

Eksploracja zasobów internetowych

Wykład 3

Pozyskiwanie informacji z sieci
WWW

Ranking stron na bazie linków

Piotr Hońko
Wydział Informatyki
Politechnika Białostocka

Wyszukiwanie informacji (ang. *information retrieval (IR)*)

- Dyscyplina, która wspomaga proces wyszukiwania informacji w dużych zbiorach dokumentów.
- Wyszukiwanie informacji polega na znalezieniu dokumentów, które pasują do zapytania użytkownika.
- W przypadku sieci WWW dokumentami są strony internetowe.

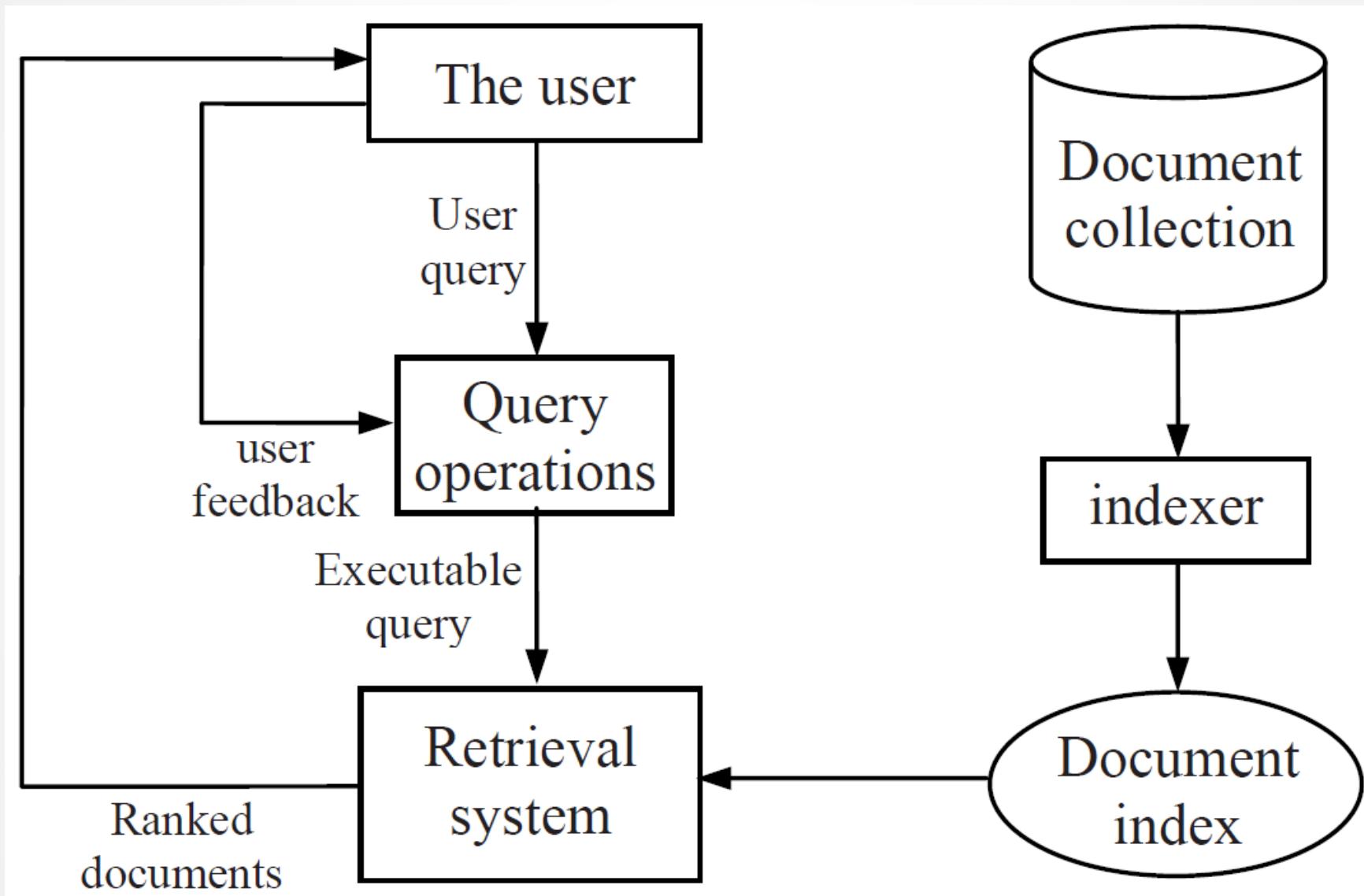
Zapytania SQL a zapytania w IR

- Baza relacyjna zawiera dane strukturalne, które wymagają stosowanie strukturalnego języka zapytań (SQL).
- Baza dokumentów, tworzona na potrzeby IR, to dane niestrukturalne, które nie są przystosowane do przeszukiwania za pomocą strukturalnego języka zapytań.

Dokument tekstowy a strona internetowa

- Pojedynczy dokument tekstowy nie jest ustrukturyzowany, składa się z tekstu, tabel, wykresów, rysunków. Może zawierać cytowania innych dokumentów.
- Strona internetowa jest częściowo ustrukturyzowana za pomocą znaczników kodu HTML, który stanowi dodatkowy składnik w stosunku do dokumentu tekstowego. Zwykle zawiera odnośniki do innych stron.

Schemat procesu wyszukiwania informacji



Rodzaje zapytań (1)

- Zapytanie logiczne
 - Zapytanie jest tworzone z wykorzystaniem słów kluczowych i operatorów logicznych AND, OR, NOT.
- Zapytanie ze słów kluczowych
 - Użytkownik wprowadza dowolną liczbę słów kluczowych oddzielonych spacją. Lista ta jest przekształcana do zapytania logicznego z wykorzystaniem operatora AND/OR.

Rodzaje zapytań (2)

- Zapytanie frazowe
 - Zapytanie tworzone z wykorzystaniem cudzysłowu. Wszystkie słowa muszą wystąpić w dokumencie w takiej kolejności, jaką jest we frazie.
