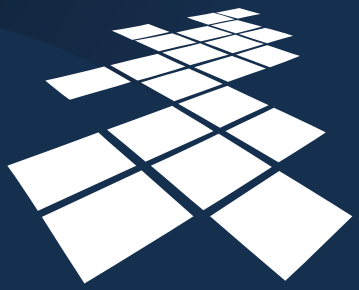
Zagadnienia zaawansowane

Wykład prowadzą:
Maciej Zakrzewicz
Mikołaj Morzy



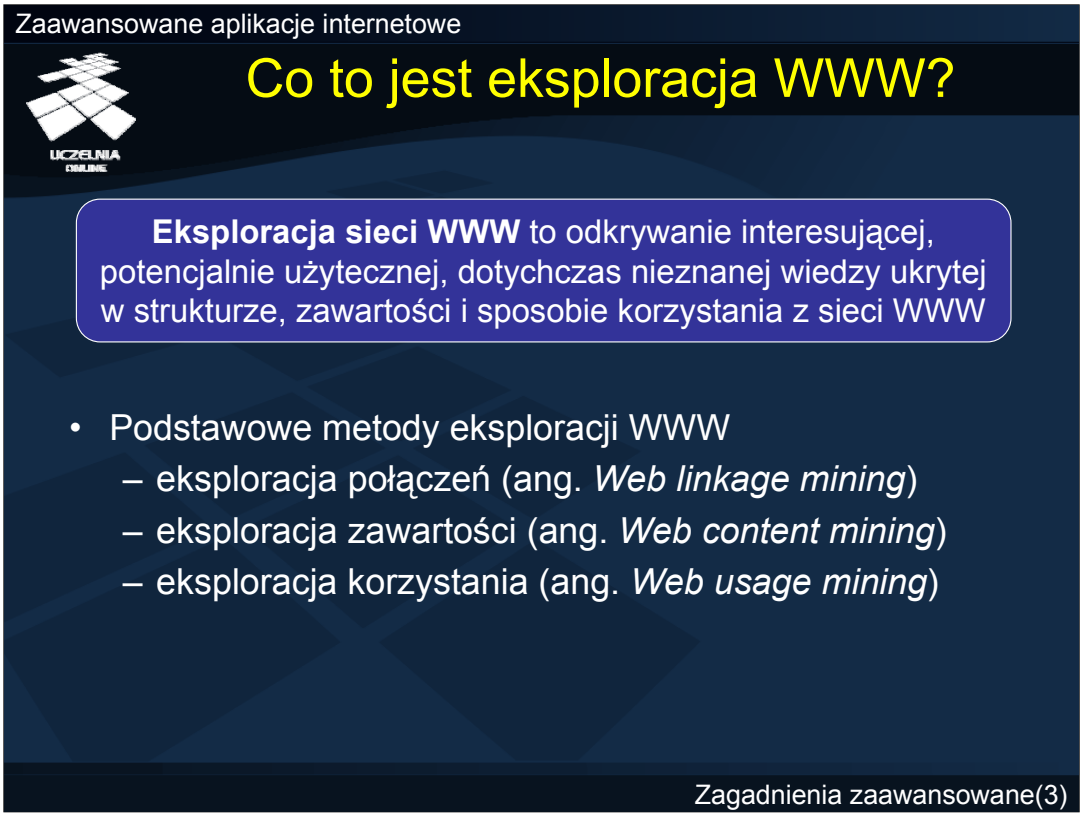
UCZELNIA
ONLINE

Zagadnienia zaawansowane. Wykład prowadzą: Maciej Zakrzewicz i Mikołaj Morzy

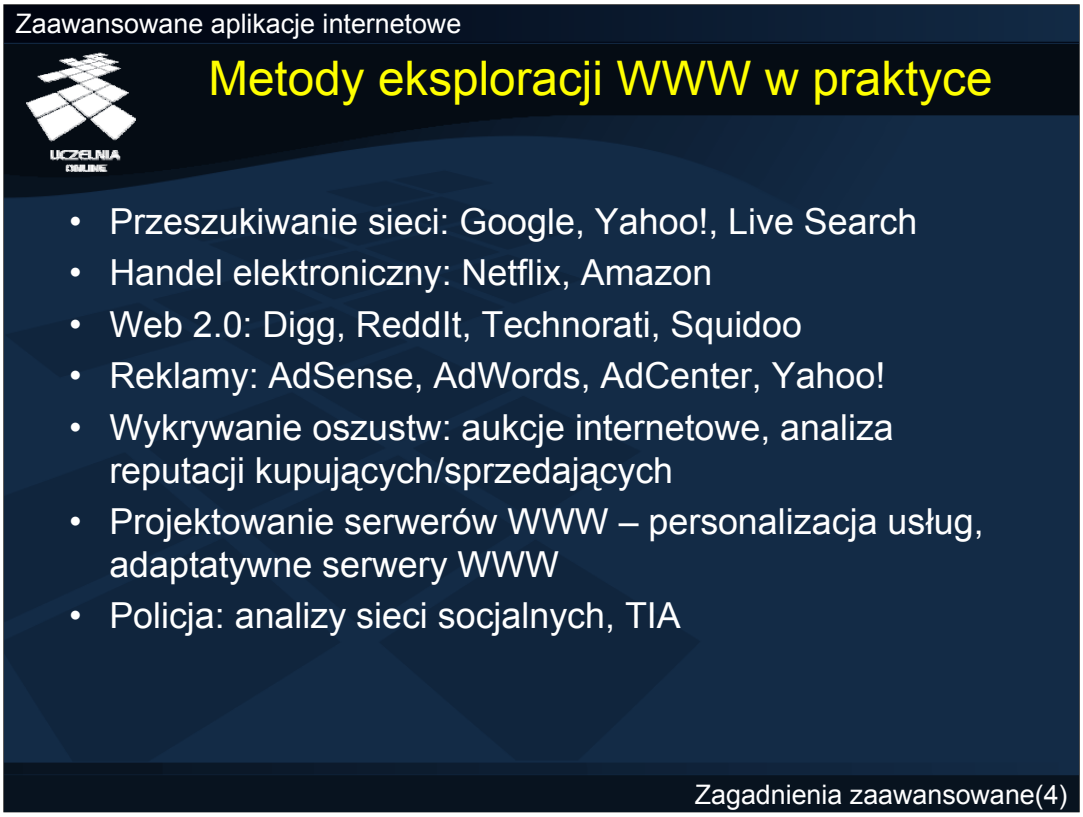
Plan wykładu

- Wprowadzenie do eksploracji sieci WWW
- Podstawy wyszukiwania
- Eksploracja połączeń
  - PageRank
  - Topic-Specific PageRank
  - Hubs & Authorities

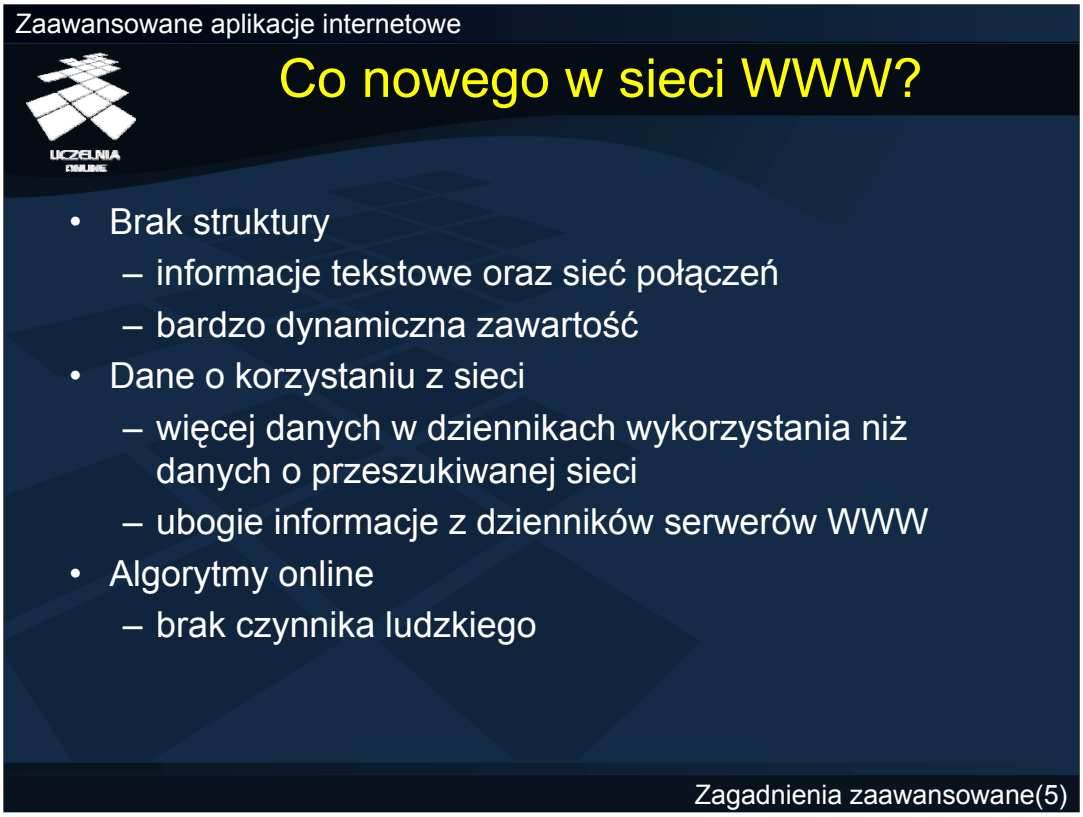
Celem wykładu jest przedstawienie zaawansowanych zagadnień związanych z aplikacjami internetowymi. Pierwsza część wykładu jest poświęcona eksploracji sieci WWW. Na wstępie przedstawione zostanie krótkie wprowadzenie do zagadnień eksploracji, ze szczególnym uwzględnieniem charakterystyki sieci WWW. Następnie, przedstawione zostaną podstawowe pojęcia związane z przeszukiwaniem sieci WWW. Pierwszą część wykładu zakończy prezentacja trzech różnych algorytmów rankingu stron WWW: algorytmu PageRank wykorzystywanego przez popularną wyszukiwarkę Google, algorytmu Topic-Specific PageRank, oraz algorytmu HITS, zwanego także "Hubs & Authorities".



