Eksploracja zasobów internetowych

Wykład 4

Ranking wyników na bazie linków

mgr inż. Maciej Kopczyński

Białystok 2014

Wstęp

Poznane do tej pory mechanizmy sortowania istotności zwróconych wyników bazowały na zawartości tekstowej dokumentów.

Jednak cechą sieci WWW jest grafowy charakter, co umożliwia wzbogacenie sortowania o zależności krawędziowe pomiędzy wierzchołkami grafu.

Dzięki takiemu rozwiązaniu możliwe jest bardziej niezawodne sortowanie wyników otrzymanych na bazie zapytania.

Wstęp

Idea rozwiązania opiera się na przydzielaniu każdej stronie WWW pewnego współczynnika określającego jej istotność.

Analiza połączeń pomiędzy stronami może być wykonywana w trybie offline, bazując na już zindeksowanych stronach WWW.

Mechanizmy wykorzystujące zależności pomiędzy dokumentami są szeroko wykorzystywane w systemach bibliograficznych związanych np. z artykułami naukowymi i ich wzajemnymi cyto-waniami.

Social networks

Sieć WWW jest przykładem sieci typu *Social networks*, w których popularność danego obiektu jest uzależniona od ilości połączeń do tego elementu (w sensie krawędzi grafu) pochodzących z innych elementów.

Istotnymi pojęciami występującymi w sieciach *Social networks* są:

- popularność,
- autorytet,
- współczynnik prestiżu.

Współczynnik prestiżu

Dla każdej zindeksowanej strony WWW przypisywany jest współczynnik prestiżu. Jest on obliczany na podstawie ilości linków wskazujących na daną stronę, a także współczynników prestiżu stron linkujących.

Współczynnik prestiżu oznaczany jest jako $p(u)$.

Jak jest obliczana wartość współczynnika prestiżu?

Wartość współczynnika prestiżu jest wyliczana rekurencyjnie.

Obliczanie współczynnika prestiżu

Sieć WWW może być reprezentowana jako graf skierowany.

Przyjmijmy następujące oznaczenia:

- A – macierz sąsiedztwa grafu,
- $A(u,v)$ – krawędź łącząca wierzchołek u i v.

Wartość współczynnika prestiżu dla dokumentu u jest zdefiniowana jako:

$$p(u)=\sum_{v} A(v,u)\,p(v)$$

Obliczanie współczynnika prestiżu

Wartości *p*(*u*) dla wszystkich dokumentów mogą być zapisane jako wektor kolumnowy *P*. Po założeniu pewnej wartości początkowej wektora *P*, nowy wektor *P′* jest równy:

$$P' = A^T P$$

Po wielokrotnym obliczeniu wartości wektora *P′* (podstawienie *P′* za *P* po każdym kroku), otrzymamy ostateczną wartość współczynników prestiżu, co jest tożsame z rozwiązaniem równania:

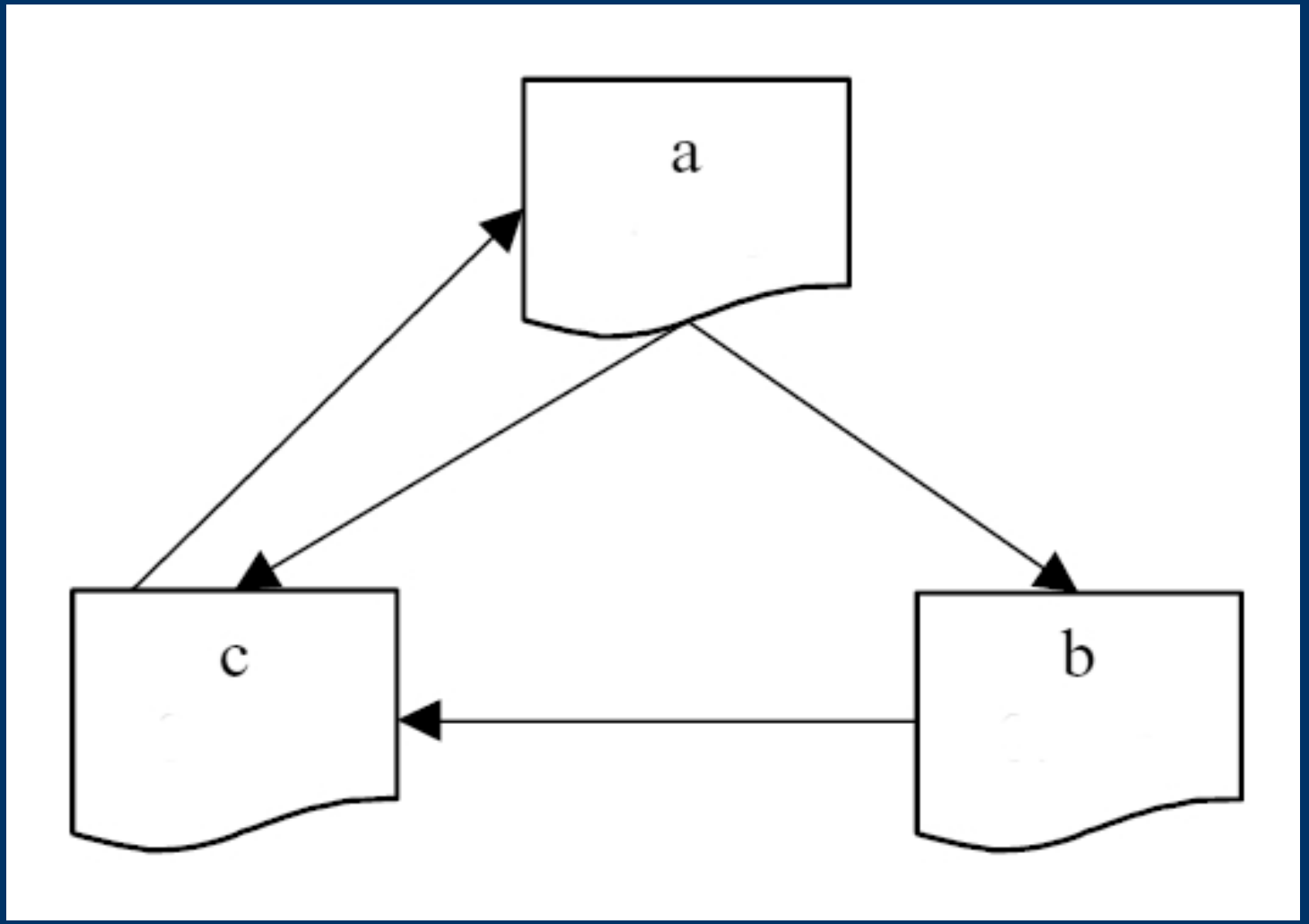
$$\lambda P = A^T P$$

Obliczanie współczynnika prestiżu

Rozwiązanie tego równania jest znane w algebrze liniowej jako problemem obliczania wartości własnych i wektorów własnych. Dla macierzy o rozmiarze *n* x *n*, istnieje *n* takich wektorów.

W przypadku obliczania współczynników prestiżu istotny jest wektor związany z największą wartością własną. Poszczególne współczynniki tego wektora określają wartości współczynników prestiżu powiązanych ze zindeksowanymi stronami WWW.

Współczynniki prestiżu – przykład



Współczynniki prestiżu – przykład

Załóżmy, że macierz sąsiedztwa *A* wygląda następująco:

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Po wykonaniu transpozycji wektor A^T:

$$A^T = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}$$

Współczynniki prestiżu – przykład

Po rozwiązaniu równania:

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

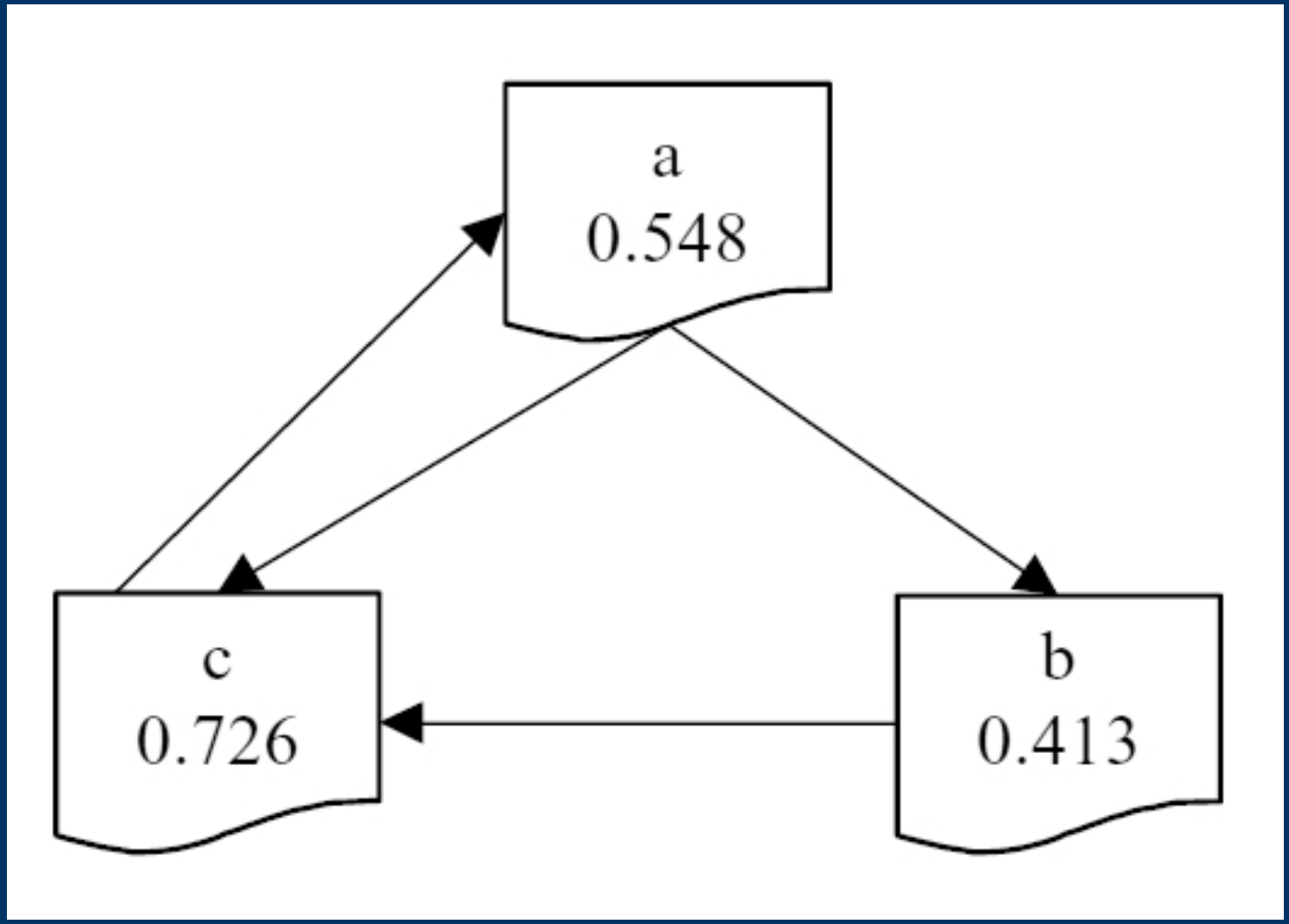
wektor współczynników prestiżu dla największej wartości własnej $\lambda = 1{,}325$ jest równy:

$P = (0{,}548 \;\; 0{,}414 \;\; 0{,}726)^T$

Po podstawieniu do wzoru:

$$1{,}325 \begin{pmatrix} 0{,}548 \\ 0{,}414 \\ 0{,}726 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0{,}548 \\ 0{,}414 \\ 0{,}726 \end{pmatrix}$$

Współczynniki prestiżu – przykład



Współczynniki prestiżu – przykład

Zgodnie z intuicją, dokument *c* uzyskał największą wartość współczynnika prestiżu, gdyż wskazują na niego linki z dwóch dokumentów.

Dokumenty *a* oraz *b* uzyskały niższe wartości współczynnika, gdyż są wskazywane tylko z jednego dokumentu.

Dokument *a* dostał większą wartość prestiżu niż dokument *b*, gdyż wskazuje na niego link z dokumentu o największej wartości współczynnika prestiżu.

Algorytm PageRank

Linki nie są jedynym wyznacznikiem propagowania współczynnika prestiżu. Ważne jest również to, w jaki sposób użytkownicy poruszają się pomiędzy stronami.

Do przybliżenia sposobu poruszania się pomiędzy stronami można wykorzystać model *Random Web Surfer*. Symuluje on działanie użytkownika, który w sposób losowy przenosi się pomiędzy stronami.