- Zapytanie przybliżone
 - Wersja zapytania frazowego uwzględniająca niepełną zgodność wyniku z zapytaniem. Dopuszcza się występowanie innych słów niż podane we frazie.

Rodzaje zapytań (3)

- Zapytanie pełnodokumentowe
 - Zapytaniem jest istniejący w sieci i zindeksowany dokument. Wynikiem zapytania są dokumenty najbardziej do niego podobne.
- Zapytanie w języku naturalnym
 - Użytkownik przekazuje zapytanie w postaci zwykłego pytania. W celu znalezienia wyników wykorzystywane są metody przetwarzania języka naturalnego.
W praktyce stosowane w uproszczonej wersji, np. strona obsługująca zapytania o definicje technicznych pojęć.

Reprezentacja zapytania

- Zapytanie użytkownika jest przekształcane do postaci zgodnej z tą, w jakie jest zapisana baza, utworzona w oparciu o dokumenty.
- Zapytanie składające się z termów t_1, t_2, \dots, t_n może zostać zapisane za pomocą wektora

$$\vec{q} = (w_1, \dots, w_m),$$

gdzie każdy term $t_i (1 \leq i \leq n)$ jest reprezentowany przez dokładnie jedną wagę $w_j (1 \leq j \leq m)$. Liczba wag odpowiada liczbie wszystkich termów zgromadzonych w bazie.

Zapytanie w modelu boolowskim – przykład

- Zapytanie: $q = \{ \text{computer}, \text{program} \}$.
- Wektor zapytania $\vec{q} = (0,0,0,1,1)$.
- Wynik zapytania d_4, d_6 .

DID	lab	laboratory	programming	computer	program
d_1	0	0	0	0	1
d_2	0	0	0	0	1
d_3	0	1	0	1	0
d_4	0	0	0	1	1
d_5	0	0	0	0	0
d_6	0	0	1	1	1
d_7	0	0	0	0	1
d_8	0	0	0	0	1
d_9	0	0	0	0	0
d_{10}	0	0	0	0	0

Zapytanie w modelu ilościowym – przykład

- Zapytanie: $q = \{computer, program\}$.
- Wektor zapytania $\vec{q} = (0,0,0,1,1)$.
- Wynik zapytania $d_4(1.5), d_6(4.5)$.