Eksploracja sieci WWW (ang. Web mining) to odkrywanie nowej, interesującej, potencjalnie użytecznej i dotychczas nieznanej wiedzy ukrytej w strukturze, zawartości i sposobie korzystania z sieci WWW. Wiedza odkrywana w procesie eksploracji może być prezentowana za pomocą wielu różnych modeli, np. w postaci reguł, wzorców, wyjątków, czy zależności. Sieć WWW zawiera, poza użyteczną wiedzą, ogromną ilość szumu informacyjnego i śmieci. Pozyskanie użytecznej i wartościowej wiedzy z gigantycznego repozytorium, jakim jest sieć WWW, jest zadaniem trudnym. Metody eksploracji sieci WWW można z grubsza podzielić na trzy podstawowe kategorie. Metody eksploracji połączeń (ang. Web linkage mining) analizują strukturę połączeń i odnośników między dokumentami WWW w celu opracowania rankingu dokumentów. Metody te znajdują zastosowanie przede wszystkim w wyszukiwarkach internetowych. Drugą grupą metod są metody eksploracji zawartości dokumentów WWW (ang. Web content mining). Celem metod eksploracji zawartości jest automatyczne pozyskiwanie wiedzy o treści dokumentów. Zastosowania tych metod to, między innymi, automatyczne katalogi recenzji produktów i usług, agregatory wiadomości, itp. Ostatnią kategorią metod eksploracji sieci WWW są metody eksploracji sposobów korzystania z sieci WWW (ang. Web usage mining). Metody te mają na celu przede wszystkim analizę zawartości dzienników serwerów WWW w poszukiwaniu wzorców opisujących częste sekwencje odwołań do dokumentów przechowywanych na serwerze WWW. Ze względu na ograniczenia czasowe w trakcie tego wykładu zajmiemy się tylko pierwszą grupą, czyli algorytmami analizy połączeń między dokumentami WWW.



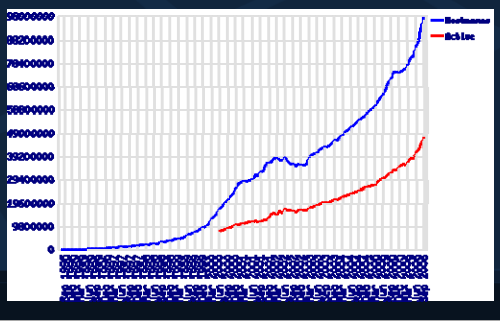
Jak wcześniej wspomniano, sieć Web jest olbrzymim repozytorium różnorodnej wiedzy. Eksploracja WWW pozwala na odkrycie tej wiedzy i wykorzystanie odkrytych wzorców w praktyce. Bieżący slajd przedstawia (rzecz jasna, niepełną) listę przykładów zastosowania wiedzy odkrytej za pomocą metod eksploracji WWW w praktyce. Bez wątpienia najbardziej udanym przykładem wykorzystania metod eksploracji WWW są wyszukiwarki internetowe. We wszystkich wiodących wyszukiwarkach (m.in. Google, Yahoo!, Live Search) zastosowano zaawansowane algorytmy rankingu dokumentów WWW. Automatyczne rekomendacje produkowane przez internetowe księgarnie (Amazon) i wypożyczalnie wideo (Netflix) są także dziełem algorytmów eksploracji WWW. Metody eksploracji danych wspierają także bardzo popularne ostatnimi czasy aplikacje Web 2.0, takie jak agregatory wiadomości (Digg, ReddIt, Squidoo) czy serwisy przeszukiwania blogów (Technorati). Nie bez znaczenia jest także, z racji ogromnego rozmiaru rynku, zastosowanie eksploracji WWW do optymalizacji wyświetlania reklam i linków sponsorowanych (Google: AdSense i AdWords, Microsoft: AdCenter, Yahoo!: Search Marketing). Inną dziedziną zastosowania metod eksploracji WWW jest ocena wiarygodności i reputacji uczestników aukcji internetowych (eBay) oraz analiza wiarygodności opinii wyrażanych publicznie (ePinions). Analiza sposobów wykorzystywania sieci WWW (a przede wszystkim analiza dzienników serwera WWW) umożliwia personalizowanie usług i tworzenie tzw. adaptatywnych serwerów WWW, które potrafią dynamicznie generować zawartość zgodnie z preferencjami indywidualnych użytkowników. Dziedziną, w której eksploracja jest stosowana od dawna, jest bezpieczeństwo narodowe. Metody eksploracji umożliwiają analizę sieci socjalnych, a co za tym idzie, umożliwiają skuteczną walkę z przestępczością. W zakresie bezpieczeństwa narodowego w ostatnich latach największy rozgłos uzyskał amerykański projekt Total Information Awareness (TIA), polegający na poddaniu praktycznie całej działalności internetowej skrupulatnej kontroli.



Wiele osób postrzega sieć WWW jako gigantyczną, heterogeniczną bazę danych. Wydawać by się mogło, że można do niej zastosować tradycyjne metody eksploracji danych. W rzeczywistości jednak sieć WWW posiada pewne cechy, które uniemożliwiają proste przeniesienie metod eksploracji z baz danych i hurtowni danych do środowiska WWW. Przede wszystkim, w przeciwieństwie do tradycyjnych składnic danych, dokumenty w sieci WWW nie posiadają dobrze określonej struktury, lub posiadają tylko częściową strukturę (tzw. dane semistrukturalne). Wartościowe informacje są ukryte w dokumentach WWW zarówno w ich zawartości (najczęściej są to informacje tekstowe), jak i w strukturze połączeń między danym dokumentem a innymi dokumentami. Dodatkowo, zawartość wielu dokumentów jest dynamiczna i zmienia się w czasie. Ilość informacji o wykorzystaniu sieci WWW również stanowi istotny problem. Wg firmy Google, dzienniki wykorzystania są obszerniejsze niż indeks odwiedzonych stron WWW (!). W jednym z doniesień pracownicy firmy Google oceniali rozmiar dziennego przyrostu dzienników wykorzystania na porównywalny z rozmiarem sporej hurtowni danych. Z drugiej strony, informacje trafiające do dzienników wykorzystania są stosunkowo ubogie i nie zawierają dużo użytecznych danych. Wreszcie, czynnikiem znacznie utrudniającym prostą adaptację metod eksploracji danych do środowiska WWW jest częste wymaganie, aby algorytmy eksploracji sieci WWW działały w trybie online, bez nadzoru i bez udziału czynnika ludzkiego. Dobrym przykładem takich algorytmów są algorytmy alokacji reklam do wyświetlanych stron.

