasycznej wersji jest on jednorodny, ale już w pierwszej, oryginalnej publikacji na temat PageRank rozważano tę możliwość.

Personalizacja w tym wypadku polega na odpowiednim zwiększeniu prawdopodobieństw przejścia do dokumentów "bardziej interesujących" kosztem zmniejszenia pozostałych prawdopodobieństw.

Pomysł rozwiązania problemu skalowalności masowej personalizacji wektorów ucieczki jest zaprezentowany w: G.Jeh et al. "Scaling Personalized Web Search", WWW Conference 2003 (best paper award)

Topic-Sensitive PageRank

Klasyczny PageRank jest “statyczny” tzn. niewrażliwy na kontekst zapytania przychodzącego do wyszukiwarki.

Zaproponowano wersję “kontekstową” - wrażliwą na temat zapytania. Ranking dokumentu zależy wtedy nie tylko od struktury linków ale i od tematu zapytania.

T.Haveliwala “Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search”, WWW Conference 2002

TSPR - Idea

W klasycznym PageRanku liczy się (przed przetwarzaniem zapytania) **1 wektor** rankingu dla wszystkich dokumentów w kolekcji WWW.

W wersji Topic-Sensitive zaproponowano policzenie **wielu wektorów** (oryginalnie 16) - każdy z innym wektorem ucieczki - specjalnie dobranym do wybranej, "reprezentacyjnej" grupy tematycznej. Oryginalnie zaproponowano wykorzystanie 16 głównych kategorii ODP (Open Directory Project).

Przy obliczaniu rankingu dokumentu **w kontekście** zapytania q, bierze się kombinację liniową 16 rankingów, gdzie współczynniki wyrażają "bliskość" zapytania q do każdego z 16 składników tematycznych.

W pracy wykazano eksperymentalnie efektywność tego podejścia i jego przewagę jakościową nad klasycznym algorytmem PageRank.

Rozszerzanie modelu losowego internauty

Innym kierunkiem rozszerzania klasycznego algorytmu PageRank jest rozszerzanie bazowego modelu losowego inernauty poprzez dozwalanie na więcej akcji (niż wybór linku i skok do losowej strony)

Na przykład, oprócz 2 w/w akcji bardzo często wykonywaną akcją w przeglądarkach jest użycie klawisza "wstecz" (ang. "back-button").

Okazuje się, że da się tak zmodyfikować klasyczny model, żeby rozwiązanie było matematycznie zbieżne i zarazem efektywnie obliczalne na dużych grafach (mimo, że wynikowy proces nie jest już łańcuchem Markowa). Algorytm (RBS) pracuje na rzeczywistych grafach WWW. ("Random Surfer with back-step", M.Sydow, WWW Conference 2004, (oraz Fundamenta Informaticae, 2005))

Porównanie HITS z PageRank

Algorytm rankingowy HITS ma wiele podobieństw do PageRank.
Podobieństwa:

- Używa struktury linków (pierwszego rzędu) do automatycznego obliczenia wartości strony w grafie WWW.
- matematycznie sprowadza się do obliczenia głównych wektorów własnych pewnych macierzy powstałych z macierzy sąsiedztwa grafu

Różnice:

- HITS oblicza *ranking dynamiczny* — działa w kontekście konkretnego zapytania, a PageRank działa niezależnie od zapytań (*ranking statyczny*)
- HITS (oprócz autorytetu) modeluje pojęcie koncentratora. PageRank daje dobre rezultaty bez tego (być może dobre koncentratory w praktyce szybko stają się dobrymi autorytetami [Chakrabarti, "Mining the Web"]).

Stabilność i odporność na manipulacje

W HITS tempo zbieżności (podobnie jak w PageRank) zależy od tego jak daleko druga wartość własna leży od głównej.
PageRank jest znacznie stabilniejszy od HITS.
Matematyczna analiza stabilności HITS w porównaniu z PageRank:

- Ng, A. and A.Zheng and M.Jordan, "Stable Algorithms for Link Analysis", Proceedings of SIGIR'01, 2001
- Ng, A. and A.Zheng and M.Jordan, "Link analysis, eigenvectors and stability", Ijcai, pp. 903-910, 2001

Jak wspominaliśmy HITS jest łatwo podatny na manipulacje.
Aby sztucznie zwiększyć wartość autorytatywności jakiejś strony s wystarczy:

1. Stworzyć wiele stron linkujących do dobrych autorytetów (udają naturalne koncetratory)
2. Dodać linki z owych "sztucznych" koncentratorów do strony s

Na zaliczenie tego wykładu:

1. Idea PageRank
2. 3 perspektywy
3. Uproszczony PageRank i jego wady
4. "Realny" PageRank
5. Równanie PageRank i warunki jego rozwiązalności
6. Algorytm Potęgowy obliczania PageRank
7. Problem "zwisających" linków
8. Rozszerzenia PageRank
9. Personalizacja i Topic-Sensitive PageRank
10. Porównanie HITS i PageRank

Dziękuję za uwagę