DID	lab	laboratory	programming	computer	program
d_1	0	0	0	0	1
d_2	0	0	0	0	1
d_3	0	2	0	1	0
d_4	0	0	0	1	2
d_5	0	0	0	0	0
d_6	0	0	2	6	3
d_7	0	0	0	0	2
d_8	0	0	0	0	1
d_9	0	0	0	0	0
d_{10}	0	0	0	0	0

Zapytanie w modelu pozycyjnym – przykład

- Zapytanie: $q = \{computer, program\}$.
- Wektor zapytania $\vec{q} = (0,0,0,1,1)$.
- Wynik zapytania $d_4(4), d_6(2)$.

DID	lab	laboratory	programming	computer	program
d_1	0	0	0	0	[71]
d_2	0	0	0	0	[7]
d_3	0	[65,69]	0	[68]	0
d_4	0	0	0	[26]	[30,43]
d_5	0	0	0	0	0
d_6	0	0	[40,42]	[1,3,7,13,26,34]	[11,18,61]
d_7	0	0	0	0	[9,42]
d_8	0	0	0	0	[57]
d_9	0	0	0	0	0
d_{10}	0	0	0	0	0

Zapytanie w modelu wektorowym (TF) – przykład

- Zapytanie: $q = \{computer, program\}$.
- Wektor zapytania $\vec{q} = (0,0,0,0.5,0.5)$.
- Wynik zapytania $d_4(0.475), d_6(0.442)$.

Document ID	TF Coordinates				
\vec{d}_1	0	0	0	0	0.012
\vec{d}_2	0	0	0	0	0.010
\vec{d}_3	0	0.022	0	0.011	0
\vec{d}_4	0	0	0	0.017	0.034
\vec{d}_5	0	0	0	0	0.011
\vec{d}_6	0	0	0.026	0.078	0.039
\vec{d}_7	0	0	0	0	0.033
\vec{d}_8	0	0	0	0	0.013
\vec{d}_9	0	0	0	0	0
\vec{d}_{10}	0	0	0	0	0

Określanie zgodności wyniku z zapytaniem

- Mierzenie odległości/podobieństwa pomiędzy zapytaniem q a dokumentem d_j :
 - norma Euklidesowa

$$\|\vec{q} - \vec{d}_j\| = \sqrt{\sum_{i=1}^m (q^i - d_j^i)^2}$$

- podobieństwo cosinusowe

$$\vec{q} \cdot \vec{d}_j = \sum_{i=1}^m q^i \cdot d_j^i$$

Podobieństwo cosinusowe – przykład

- Zapytanie: $q = \{computer, program\}$.
- Wektor zapytania $\vec{q} = (0,0,0,0.5,0.5)$.

Document ID	TF Coordinates				$\vec{q} \cdot \vec{d}_j$
\vec{d}_1	0	0	0	0	0.012
\vec{d}_2	0	0	0	0	0.010
\vec{d}_3	0	0.022	0	0.011	0
\vec{d}_4	0	0	0	0.017	0.034
\vec{d}_5	0	0	0	0	0.011
\vec{d}_6	0	0	0.026	0.078	0.039
\vec{d}_7	0	0	0	0	0.033
\vec{d}_8	0	0	0	0	0.013
\vec{d}_9	0	0	0	0	0
\vec{d}_{10}	0	0	0	0	0

Ranking wyników zapytania

- Ranking wyników może być utworzony za pomocą miary zastosowanej do określania stopnia zgodności wyniku z zapytaniem.
- Przy wyznaczaniu rankingu należy również uwzględnić to, ile wyszukiwanych słów znajduje się w zwróconych dokumentach.
- Przykład: $\vec{d} = (0, 0, 0, 0.017, 0.034)$, $\vec{d}' = (0, 0, 0, 0, 0.072)$, $\vec{q} = (0, 0, 0, 0.5, 0.5)$.