Model ten został po raz pierwszy wykorzystany w algorytmie *PageRank* w wyszukiwarce *Google*.

Algorytm PageRank

Istotną cechą algorytmu *PageRank* jest to, że korzysta on nie tylko z linków wskazujących na daną stronę, ale również uwzględnia linki wychodzące z danej strony (ang. *out-links*).

Załóżmy, że strona *u* zawiera N_u linków do innych stron (w tym do *v*).

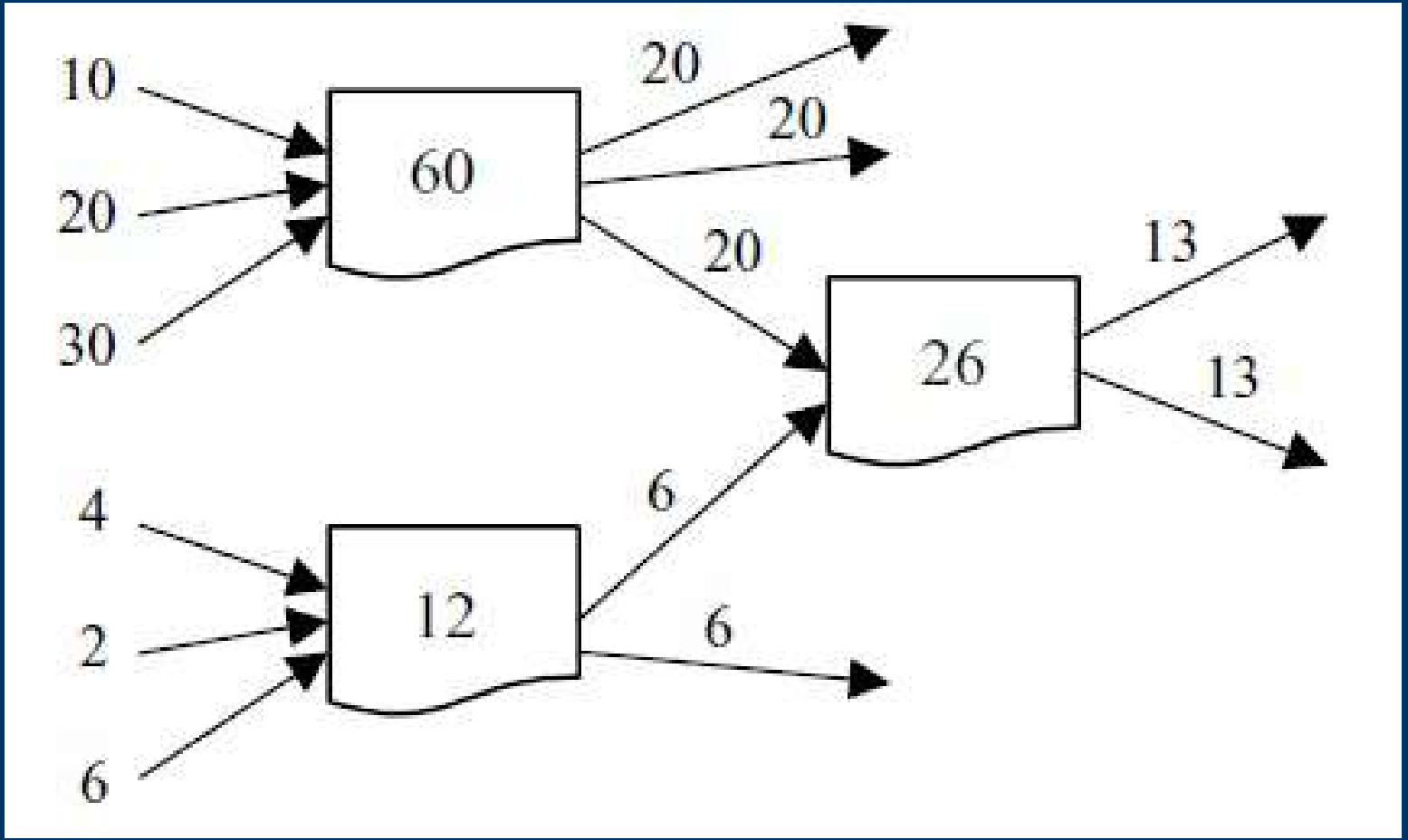
Prawdopodobieństwo przejścia do strony *v* jest równe $1 / N_u$.