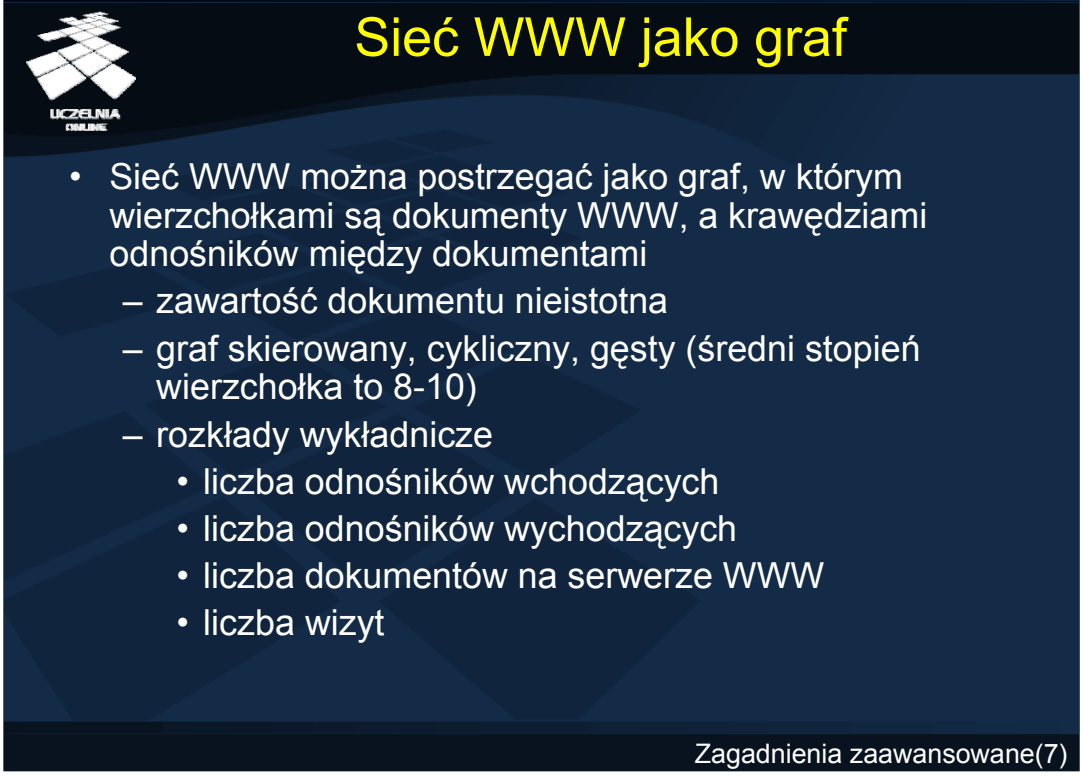
Rozmiar sieci WWW

- Liczba dokumentów WWW
  - praktycznie nieskończona, bardzo dużo duplikatów (30%-40%)
  - Google: 8 miliardów, Yahoo!: 20 miliardów
- Liczba serwerów WWW
  - Netcraft: 96 854 877 (wrzesień 2006)



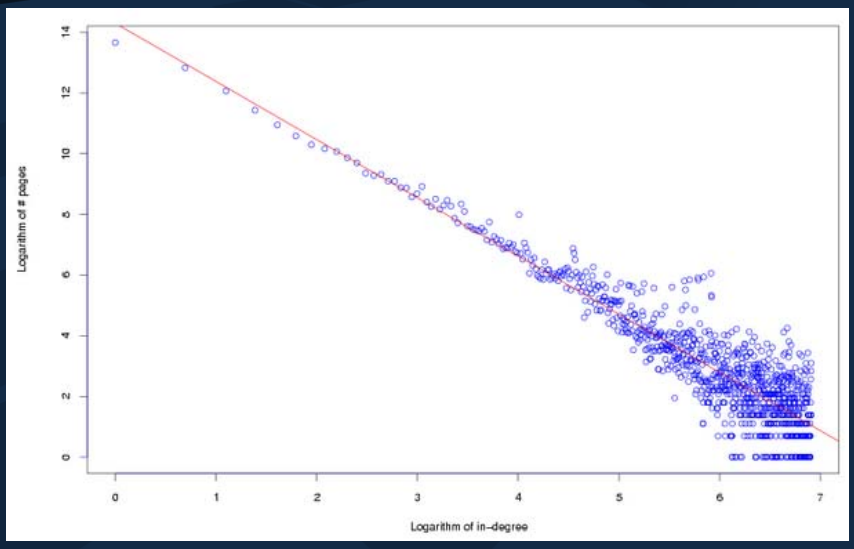
źródło: news.netcraft.com

Jak duża jest sieć WWW, która podlega eksploracji? Liczba dokumentów WWW jest w praktyce nieskończona, przede wszystkim ze względu na to, że wiele dokumentów jest dynamicznie generowanych. Jednak bardzo duża część dokumentów (szacuje się, że liczba to sięga nawet 40%) jest duplikatami innych dokumentów. Najlepszym przybliżeniem rzeczywistej liczby dokumentów WWW są rozmiary indeksów największych wyszukiwarek internetowych. Yahoo! podaje, że ich indeks obejmuje 20 miliardów statycznych dokumentów HTML, Google szacuje liczbę poindeksowanych dokumentów HTML na ok. 8 miliardów. Znacznie bardziej wiarygodne są szacunki liczby różnych serwerów WWW. Wg najnowszego sondażu Netcraft, we wrześniu 2006 było zarejestrowanych ponad 96 milionów serwerów (gwałtowny przyrost liczby serwerów w okresie od lipca do września 2006 jest związany z uruchomieniem platformy Microsoft Life Spaces).



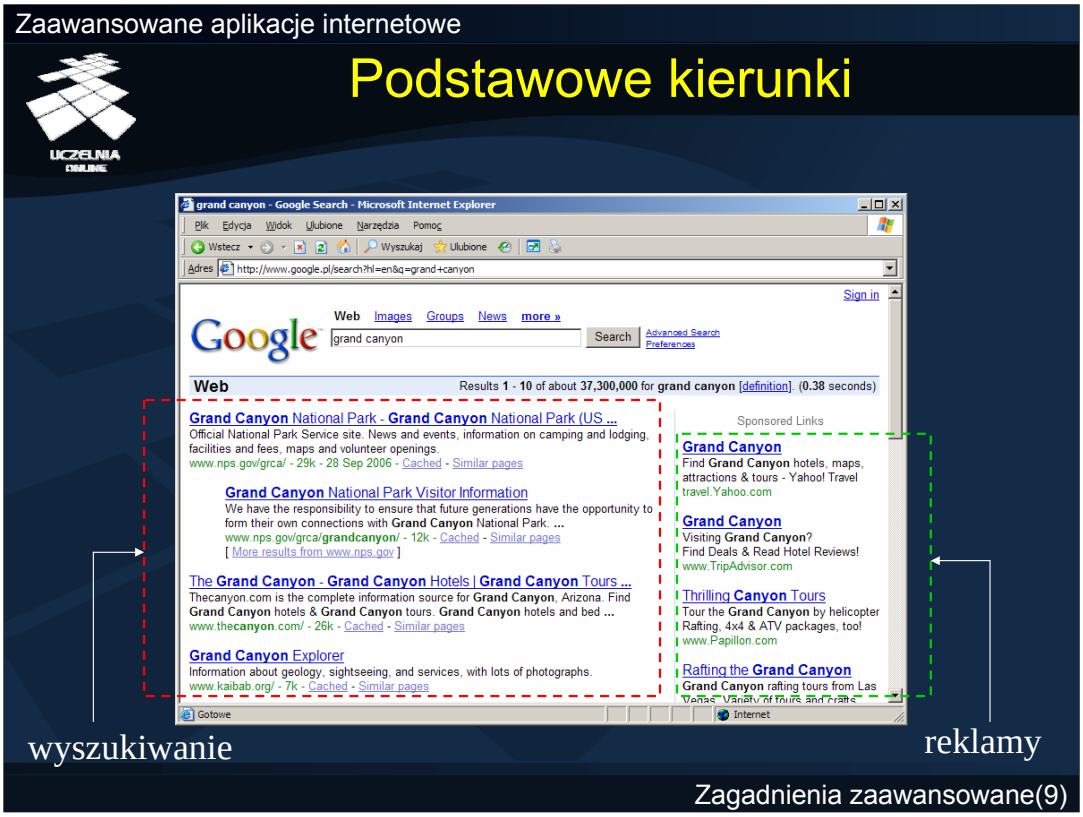
W wielu zastosowaniach, w szczególności w wielu algorytmach eksploracji sieci WWW, sieć WWW przybliżamy za pomocą prostego modelu grafowego, w którym wierzchołkom grafu odpowiadają dokumenty WWW, a krawędziom grafu odpowiadają odnośniki między poszczególnymi dokumentami. W takim modelu zawartość dokumentu (czyli wierzchołka) jest nieistotna, a liczy się tylko topologia połączeń między wierzchołkami. Graf modelujący sieć WWW jest grafem skierowanym cyklicznym. Co ciekawe, graf ten jest bardzo gęsty, gdyż, jak pokazuje wiele opracowań i eksperymentów, średnia liczba odnośników w dokumencie HTML waha się między 8 a 10. Wiele zjawisk występujących w sieci WWW modelujemy za pomocą rozkładów wykładniczych. Przykładowo, liczba odnośników wychodzących z dowolnego dokumentu HTML, liczba odnośników wskazujących na dokument HTML, liczba dokumentów przechowywanych na serwerze WWW, a nawet liczba wizyt mogą być z dużą dokładnością zamodelowane przy pomocy rozkładu wykładniczego. Na kolejnym slajdzie przedstawiono przykład rozkładu wykładniczego obrazującego liczbę odnośników wchodzących.