	$\vec{q} \cdot \vec{d}_j$	a	$a \cdot \vec{q} \cdot \vec{d}_j$
\vec{d}	0.026	1	0.026
\vec{d}'	0.036	0.5	0.018

a – uśrednione występowanie termów w dokumencie.

Mierzenie jakości wyniku zapytania

- Oznaczenia pomocnicze:
 - D – zbiór wszystkich dokumentów w bazie;
 - $D_q \subseteq D$ – zbiór dokumentów istotnych dla zapytania q ;
 - $R_q \subseteq D$ – zbiór dokumentów zwroconych jako wynik zapytania q .
- Jakość wyniku może być mierzona za pomocą:
 - czułości
 - precyzji

$$\text{recall}(R_q) = \frac{|R_q \cap D_q|}{|D_q|},$$

$$\text{precision}(R_q) = \frac{|R_q \cap D_q|}{|R_q|}.$$

Mierzenie jakości wyniku zapytania – przykład

- Niech $D_q = \{d_2, d_3, d_4, d_6\}$, $R_q = \{d_3, d_6, d_7\}$.
- Ocena jakości wyniku zapytania q :

$$recall(R_q) = \frac{|R_q \cap D_q|}{|D_q|} = \frac{|\{d_3, d_6\}|}{|\{d_2, d_3, d_4, d_6\}|} = 0.5,$$

$$precision(R_q) = \frac{|R_q \cap D_q|}{|R_q|} = \frac{|\{d_3, d_6\}|}{|\{d_3, d_6, d_7\}|} = 0.67.$$

Ranking stron na bazie linków – Motywacje

- Popularność, autorytet i prestiż są kluczowymi pojęciami w sieciach społecznościowych.
- Bibliometria jest przykładem podejścia, które dostarcza ocenę artykułów, czasopism naukowych oraz autorów zbudowaną na bazie ich cytowań: liczba cytowań danego artykułu lub autora, prestiż czasopism cytujących.

Ranking stron na bazie linków – Idea

- Struktura linków stron WWW może zostać wykorzystana do mierzenia popularności, autorytetu czy prestiżu stron internetowych.
- Większa liczba linków wskazujących na daną stronę przekłada się na większe zainteresowanie stroną.
- Prestiż danej strony jest budowany na bazie prestiżów stron na nią wskazujących – rekurencyjna natura prestiżu strony.

Wyznaczenie prestiżu stron na podstawie macierzy sąsiedztwa

- Niech A będzie macierzą sąsiedztwa grafu, reprezentującego fragment sieci WWW, zdefiniowaną następująco:

$$A(u,v) = \begin{cases} 1 & \text{gdy dokument } u \text{ wskazuje na dokument } v, \\ 0 & \text{w przeciwnym przypadku.} \end{cases}$$

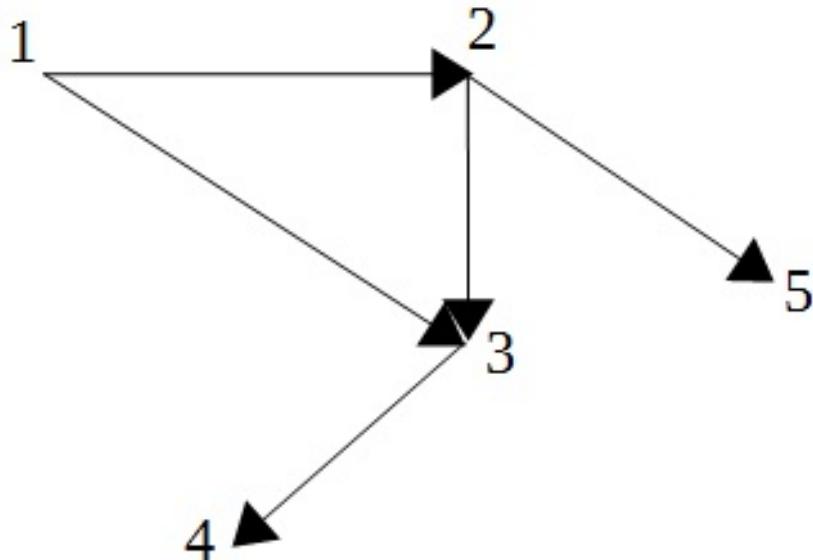
- Prestiż strony u zdefiniowany jest następująco

$$p(u) = \sum_v A(v,u) p(v)$$

gdzie $p(v)$ jest prestiżem strony v .