Jaką wartość współczynnika prestiżu strony *u* dostaje strona *v*?

Wartość współczynnika prestiżu, który otrzymuje strona *v* jest równa $1 / N_u$ wartości współczynnika prestiżu strony *u*.

Algorytm PageRank

Propagacja współczynnika prestiżu w algorytmie *PageRank*:



Algorytm PageRank

Współczynniki prestiżu w algorytmie *PageRank* otrzymywane przez stronę *u* można opisać zależnością:

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

gdzie λ jest równa 1.

Wykorzystanie notacji macierzowej wymaga przyjęcia w macierzy sąsiedztwa wartości 1 / N_u, zamiast wartości 1 i 0 definiujących istnienie krawędzi.

Algorytm PageRank - przykład

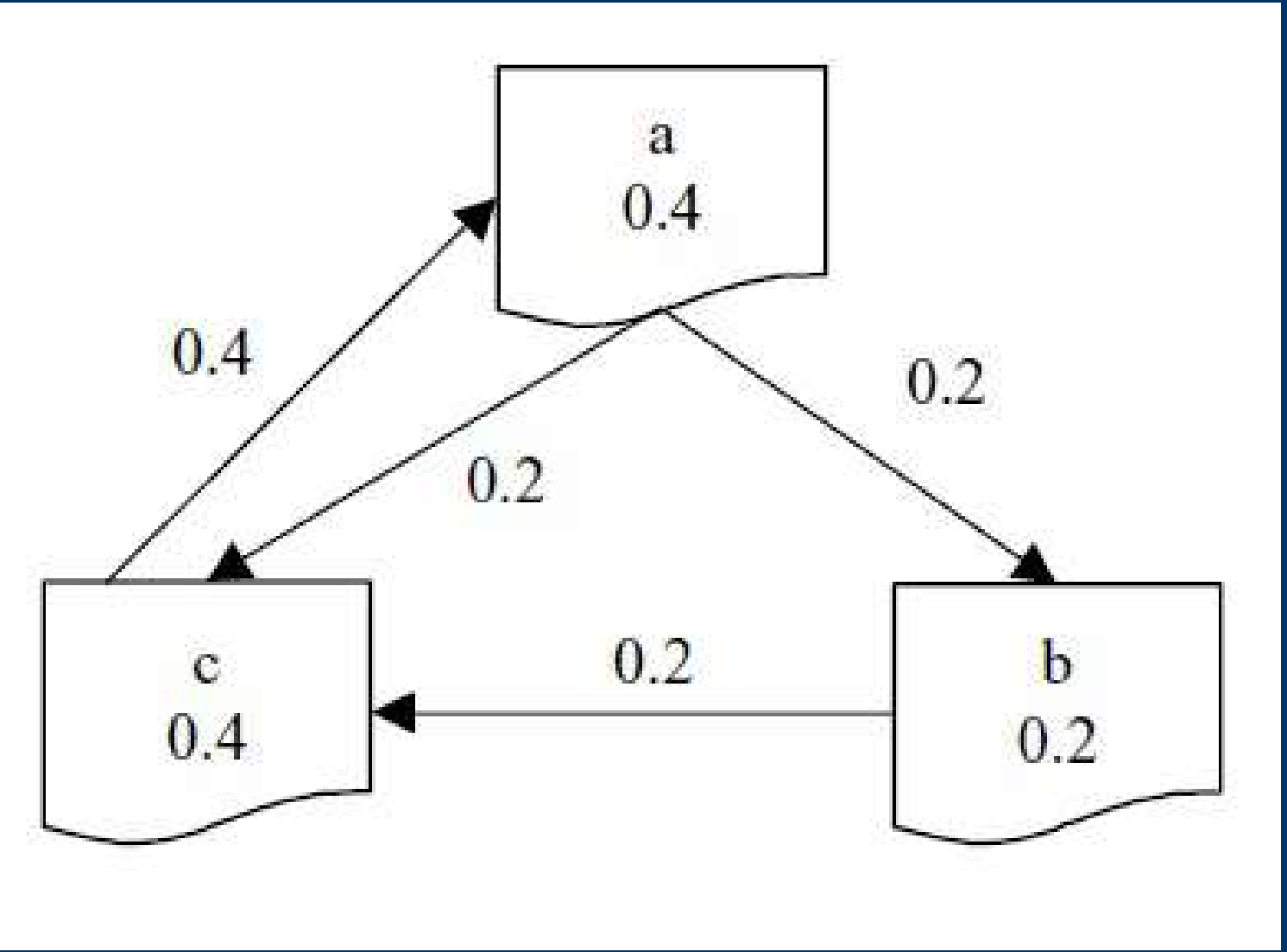
Po przeliczeniu macierzy *A*:

$$A = \begin{pmatrix} 0 & 0{,}5 & 0{,}5 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Rozwiązanie równania podanego na slajdzie 7 daje wynik:

$$\begin{pmatrix} 0{,}4 \\ 0{,}2 \\ 0{,}4 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 0{,}5 & 0 & 0 \\ 0{,}5 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0{,}4 \\ 0{,}2 \\ 0{,}4 \end{pmatrix}$$

Algorytm PageRank - przykład



Algorytm PageRank – pętle

Algorytm *PageRank* jest odporny na występowanie pętli, które składają się z dowolnej ilości stron, pod warunkiem, że pętle te posiadają linki, które pozwalają z nich wyjść.

Algorytm nie będzie sobie w stanie poradzić w sytuacji, gdy dwie strony będą posiadały linki wskazujące na siebie nawzajem bez możliwości opuszczenia takiej pętli.

Ten układ stron nazywany jest *rank sink*.

Algorytm PageRank – pętle

Istnienie pętli bez linków wyjściowych będzie powodowało sztuczne zwiększanie współczynnika prestiżu tych stron z uwagi na zastosowanie modelu *Random Web Surfer*.

Jak sobie poradzić z tym problemem?

Sposobem radzenia sobie z tą sytuacją jest wprowadzenie cechy „znudzonego surfera”. Polega to na ograniczeniu liczby przejść pomiędzy tymi samymi witrynami i przejściu na inną, losowo wybraną stronę.

Algorytm PageRank – wektor E

Wprowadzenie modelu „znudzonego surfera” wymaga modyfikacji zależności opisującej wartość współczynnika prestiżu algorytmu *PageRank* poprzez uwzględnienie współczynników wektora źródeł prestiżu *E*:

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

Wektor źródeł prestiżu *E* ma ilość współrzędnych równą ilości wszystkich zindeksowanych stron WWW i jego współczynniki reprezentują rozkład prawdopodobieństwa przejścia do losowego dokumentu sieci WWW.

Algorytm PageRank – wektor E

Współczynniki wektora E dobierane są w taki sposób, aby norma wektora E wynosiła ok. 0.15. Im wyższa norma, tym skoki do losowych stron występują częściej.

Przyjęcie współczynników, które generują zbyt dużą wartość normy dla wektora E powoduje zmniejszenie zależności wartości współczynników prestiżu w odniesieniu do rzeczywistej struktury sieci WWW.

Zbyt mała wartość normy przekłada się z kolei na zbyt długie zatrzymywanie się w układach stron typu *rank-sink*.

Algorytm PageRank – wektor E

Zastosowanie wektora *E* zabezpiecza również przed celowymi manipulacjami linkami powodującymi zwiększanie wartości współczynnika prestiżu.

Bez wektora *E* byłoby możliwe zwiększanie wartości współczynnika danej strony poprzez tworzenie linków ze stron o dużym prestiżu, bądź tworzenie wielu linków ze stron o małym prestiżu.

Dodatkową zaletą wykorzystania wektora *E* jest przechodzenie do stron, które tworzą podgrafy niepołączone z głównym grafem.

Algorytm PageRank – zastosowanie

Uniwersalność algorytmu *PageRank* pozwala na zastosowanie go również w innych aspektach badania sieci niż sam rozdział współczynników prestiżu. Inne zastosowania algorytmu:

- przewidywanie obciążenia stron WWW i serwerów HTTP,
- modelowanie zachowań użytkowników serwisów internetowych,
- optymalizacja przeszukiwania sieci WWW,
- wpływanie na częstość powtórnej indeksacji już odwiedzonych stron WWW.

Algorytm HITS

Pomimo tego, że istotność dokumentów wynikowych bazująca na wartościach prestiżu jest najpowszechniej wykorzystywana w systemach wyszukiwania danych w sieciach, to bazowanie tylko na tej wartości może stwarzać potencjalne problemy.