Rozkład wykładniczy liczby odnośników wchodzących

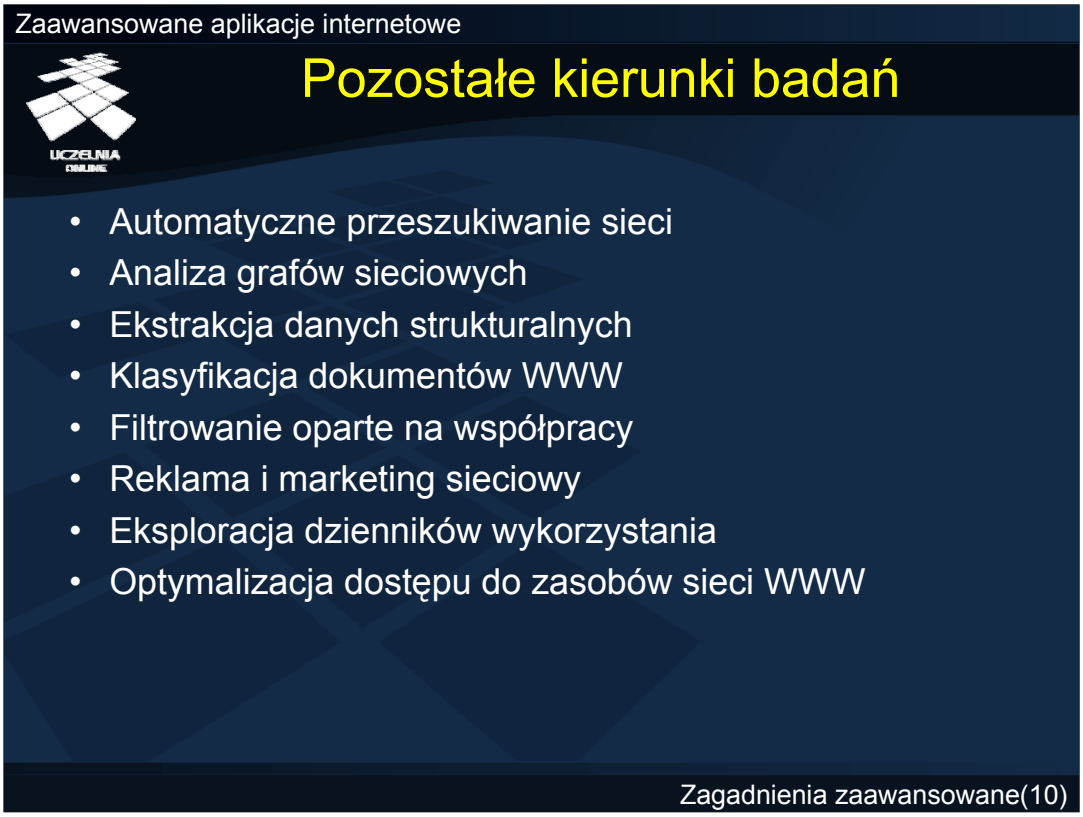


źródło: www2002.org

Na slajdzie przedstawiono rozkład wykładniczy liczby odnośników wchodzących (czyli wskazujących na dany dokument). Należy zwrócić uwagę na to, że współrzędne na obu osiach są wykreślone w skali logarytmicznej. Interpretacja wykresu jest następująca: znakomita większość dokumentów posiada bardzo niewielką liczbę odnośników wchodzących (punkty w lewym górnym rogu wykresu), podczas gdy bardzo niewielka część dokumentów posiada ogromną liczbę odnośników wchodzących (skupisko punktów w prawym dolnym rogu wykresu).



Dwa podstawowe kierunki rozwoju algorytmów eksploracji sieci WWW to wyszukiwanie i reklamy. Dochody z reklam stanowią podstawę działalności biznesowej największych wyszukiwarek. Poza budowaniem ogromnych indeksów dokumentów HTML oraz dopasowywaniem poindeksowanych dokumentów do kryteriów wyszukiwania, niebagatelne znaczenie ma także alokacja reklam do wyświetlanych wyników wyszukiwania. Aktualnie wykorzystywany jest model reklamowy opracowany przez firmę Overture Services (przejętą później przez Yahoo! Search Marketing) i wykorzystujący mechanizm mikroaukcji, gdzie uczestnikami aukcji są reklamodawcy, a oferowanym dobrem jest "impresja" reklamy w dokumencie HTML. Reklamodawca płaci za impresję tylko i wyłącznie w przypadku, gdy reklamowany odnośnik faktycznie zostanie kliknięty. Stąd, wyszukiwarka musi w taki sposób dobierać reklamy do wprowadzonych przez użytkownika kryteriów wyszukiwania, aby maksymalizować prawdopodobieństwo kliknięcia na link przy uwzględnieniu kwoty, jaką reklamodawca jest gotów zapłacić za pojedyncze kliknięcie. Z punktu widzenia reklamodawcy problem polega na tym, z jakimi słowami kluczowymi związać swoje reklamy oraz ile oferować za pojedyncze kliknięcie na reklamę. Niestety, ten prosty model jest bardzo podatny na manipulację i oszustwo, w szczególności bardzo łatwo jest generować "puste" kliknięcia na reklamy konkurencji. Aktualnie trwają bardzo wytężone prace nad identyfikacją fałszywych kliknięć w reklamowane linki oraz nad ulepszaniem protokołów mikroaukcyjnych wybierających reklamy właściwe dla kontekstu wyszukiwania. Na marginesie warto wspomnieć tutaj o interesującej dyskusji prawnej, która rozgorzała wokół przedstawionego wyżej modelu wyświetlania reklam. Firma Geico podała do sądu firmę Google twierdząc, że nazwa "Geico" jest nazwą zastrzeżoną, stąd nie może być sprzedawana jako słowo kluczowe innym reklamodawcom. Wyrok sądowy ustalił, że wyszukiwarki mogą sprzedawać reklamy dla słów kluczowych będących zastrzeżonymi znakami towarowymi, natomiast nie jest jeszcze ustalone, czy same sprzedawane reklamy mogą zawierać (np. w postaci fragmentu adresu URL) nazwy innych zastrzeżonych towarów.