Wyznaczenie prestiżu stron na podstawie macierzy sąsiedztwa – przykład

- Graf



- Macierz sąsiedztwa

	1	2	3	4	5
1	0	1	1	0	0
3	0	0	0	1	0
4	0	0	0	0	0
5	0	0	0	0	0

- Niech $p(1)=1$.
- Otrzymujemy
$$p(4)=p(3)=$$
$$p(1)+p(2)=$$
$$p(1)+p(1)=2.$$

Wyznaczenie prestiżu stron za pomocą algebry liniowej

- Prestiż wszystkich dokumentów jest zapisywany w postaci wektora kolumnowego P (prestiż początkowy).
- Nowy wektor prestiżów obliczany jest następująco

$$P' = A^T P$$

gdzie A^T jest macierzą transponowaną macierzy A .

Wyznaczenie prestiżu stron za pomocą algebry liniowej c.d.

- Wielokrotne, rekurencyjne przeliczanie wektora prestiżów poprzez podstawianie P' w miejsce P prowadzi do tzn. punktu stałego dla P .
- Wektor ten jest wektorem własnym, który jest rozwiązaniem następującego równania

$$\lambda P = A^T P$$

gdzie λ jest wartością własną macierzy A^T .

- W celu ustalenia prestiżów stron, poszukiwany jest wektor własny powiązany z największą wartością własną.

Liczenia wartości własnych i wektorów własnych (1)

- Dana jest macierz $A = \begin{pmatrix} 3 & 4 \\ 5 & 2 \end{pmatrix}$.
- Wartości własne są rozwiązaniem równania $\det(A - \lambda I) = 0$.
- Otrzymujemy

$$\det(A - \lambda I) = \det \left(\begin{pmatrix} 3 & 4 \\ 5 & 2 \end{pmatrix} - \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \right) = \det \begin{pmatrix} 3-\lambda & 4 \\ 5 & 2-\lambda \end{pmatrix} = (3-\lambda)(2-\lambda) - 20.$$

Liczenia wartości własnych i wektorów własnych (2)

- Rozwiązaniem równania $(3-\lambda)(2-\lambda)-20=0$ są wartości własne $\lambda_1=-2, \lambda_2=7$.
- Wektor własny \vec{v}_i wyznaczamy za pomocą wzoru $(A - \lambda_i I) \vec{v}_i = \vec{0}$ gdzie $i=1,2$.
- Dla $\lambda_1=-2$ otrzymujemy

$$\begin{pmatrix} 3-(-2) & 4 \\ 5 & 2-(-2) \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 5 & 4 \\ 5 & 4 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 5x+4y \\ 5x+4y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Liczenia wartości własnych i wektorów własnych (3)

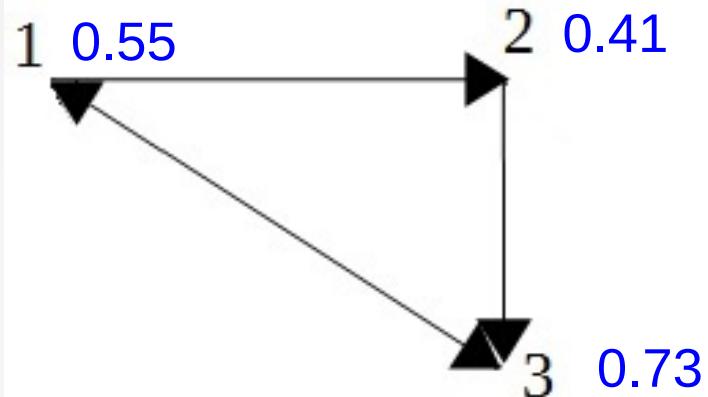
- Z równania $5x+4y=0$ otrzymujemy $x=-\frac{4}{5}y$.
- Zbiorem wektorów własnych związanych

$$\text{z } \lambda_1 = -2 \text{ jest } \vec{v}_1 = \left(-\frac{4}{5}y, y \right) = y \left(-\frac{4}{5}, 1 \right).$$

- Przykładowym wektorem własnym odpowiadającym wartości własnej $\lambda_1 = -2$ jest $\left(-\frac{4}{5}, 1 \right)$.

