

6 Wyszukiwarki internetowe.

Zapytania w wyszukiwarkach

Zapytania są z natury krótkie:

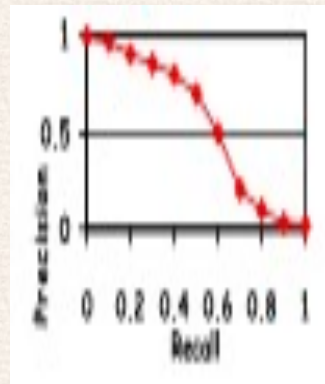
- 1,7 słowa 5 lat temu
- 2,4 obecnie

Oczekiwania użytkownika:

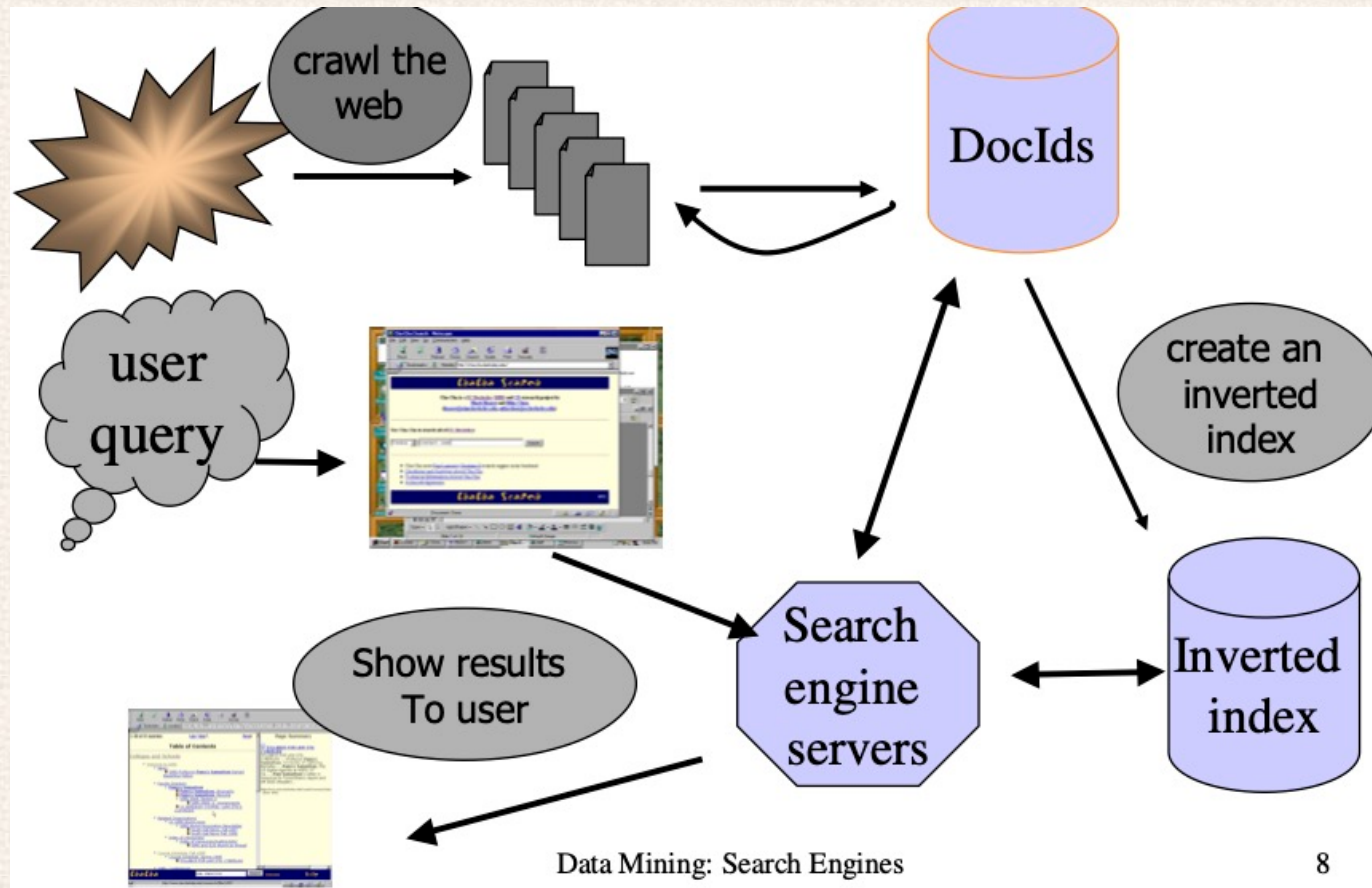
- "Pierwszy wynik powinien być tym, którego szukam"
- To działa, o ile użytkownik ma na myśli najbardziej popularne zapytanie, ale nie dla zapytań rzadkich

Ranking relewantności

- Jaka część zgłoszonych dokumentów była relewantna? (precision)
- Jaka część relewantnych dokumentów została zgłoszona? (recall)



Typowa architektura silnika wyszukiwarki internetowej



Heurystyczny ranking stron internetowych

Naiwne podejście stosowane przez wiele silników wyszukiwarek

Zastosowana heurystyka:

- liczba wystąpień danego zapytania na jednej stronie
- preferencja dla stron, gdzie zapytanie pojawia się wcześniej
- zwiększona waga dla stron, gdzie zapytanie pojawia się w wyróżnionym miejscu (np. w tytule)

Takie heurystyki epicko zawodzą:

- z powodu spamu
- z powodu synonimów etc.

Analiza grafu hiperłączy

Hypermedia jest siecią społecznościową.