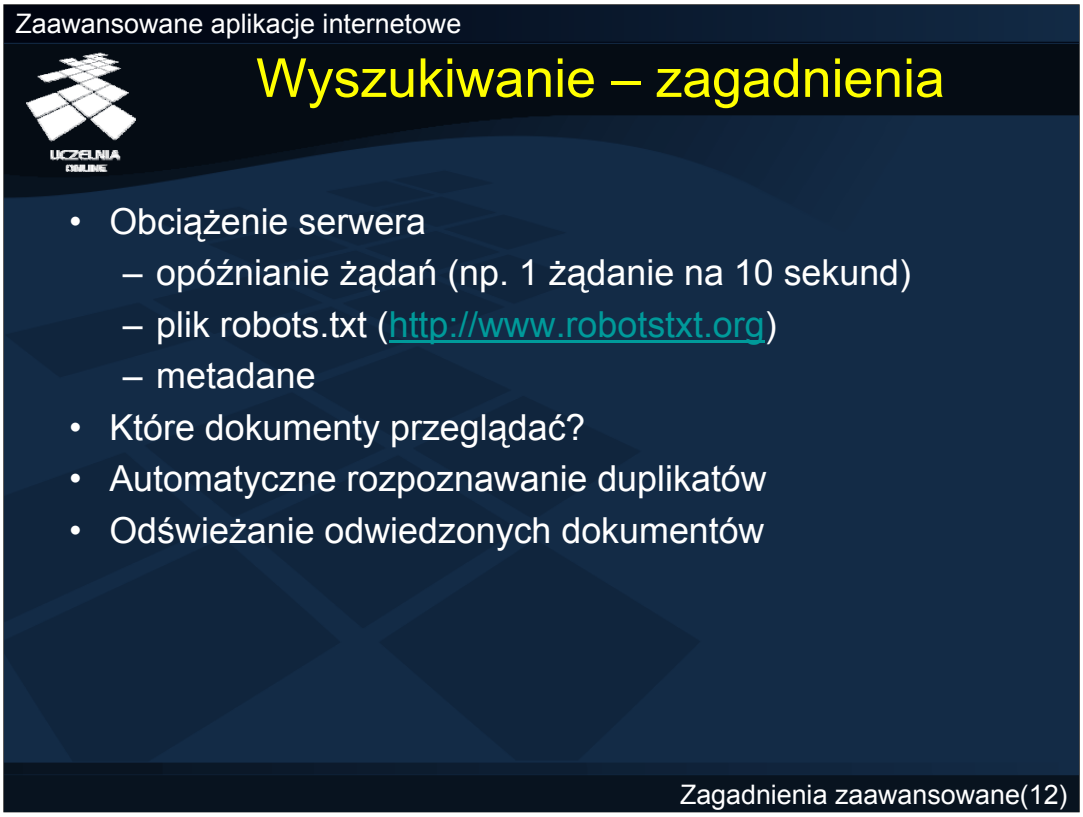


Poza kierunkami przedstawionymi na poprzednim slajdzie w zakresie eksploracji sieci WWW można pokusić się o wyróżnienie następujących kierunków badań. Automatyczne przeszukiwanie sieci obejmuje zagadnienia związane z pracą tzw. pajączków (ang. crawlers), czyli agentów indeksujących i przeszukujących dokumenty WWW. W ramach analizy grafów sieciowych bada się topologie połączeń, sieci powiązań socjalnych, itp. Ekstrakcja danych strukturalnych obejmuje wszystkie zagadnienia dotyczące automatycznego odkrywania i odczytywania informacji zawartej w strukturalnych dokumentach WWW, a w szczególności, w dokumentach XML. Jednym z najlepszych przykładów takich zastosowań jest bardzo prężnie rozwijający się rynek usług opartych na standardach RSS i Atom. Automatyczne etykietowanie i rozpoznawanie dokumentów WWW to elementy klasyfikacji dokumentów WWW. W ramach tej grupy zagadnień warto wymienić choćby systemy dokonujące automatycznego podsumowywania wyników wyszukiwania czy filtry niechcianej zawartości. Bardzo intensywnie bada się możliwość efektywnego wykorzystania metod eksploracji do celów marketingowych, m.in. dotyczy to prób wyszukiwania w sieci WWW autorytetów odpowiedzialnych za promowanie produktów i usług. Osobną gałąź badań stanowią metody odkrywania wiedzy o sposobach i typach wykorzystania sieci WWW czerpane z dzienników wykorzystania (tzw. weblogs). Metody te wykorzystuje się przede wszystkim do automatycznej personalizacji zawartości serwisów oraz do budowy tzw. adaptatywnych serwerów WWW, które potrafią automatycznie dostosowywać się do oczekiwań użytkowników pod względem rozkładu zawartości, połączeń między poszczególnymi częściami serwisu, itp.

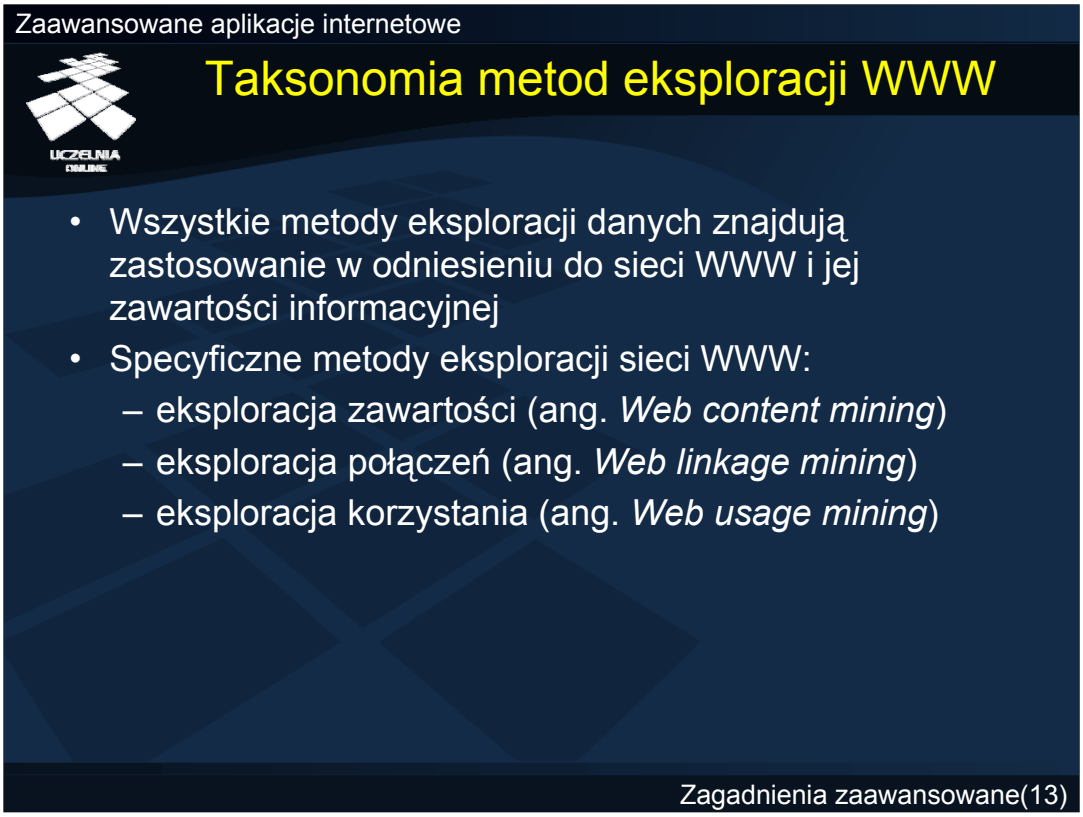
Wyszukiwanie – podstawowe pojęcia

- Pająk (ang. *spider*, *crawler*, *agent*)
  - rekurencyjne przeszukiwanie sieci WWW
  - ręczna aktualizacja katalogu
- Moduł indeksujący (ang. *indexer*)
  - lematyzatory, filtry, konwertery, indeksy odwrócone
- Moduł przetwarzania zapytań (ang. *query processor*)
  - asystent zapytań, transformator, baza danych dokumentów i reklam

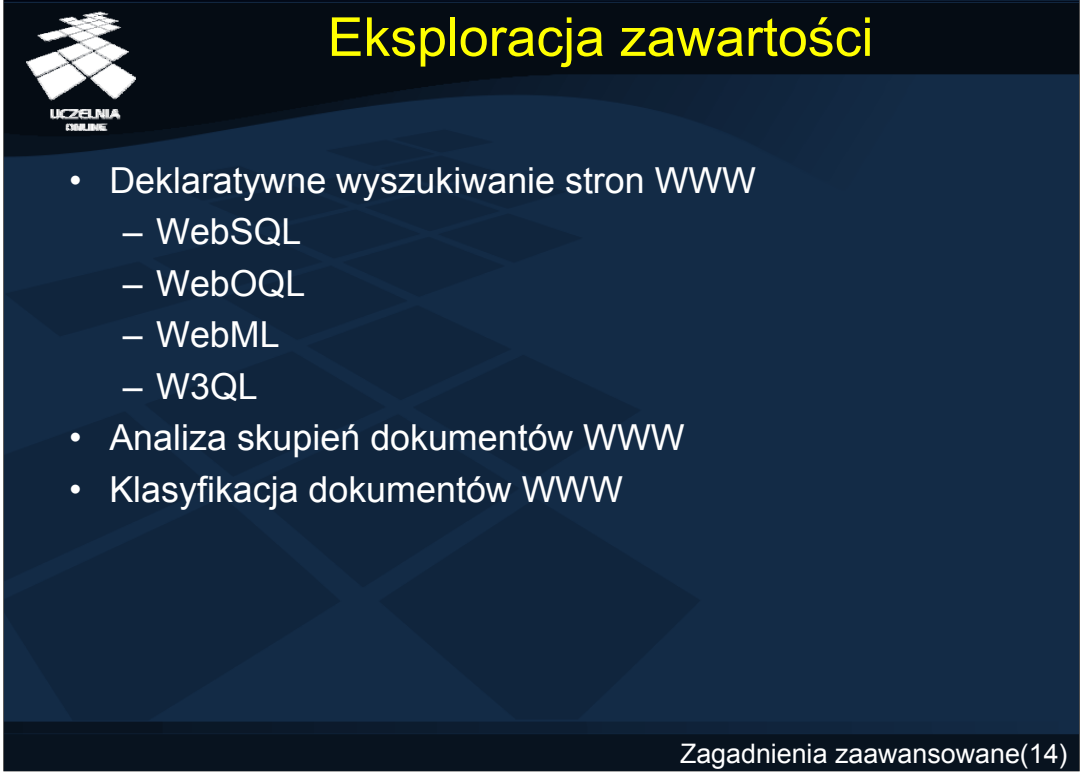
Slajd przedstawia podstawowe elementy składające się na wyszukiwarkę internetową. Na najniższym poziomie działa agent zwany pająkiem (ang. spider, crawler, agent). Jest to rozproszony program dokonujący rekurencyjnego przeszukiwania sieci WWW. Najczęściej, procedura wyszukiwania jest następująca: proces nadrzędny pająka rozpoczyna pracę od odczytania i sparsowania dokumentu za pomocą znanego początkowego adresu URL. Z początkowego dokumentu zostają odczytane kolejne adresy URL odnośników. Adresy te trafiają do kolejki, z której są asynchronicznie pobierane przez współbieżne procesy usługowe pająka. Każdy z procesów usługowych pobiera kolejny adres URL, odczytuje dokument i analizuje dokument w poszukiwaniu nowych adresów. W rzadkich przypadkach możliwa jest także ręczna aktualizacja katalogu adresów URL. Drugim modułem wyszukiwarki jest moduł indeksujący. Jest to moduł odpowiedzialny za znalezienie w odczytanych dokumentach termów (np. za pomocą lematyzatora), zastosowanie filtrów (np. odrzucenie słów pospolitych występujących na stop-listach), konwersję między różnymi stronami kodowymi, i wreszcie za stworzenie indeksów odwróconych. Trzecim modułem wyszukiwarki jest moduł przetwarzania zapytań, odpowiedzialny za podpowiedzi (wiele wyszukiwarek modyfikuje słowa kluczowe podawane przez użytkownika), transformacje zapytania, oraz za fizyczne wykonanie zapytania, odczytanie listy właściwych trafień w kolejności określonej przez ranking, oraz dopasowanie zbioru odnośników reklamowych do kontekstu zapytania.