Wyznaczenie prestiżu stron na podstawie macierzy sąsiedztwa – przykład

- Graf



- Macierz sąsiedztwa A

	1	2	3
1	0	1	1
2	0	0	1
3	1	0	0

- Otrzymujemy

$$\lambda \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \quad \text{stąd} \quad 1.33 \begin{pmatrix} 0.55 \\ 0.41 \\ 0.73 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0.55 \\ 0.41 \\ 0.73 \end{pmatrix}$$

Wyznaczenie prestiżu stron za pomocą metody „power iteration”

- $P \leftarrow P_0$ – wektor początkowy
- loop:
 - $Q \leftarrow P$
 - $P \leftarrow A^T Q$
 - $P \leftarrow \frac{1}{\|P\|} P$ – normalizacja wektora
- while $\|P - Q\| > \varepsilon$ – dopuszczalna różnica
- Denormalizacja wartości własne i wektora własne:
$$\lambda = \|A^T P\|, \lambda P.$$

Metoda „power iteration” – przykład

- Dane $A^T = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}, P_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$.

$$P_1 = (1 \ 1 \ 2)^T, \lambda_1 = 4, \|P_1 - P_0\| = 1,$$

$$P_2 = (0.5 \ 0.25 \ 0.5)^T, \lambda_2 = 1.25, \|P_2 - P_1\| = 2.75,$$

$$P_3 = (0.4 \ 0.4 \ 0.6)^T, \lambda_3 = 1.4, \|P_3 - P_2\| = 0.15,$$

...

$$P_{10} = (0.429 \ 0.321 \ 0.571)^T, \lambda_{10} = 1.321, \|P_{10} - P_9\| = 0.12,$$

...

$$P_{16} = (0.43 \ 0.325 \ 0.57)^T, \lambda_{16} = 1.325, \|P_{16} - P_{15}\| = 0.$$

$$P = \lambda_{16} P_{16} = (0.57 \ 0.43 \ 0.76)^T.$$

Wyznaczanie prestiżu stron za pomocą algorytmu PageRank (1)

- Modyfikacja poprzedniego algorytmu, polegająca na uwzględnieniu prawdopodobieństwa przejścia z jednej strony na drugą.
- Ranking stron zbudowany za pomocą PageRank ma w większym stopniu odpowiadać sposobowi poruszania się użytkownika po sieci WWW.

Wyznaczanie prestiżu stron za pomocą algorytmu PageRank (2)

- Prestiż strony u zdefiniowany jest następująco

$$R(u) = \sum_v \frac{A(v, u)}{N_v} R(v)$$

gdzie $N_v = \sum_w A(v, w)$ jest liczbą stron, na które wskazuje strona v , a $R(v)$ jest prestiżem strony v .

Wyznaczanie prestiżu stron za pomocą algorytmu PageRank (3)

- Niech A będzie macierzą sąsiedztwa grafu, reprezentującego fragment sieci WWW, zdefiniowaną następująco:

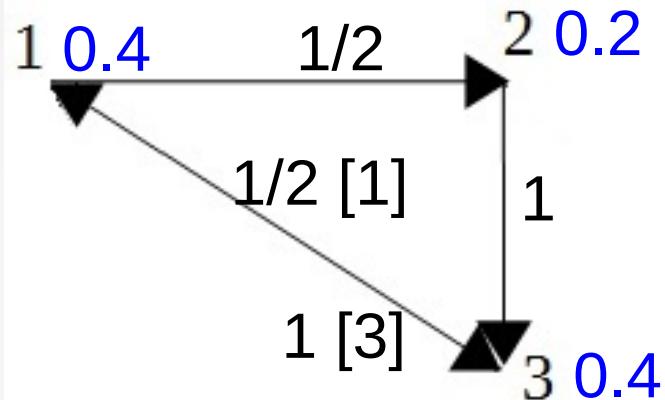
$$A'(u,v) = \begin{cases} 1/N_v & \text{gdy dokument } u \text{ wskazuje na dokument } v, \\ 0 & \text{w przeciwnym przypadku.} \end{cases}$$