Nie zawsze strony, które posiadają najwyższe współczynniki prestiżu zawierają treści zgodne z zapytaniem, zaś dokumenty, które dokładnie wpasowują się zapytanie często mają bardzo niski współczynnik prestiżu.

Algorytm HITS

Ranking na bazie współczynników prestiżu musi iść w parze z rankingiem opierającym się na treści zawartej w dokumentach sieci WWW.

Przykładem takiego rozwiązania jest algorytm *HITS* (*Hyperlink Induced Topic Search*). Idea jego działania opiera się na:

- wyszukaniu i posortowaniu stron bazujące na treści dokumentów,
- bazując na zwróconych wynikach obliczane są współczynniki prestiżu.

Algorytm HITS

Schemat działania algorytmu *HITS* można opisać jako:

- z wykorzystaniem metod *IR* zwracany jest mały zbiór istotnych stron R_q,
- zbiór R_q jest rozszerzany o strony, które wskazują linki ze zbioru bazowego, a także o strony, z których wychodzą linki wskazujące na zbiór bazowy,
- na bazie utworzonego zbioru obliczane są współczynniki prestiżu,
- wykonywane jest sortowanie wyników z uwzględnieniem prestiżu oraz metod systemu *IR*.

Algorytm HITS

Takie podejście ma dużą wadę, ale jednocześnie dużą zaletę.

Dużą wadą algorytmu jest potrzeba każdorazowego przeliczenia współczynników prestiżu na bazie zwróconych wyników, co zwiększa koszt obliczeniowy rozwiązania.

Jest to jednocześnie duża zaleta, gdyż współczynniki prestiżu są przeliczane jedynie dla stron, które w sposób jednoznaczny wiążą się z zapytaniem złożonym ze słów kluczowych, dzięki czemu zbiór wynikowy jest lepiej posortowany.

Algorytm HITS

Dobrym przykładem porównania wyników zwracanych przez algorytmy *HITS* i *PageRank* jest zapytanie złożone ze słów kluczowych *music program*. *Google* bazujące na *PageRank* umieszcza na pierwszych miejscach strony odnoszące się do komputerów i muzyki, natomiast na dalszych pozycjach strony dotyczące programów muzycznych w radiu czy telewizji.

Skąd to wynika?
Strony zawierające treści powiązane z komputerami oraz oprogramowaniem mają duże wyższe wartości współczynników prestiżu niż strony dotyczące programów radiowych czy telewizyjnych.