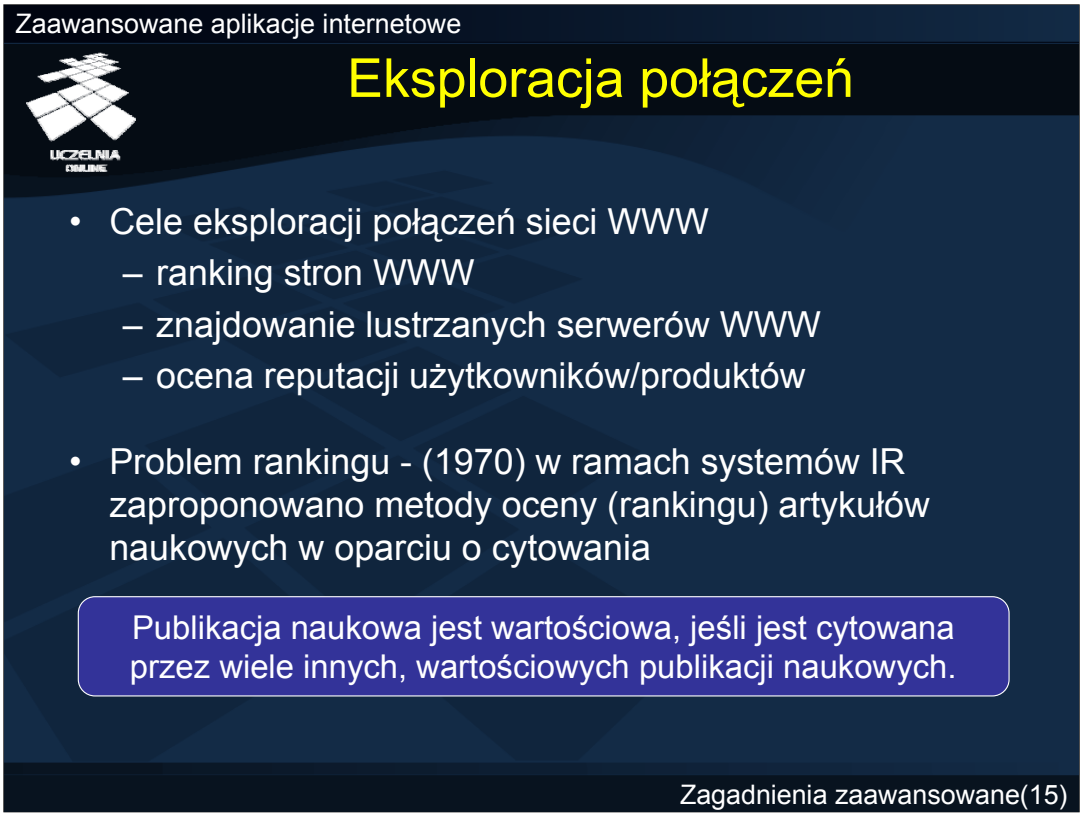
Przeszukiwanie sieci WWW nastręcza wiele problemów. Podstawowy problem, ściśle związany z "netykietą", to generowanie dodatkowego obciążenia odwiedzanych serwerów. Zakłada się, że automatyczny pająk nie powinien nadmiernie obciążać obcych serwerów i zaleca się, aby programistycznie odciążać obce serwery, np. przez opóźnianie kolejnych żądań do danego serwera. Dodatkowo, dobrym zwyczajem jest akceptowanie dyrektyw umieszczonych w pliku robots.txt. Jest to prosty plik tekstowy implementujący protokół Robots Exclusion Protocol i zawierający informację o tym, czy i które części serwera WWW są dostępne dla automatycznych pająków. Alternatywą dla stosowania pliku robots.txt jest wykorzystanie znacznika <meta> z atrybutami name='robots' i contents="nofollow|noindex". Ponieważ fizycznie nie jest możliwe przejrzenie całej sieci WWW, automatyczny pająk musi zdecydować, które fragmenty sieci należy indeksować. Jak już wcześniej wspomniano, bardzo duża część dokumentów WWW to duplikaty występujące w identycznej postaci w wielu różnych miejscach sieci (np. serwery lustrzane, tzw. mirrory). Rozpoznawanie takich dokumentów jest zadaniem trudnym i kosztownym. Jeszcze inna trudność wiąże się z koniecznością okresowego odświeżania informacji zebranych podczas pracy pająka. Dla poszczególnych serwerów należy ustalić częstotliwość odświeżania, a ta z kolei zależy od tempa zmian zachodzących w dokumentach źródłowych.



Jak już wspomnieliśmy, praktycznie wszystkie metody eksploracji danych znajdują zastosowanie w odniesieniu do sieci Web i jej zawartości informacyjnej. Niemniej, w literaturze wyróżnia się trzy podstawowe grupy metod eksploracji sieci Web: eksploracja zawartości sieci Web (ang. Web content mining), eksploracja połączeń sieci Web (ang. Web linkage mining), oraz eksploracja korzystania z sieci Web (ang. Web usage mining). W trakcie bieżącego wykładu pokrótce omówimy zarys metod eksploracji zawartości cieci i skupimy się przede wszystkim na metodach eksploracji połączeń. Z racji ograniczonego czasu w bieżącym wykładzie pominiemy zagadnienia związane z eksploracją korzystania z sieci.