- Prestiż strony u może być zdefiniowany następująco

$$R(u) = \sum_v A'(v,u) R(v).$$

Wyznaczenie prestiżu stron na podstawie macierzy sąsiedztwa – przykład

- Graf



- Macierz sąsiedztwa A

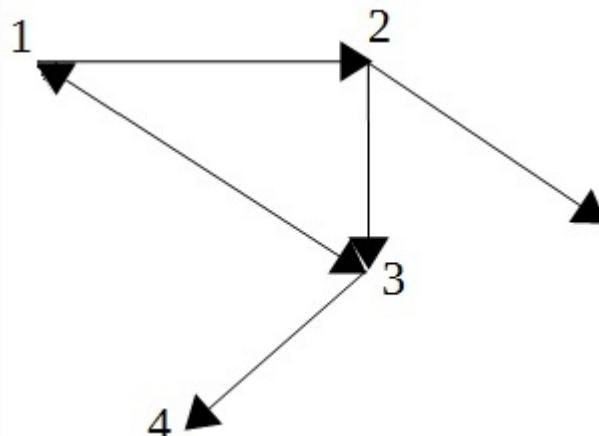
	1	2	3
1	0	1/2	1/2
2	0	0	1
3	1	0	0

- Otrzymujemy

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 1/2 & 0 & 0 \\ 1/2 & 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} \quad \text{stąd} \quad \begin{pmatrix} 0.4 \\ 0.2 \\ 0.4 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 1/2 & 0 & 0 \\ 1/2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0.4 \\ 0.2 \\ 0.4 \end{pmatrix}$$

Wyznaczenie prestiżu stron za pomocą algebry liniowej c.d.

- Czy wektor własny jest dobrym rozwiązaniem w przypadku każdego grafu?



- Dla grafu

$$\lambda \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}$$

stąd $\lambda = 0$ i

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

Modyfikacja algorytmu PageRank

- Algorytm PageRank jest odporny na występowanie pętli, które składają się z dowolnej liczby stron, pod warunkiem, że pętle te posiadają linki, które pozwalają z nich wyjść.
- Wielokrotne przemieszczanie się po stronach izolowanej pętli, tj. pętli bez linków umożliwiających wyjście z niej, sztucznie zwiększa prestiż stron.
- W celu radzenia sobie z izolowanymi pętlami, wykorzystywany jest wektor źródeł prestiżu stron (ang. *rank source vector*).

Modyfikacja algorytmu PageRank c.d.

- Prestiż strony u zdefiniowany jest następująco

$$R(u) = \lambda \left[\sum_v \frac{A(v, u)}{N_v} R(v) + E(u) \right]$$

gdzie E jest wektorem źródeł prestiżu wszystkich stron, który definiuje rozkład prawdopodobieństwa przejścia do losowo wybranego dokumentu sieciowego.

- Im wyższa norma, tym skoki do losowych stron występują częściej. Zwykle $\|E\| = 0.15$.

Zmodyfikowana metoda „power iteration”

- $R \leftarrow R_0$
- loop:
 - $Q \leftarrow R$
 - $R \leftarrow A^T Q$
 - $R \leftarrow R + E$
 - $R \leftarrow R / \|R\|$
- while $\|R - Q\| > \varepsilon$

Zmodyfikowana metoda „power iteration” – przykład

- Dane $A^T = \begin{pmatrix} 0 & 0 & 1 \\ 1/2 & 0 & 0 \\ 1/2 & 1 & 0 \end{pmatrix}$, $R_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$, $E = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$.

$$R_1 = (1 \ 0.5 \ 1.5)^T, \lambda_1 = 3, \|R_1 - R_0\| = 0,$$

$$R_2 = (0.5 \ 0.167 \ 0.33)^T, \lambda_2 = 1, \|R_2 - R_1\| = 2,$$

$$R_3 = (0.33 \ 0.25 \ 0.417)^T, \lambda_3 = 1, \|R_3 - R_2\| = 0,$$

...