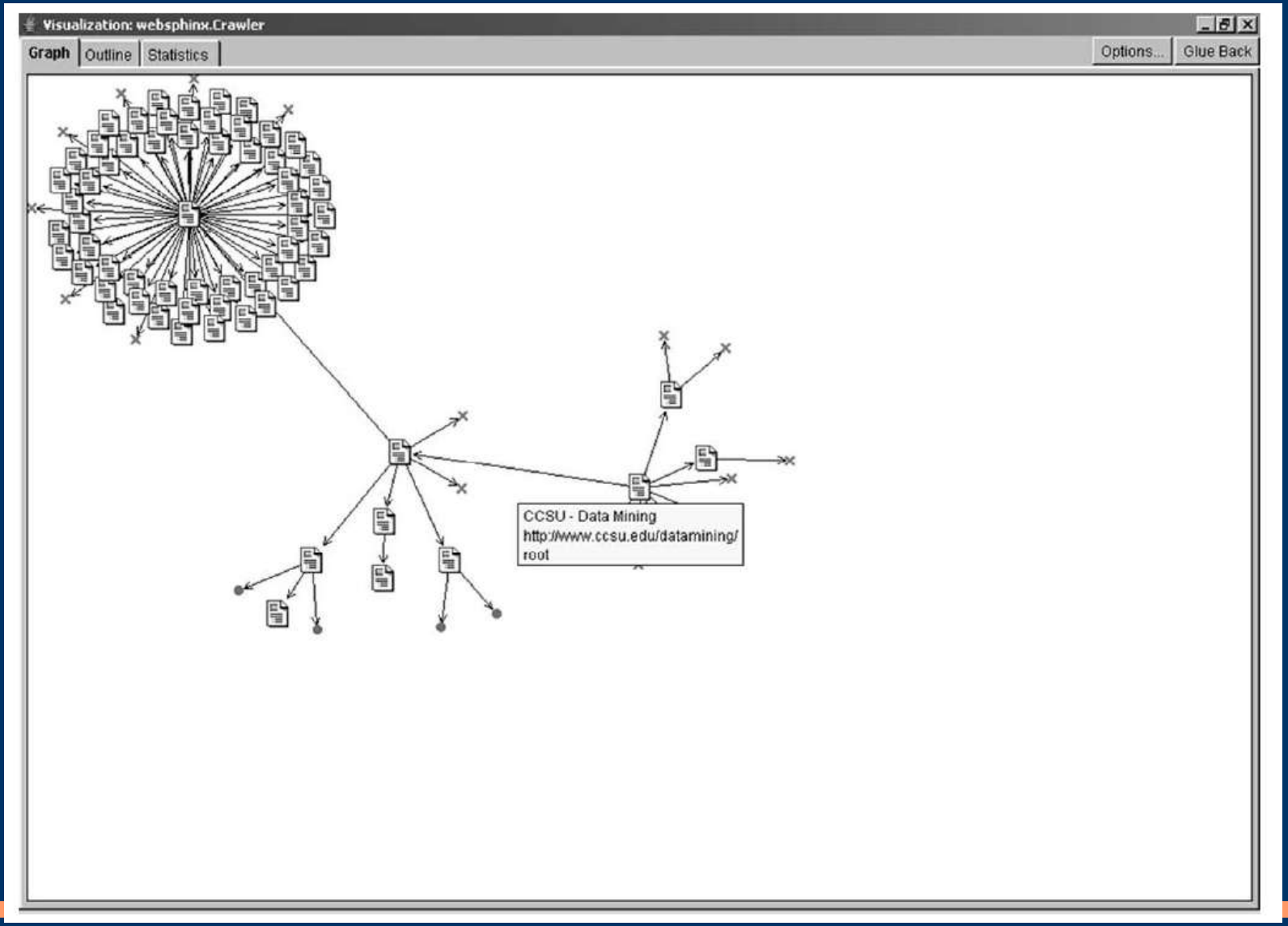
Wyniki na bazie linków

Badania pokazują, że zmiana tematu danej strony następuje średnio po przejściu przez dwa kolejne linki. Bazując na tej obserwacji, można zaproponować algorytm zwracający wyniki na bazie linków.

Jest to modyfikacja algorytmu *HITS*, opierająca się w dużej mierze na odnajdowaniu stron typu *authorities* i typu *hubs*. Strony typu *hub* są to strony zawierające duże ilości linków wychodzących, zaś strony typu *authorities* są stronami, na które wskazuje duża ilość linków.

Wyniki na bazie linków

Strona typu *hub* (lewa górna cześć grafu):



Wyniki na bazie linków

Schemat działania algorytmu przy zadanej stronie *u* oraz określonym parametrze *k* jest następujący:

- odnalezienie *k* stron wskazujących na stronę *u* oraz wykorzystanie ich do stworzenia zbioru głównego R_u,
- wygenerowanie zbioru bazowego S_u na bazie zbioru głównego R_u,
- odnalezienie stron typu *hub* i stron typu *authority* w zbiorze S_u,
- zwrócenie jako wyników stron typu *authority* i stron typu *hub* z najwyższymi współczynnikami prestiżu.

Wyniki na bazie linków

Zaletą algorytmu zwracającego wyniki na bazie linków jest umieszczanie w zbiorze wynikowym stron zawierających niewielką ilość tekstu oraz stron zawierających treść inną niż tekst (np. obrazki czy multimedia).

Dużą wadą takiego rozwiązania jest problem związany z poprawnym doborem strony startowej *u* dla całego algorytmu. Niewłaściwy dobór takiej strony powoduje zwrócenie wyników niezwiązanych z zapytaniem.

Podsumowanie

Użycie wartości współczynników prestiżu może być wykorzystywane w procesie sortowania stron pod kątem istotności zwróconych wyników zapytania. Jednak ograniczenie się tylko do tego typu mechanizmów powoduje problemy związane z jakością tworzonego zbioru wynikowego dokumentów.

Każda licząca się na rynku wyszukiwarka łączy poznane do tej pory mechanizmy w celu zwracania jak najlepszych wyników. Wiele wykorzystywanych mechanizmów jest jednak tajemnicą handlową poszczególnych korporacji.

Dziękuję za uwagę!