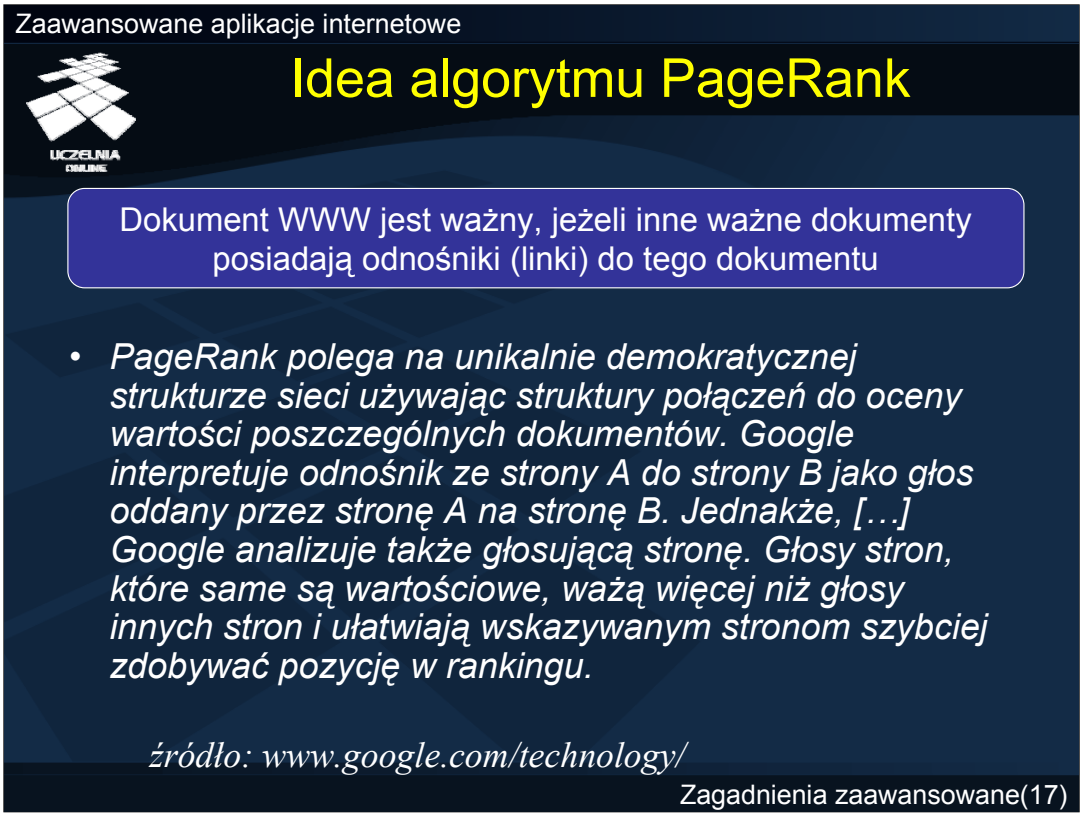
Większość zawartości dokumentów WWW do zawartość tekstowa z pewnymi elementami struktury. Część dokumentów WWW charakteryzuje się dobrze określoną strukturą (dotyczy to przede wszystkim dokumentów XML wykorzystujących DTD lub XMLSchema), natomiast dokumenty HTML zazwyczaj posiadają śladową strukturę (np. wyróżniony tytuł i autor dokumentu, nagłówki, paragrafy). Stąd metody eksploracji zawartości dokumentów WWW są najczęściej adaptacją tradycyjnych metod eksploracji tekstu, wzbogaconych o możliwość wykorzystania informacji o elementach struktury dokumentu. Przykładami takich metod są metody analizy skupień (ang. clustering) dokumentów w celu znalezienia grup dokumentów podobnych, lub metody automatycznej klasyfikacji dokumentów WWW. W obu przypadkach zachodzi konieczność opracowania nowych miar podobieństwa między dokumentami tekstowymi. Oryginalnym kierunkiem badań w zakresie eksploracji zawartości jest kwestia deklaratywnego wyszukiwania interesujących stron WWW. W ramach prowadzonych badań zaproponowano wiele języków wyszukiwania, najczęściej wzorowanych na języku SQL. Spośród wielu propozycji warto wymienić WebSQL i WebOQL (opracowane na Uniwersytecie Toronto), WebML (wzorowany na języku UML język modelowania zawartości dokumentów WWW), czy W3QL (język rozwijany przez Israel Institute of Technology).



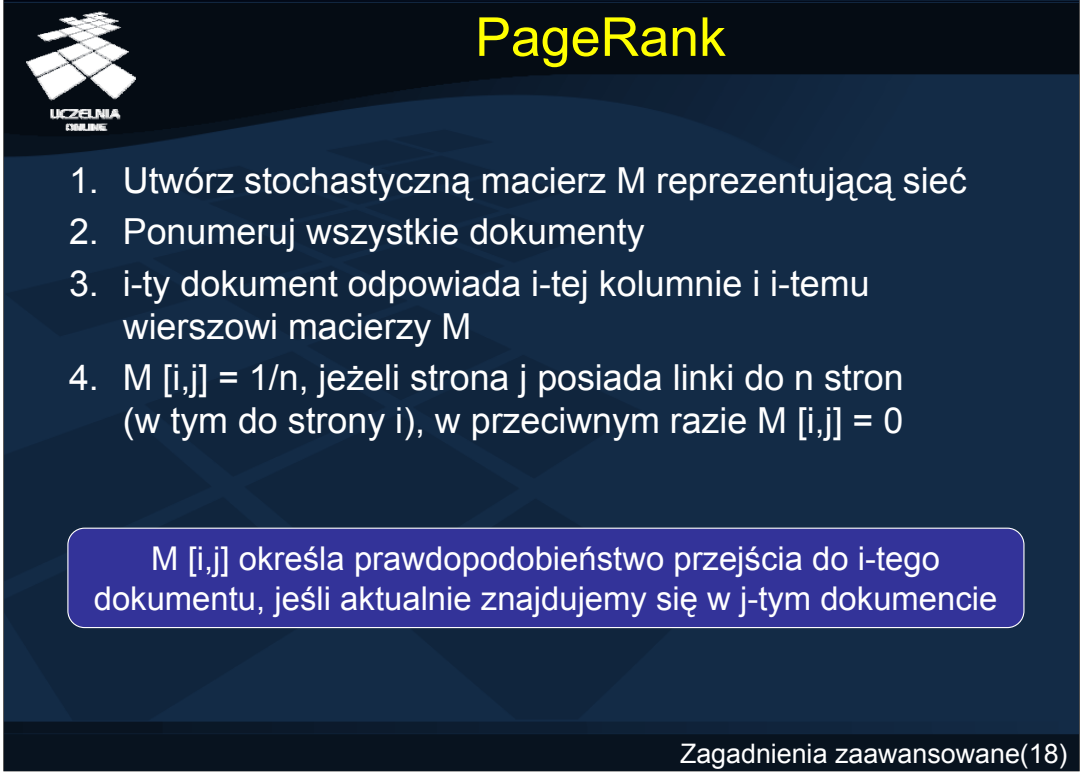
Druga grupa metod eksploracji sieci WWW wiąże się z eksploracją struktury połączeń. Początkowo, celem badań w zakresie eksploracji połączeń sieci WWW było opracowanie algorytmów umożliwiających przeprowadzenie rankingu dokumentów WWW. Okazało się jednak, że opracowane techniki są przydatne również w innych dziedzinach zastosowań. Algorytmy eksploracji sieci połączeń można wykorzystać do znajdowania lustrzanych serwerów WWW, co pozwala na implementację bardziej elastycznych optymalizatorów zapytań dla sieci rozległych, do oceny wiarygodności uczestników aukcji internetowych czy też do konstrukcji systemów rekomendacyjnych.

Algorytmy eksploracji struktury połączeń sieci WWW zostaną zilustrowane za pomocą trzech popularnych algorytmów (PageRank, Topic-Specific PageRank i H&A), których podstawowym zadaniem jest ranking, czyli ocena względnej ważności, dokumentów WWW. Problem rankingu jest znany od wielu lat i występuje w wielu dziedzinach zastosowań. Punktem wyjścia dla obu wspomnianych algorytmów rankingu stron były prace prowadzone w ramach systemów IR (ang. Information Retrieval) nad rankingiem publikacji naukowych. Na początku lat 70-tych zaproponowano metody oceny (rankingu) artykułów naukowych w oparciu o liczbę i jakość cytowań. Zaproponowano wówczas rekurencyjną definicję jakości publikacji naukowej, zgodnie z którą dana publikacja plasuje się wysoko w rankingu, jeśli jest cytowana przez wiele innych, wysoko ocenionych publikacji. Najważniejszą cechą tej definicji jest fakt, że przy ocenie jakości publikacji nie bierze ona pod uwagę tylko liczby cytowań, ale dokonuje ważenia każdego cytowania przez względną jakość cytującej publikacji (innymi słowy, jedno cytowanie w renomowanej publikacji może być więcej warte niż sto cytowań w nieistotnych publikacjach). Czasem żartobliwie przedstawiano powyższą definicję w postaci tzw. "zasady Hollywood": jesteś na tyle ważny w showbiznesie, ile znasz innych ważnych osób w showbiznesie.