$$R_{10} = (0.396 \ 0.198 \ 0.406)^T, \lambda_{10} = 1, \|R_{10} - R_9\| = 0,$$

...

$$R_{18} = (0.4 \ 0.2 \ 0.4)^T, \lambda_{18} = 1, \|R_{18} - R_{17}\| = 0.$$

Algorytm PageRank – cechy

- Algorytm symuluje sposób poruszania się użytkownika, który losowo przemieszcza się pomiędzy stronami (model *Random Web Surfer*).
- Algorytm radzi sobie ze sztucznym zwiększeniem prestiżu stron (pętle wzajemnych odwołań bez linków wyjściowych) poprzez ograniczenie liczby przejść pomiędzy tymi samymi stronami i dzięki przejściu na inną, losowo wybraną stronę (wektor źródła prestiżu).

Wyznaczanie autorytetów i identyfikacja stron typu hub – Algorytm HITS (1)

- Algorytm dokonuje oceny stron fragmentu sieci powiązanej z wynikiem konkretnego zapytania.
- Graf budowany jest na podstawie istotnych stron, zwróconych przez zapytanie oraz stron z nimi powiązanych (strony wchodzące i wychodzące).
- Algorytm jednocześnie określa autorytet strony oraz to, w jakim stopniu dana strona jest typu hub (ang. *hub page*).

Wyznaczanie autorytetów i identyfikacja stron typu hub – Algorytm HITS (2)

- R_q – zbiór istotnych stron zwróconych przez zapytanie q .
- S_q – zbiór stron powstałych na bazie zbioru R_q poprzez dodanie stron wchodzących i wychodzących.
- $X = (x_1, \dots, x_n), Y = (y_1, \dots, y_n)$ – wektory stron z S_q ($n = |S_q|$) określające odpowiednio autorytety i stopnie bycia stroną typu hub.

Wyznaczanie autorytetów i identyfikacja stron typu hub – Algorytm HITS (3)

- $X \leftarrow (11 \dots 1)$
- $Y \leftarrow (11 \dots 1)$
- loop k times
 - $x_u \leftarrow \sum_{\{v, A(v,u)=1\}} y_v$, for $u = 1, 2, \dots, n$
 - $y_u \leftarrow \sum_{\{v, A(u,v)=1\}} x_v$, for $u = 1, 2, \dots, n$
 - normalize X and Y by the L_2 norm
- end loop

k – liczba iteracji określona z góry,
 $L_2 : \|X\| = \sqrt{x_1^2 + \dots + x_n^2}$.

Algorytm HITS – przykład

- Dane $A = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$, $X_0 = (1 \ 1 \ 1)$, $Y_0 = (1 \ 1 \ 1)$.
 $X_1 = (0.48 \ 0.48 \ 0.816)$, $Y_1 = (0.816 \ 0.48 \ 0.48)$,
 $X_2 = (0.267 \ 0.535 \ 0.802)$, $Y_2 = (0.802 \ 0.535 \ 0.267)$,
 $X_3 = (0.169 \ 0.507 \ 0.845)$, $Y_3 = (0.845 \ 0.507 \ 0.169)$,
 $X_{10} = (0.006 \ 0.526 \ 0.851)$, $Y_{10} = (0.851 \ 0.526 \ 0.006)$,
 $X_{15} = (0.001 \ 0.526 \ 0.851)$, $Y_{15} = (0.851 \ 0.526 \ 0.001)$.

Algorytm HITS – cechy

- Jakości stron nie mogą być wyznaczone, dopóki nie jest znany wynik zapytania (obliczenia w czasie rzeczywistym).
- Prestiż wyznaczany jest dla fragmentu sieci, który zawiera istotne strony, przez co strony o niepasującej treści nie zawyżają sztucznie rankingu.
- W wersji podstawowej algorytm nie jest odporny na niespójność grafu zbudowanego na bazie zapytania, które zwraca niepowiązane ze sobą strony.