Prowadzone są badania teorii sieci społecznościowych z użyciem teorii grafów

Wyniki są publikowane i cytowane

Zastosowania: wyszukiwarki internetowe HITS, Google, CLEVER

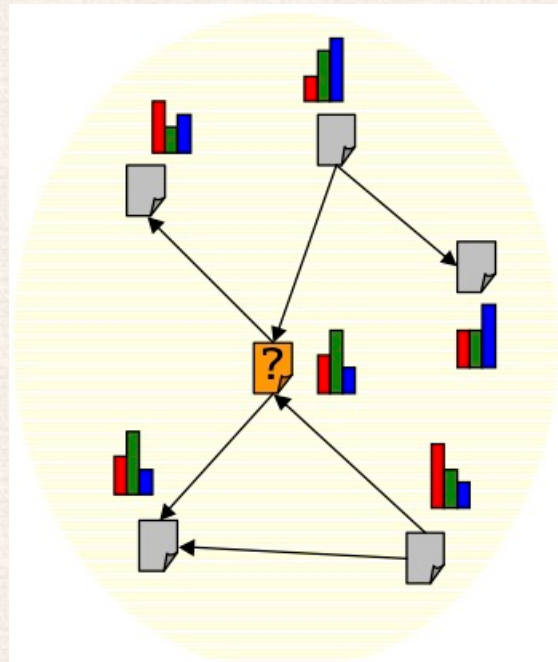
Modele hipertekstu do klasyfikacji

c = klasa, t = tekst, N = sąsiedzi

Model czystko tekstowy: $\Pr[t|c]$

Z wykorzystaniem tekstu sąsiada do oceny mojego tekstu: $\Pr[t, t(N)|c]$

Lepszy model: $\Pr[t, c(N)|c]$



Są to zadania programowania liniowego z nieliniową relaksacją

Wykorzystanie hiperlinkowej struktury sieci

Podstawowa zasada: każdy link implicite ocenia lokalizacją, do której prowadzi.

Główne założenie: Jeśli strony kierujące do danej strony są dobre, to ta strona też jest dobra.

Zasadniczo pomysł nie jest nowy i opiera się na badaniach z zakresu bibliometrii z lat 70-tych

Page rank

Intuicja: rekursywna definicja „ważności” – strona jest ważna, jeśli odpowiednio dużo ważnych stron do niej kieruje

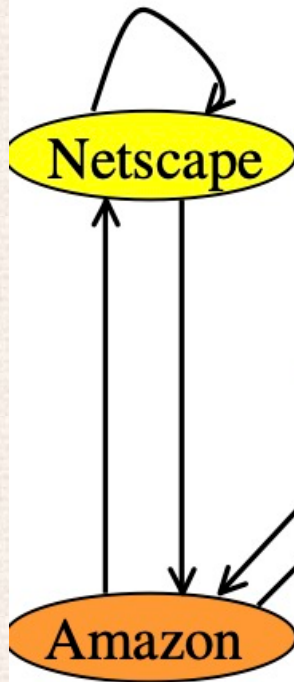
Metoda: stworzenia macierzy stochastycznej sieci.

- Każda strona odpowiada określonemu wierszowi i kolumnie macierzy
- Jeśli strona j ma n następników, to współrzędna ij macierzy równa jest $1/n$, jeśli strona i jest jednym z tych n następników strony j , lub 0 w przeciwnym razie

Page rank: intuicja

- Na początku każda strona ma jedną jednostkę „ważności”
- Przy każdym przebiegu, każda strona obdarza swoją ważnością swych następców i otrzymuje nową ważność od swych poprzedników
- Docelowo ważność osiąga wartość graniczną równą odpowiadającej mu współrzędnej głównego wektora własnego danej macierzy

Przykład: sieć w czasach prehistorycznych ;)



Equation:

$$\begin{bmatrix} N \\ M \\ A \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 \\ 1/2 & 1 & 0 \end{bmatrix} \begin{bmatrix} N \\ M \\ A \end{bmatrix}$$

Solution by relaxation:

$$\begin{array}{lcl} N = 1 & 1 & 5/4 & 9/8 & 5/4 & \dots & 6/5 \\ M = 1 & 1/2 & 3/4 & 1/2 & 11/16 & \dots & 3/5 \\ A = 1 & 3/2 & 1 & 11/8 & 17/16 & \dots & 6/5 \end{array}$$

Autorytarność i hubowość

Definiujemy je we wzajemnie rekurencyjny sposób:

- hub wskazuje do wielu wartościowych autorytetów
- autorytet jest połączony z wieloma dobrymi hubami

Tworzymy macierz podobną do tej w Page Rank, ale bez obostrzeń stochastycznych

- wiersze i kolumny odpowiadają stronom
- współrzędna ij macierzy jest równa 1 gdy strona i kieruje do strony j lub 0 w przeciwnym razie

Wielokrotne stosowanie tej macierzy prowadzi do rozbieżności

Możemy wszelako wprowadzić mnożniki skalujące, które zapewnią skończone wartości autorytarności i hubowości

Obliczanie autorytarności i hubowatości

Niech a i h będą wektorami.

i -ta współrzędna a i h odpowiada, odpowiednio, autorytarności i hubowatości i -tego wektora

Niech λ i μ będą odpowiednio dobranymi mnożnikami skalującymi

Wtedy:

- hubowatość danej strony jest sumą autorytarności stron, do których linkuje, skalowana przez λ :

$$h = \lambda A a$$

- autorytarność danej strony jest sumą hubowatości stron, które do niej linkują, skalowana przez μ :

$$a = \mu A^T h$$

Podstawiając otrzymujemy:

$$a = \lambda \mu A^T A a$$

oraz

$$h = \lambda \mu A A^T h$$

Wektory a i h możemy otrzymać przez relaksację dającą główne wektory własne macierzy $A A^T$ i $A^T A$.