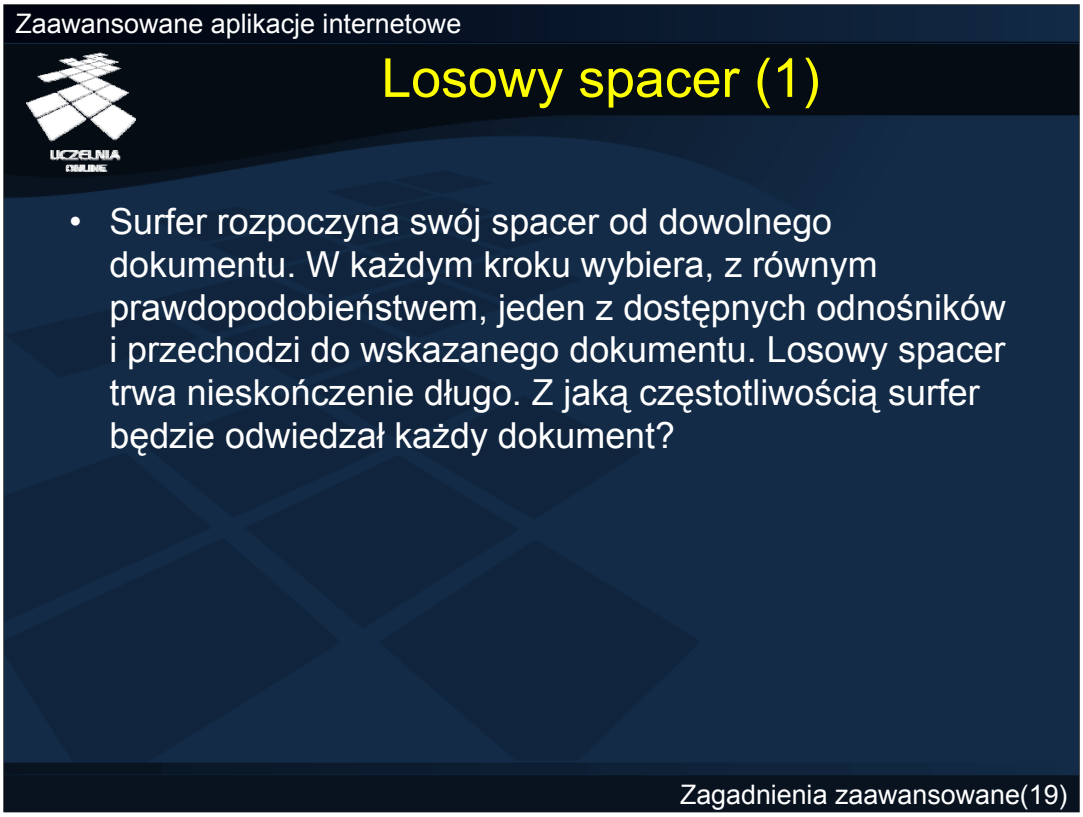
Istnieją trzy zasadnicze podejścia do problemu rankingu dokumentów WWW. Bez wątpienia najpopularniejszym algorytmem rankingu jest bezkontekstowy algorytm PageRank (PR), opracowany przez L.Page'a i S.Brina na Uniwersytecie Stanforda. Opracowany przez nich algorytm został zaimplementowany w prototypowej wyszukiwarce Google w roku 1998. Modyfikacją tego algorytmu jest zaproponowany przez Haveliwalę kontekstowy ranking dokumentów WWW o nazwie Topic-Specific PageRank (TSPR). Zupełnie inne podejście zaproponował w 1998 roku J.Kleinberg. Opracowany przez niego algorytm HITS (najczęściej nazywany Hubs & Authorities) stał się także źródłem inspiracji dla wielu kolejnych modyfikacji.



Podstawowa idea algorytmu PageRank jest bardzo podobna do przedstawionej wcześniej koncepcji oceny wartości publikacji naukowych na podstawie wartości cytowań. PageRank przypisuje wysoką wartość tym dokumentom WWW, które są wskazywane, za pomocą odnośników (hiperlinków), przez inne wartościowe dokumenty. Początkowo, każdy dokument WWW posiada tę samą jednostkową wartość. W kolejnych iteracjach każdy dokument oddaje całą posiadaną przez siebie wartość proporcjonalnie wszystkim wskazywanym przez siebie dokumentom. Oczywiście, wyliczenie wartości (czyli pozycji w rankingu) każdego dokumentu WWW jest czynnością trudną i żmudną. Na kolejnych slajdach przedstawiono zarys algorytmu PageRank i pewne podstawowe modyfikacje tego algorytmu.



Schemat działania algorytmu PageRank jest następujący. W pierwszym kroku tworzona jest stochastyczna macierz M reprezentująca całą indeksowaną sieć WWW. Wszystkie indeksowane dokumenty muszą zostać wpierw ponumerowane. Poszczególne wiersze i kolumny macierzy M reprezentują indeksowane dokumenty. Dokument i-ty odpowiada i-tej kolumnie i i-temu wierszowi macierzy M. Komórki macierzy M są inicjalizowane w następujący sposób. Jeśli dokument j-ty posiada n odnośników do innych dokumentów, to w kolumnie j-tej w wierszach reprezentujących te dokumenty umieszcza się wartość 1/n. Innymi słowy, element M [i,j] = 1/n, jeżeli dokument j-ty posiada odnośniki do n stron (w tym do dokumentu i-tego), w przeciwnym razie M [i,j] = 0. Interpretacja macierzy M jest następująca: wartość elementu M [i,j] określa prawdopodobieństwo przejścia do i-tego dokumentu, jeśli aktualnie przeglądany jest dokument j-ty. Warto tu zauważyć, że w prezentowanym modelu prawdopodobieństwo przejścia do dowolnego dokumentu wskazywanego przez j-ty dokument jest jednakowe (wszystkie odnośniki wychodzące z danego dokumentu mają tę samą wagę).



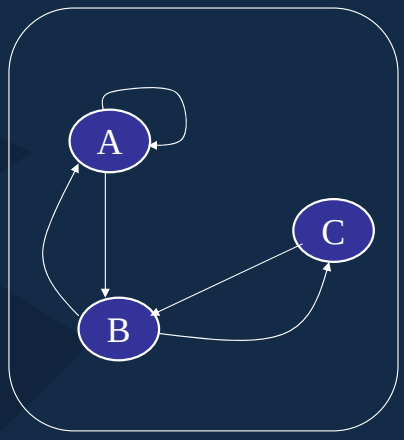
Najczęściej prezentuje się interpretację stochastycznej macierzy M wykorzystując pojęcie losowego spaceru (ang. random walk). Należy wyobrazić sobie hipotetycznego surfera, który rozpoczyna od dowolnego dokumentu w indeksowanej sieci, a następnie, w każdym kroku, przechodzi do dowolnego innego dokumentu wskazywanego przez odnośnik z danego dokumentu. Wybór odnośnika którym podąży surfer jest losowy a rozkład prawdopodobieństwa wyboru odnośnika jest jednostajny. Losowy spacer surfera trwa potencjalnie nieskończoną liczbę kroków. Pojawia się interesujące pytanie: czy można określić, z jaką częstotliwością losowy surfer będzie odwiedzał każdy dokument podczas losowego spaceru? A jeśli tak, to czy częstotliwość odwiedzin w danym dokumencie może coś powiedzieć o ważności (czyli pozycji w rankingu) danego dokumentu?

Niech **v** oznacza wektor prawdopodobieństw przebywania w dowolnym dokumencie indeksowanej sieci. Stan początkowy wektora **v** nie ma żadnego znaczenia. Najczęściej inicjalizuje się pozycje wektora **v** pewną niską wartością, obrazującą jednostajne prawdopodobieństwo rozpoczęcia losowego spaceru w dowolnym dokumencie sieci. Dysponując wektorem **v** i macierzą stochastyczną M można wyliczyć rozkład prawdopodobieństwa przebywania w każdym dokumencie po wykonaniu jednego kroku, rozkład ten jest dany przez wektor **v'**=M**v**. Zakładając nieskończony ciąg kroków, rozkład prawdopodobieństwa przebywania w dowolnym dokumencie w dowolnej chwili spaceru losowego jest dany przez granicę M(M(M(…M(M**v**)…))). A zatem, wektor wynikowy (po nieskończonej liczbie kroków) to wektor własny (ang. principal eigenvector) macierzy M. Wartości tego wektora (czyli prawdopodobieństwa przebywania w dowolnym dokumencie w dowolnej chwili spaceru losowego) nazywamy wartościami PageRank dokumentów. W algorytmie PageRank wartość i-tej pozycji wektora własnego macierzy M stanowi ważność i-tego dokumentu.

Przykład (1)

- Sieć składa się z dokumentów A, B, i C. Poniższy graf przedstawia strukturę połączeń pomiędzy dokumentami

$$M = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{pmatrix}$$



Dla ilustracji działania algorytmu PageRank rozważmy prosty przykład przedstawiony na slajdzie. Załóżmy, że sieć składa się z trzech dokumentów A, B, i C. Graf przedstawiony na slajdzie przedstawia strukturę połączeń pomiędzy dokumentami. Niech v = [a,b,c] oznacza wektor ważności dokumentów, odpowiednio, A, B, C. Macierz M reprezentującą tak zdefiniowaną sieć przedstawiono na slajdzie. Przykładowo, pierwsza kolumna macierzy M zawiera następujące elementy: 1/2, 1/2, 0. Elementy M [1,1] = 1/2 i M [2,1] = 1/2, gdyż dokument A (odpowiada mu numer 1) posiada odnośnik do siebie i odnośnik do dokumentu B (dokumentowi B odpowiada numer 2). Element M [3,1] = 0 , gdyż nie ma odnośnika z dokumentu A do dokumentu C (dokumentowi C odpowiada numer 3). Druga kolumna macierzy M zawiera elementy: 1/2, 0 i 1/2, gdyż dokument B posiada odnośnik do dokumentów A i C. Zatem 1/2 ważności dokumentu B jest przekazywane dokumentowi A i 1/2 ważności dokumentu B wędruje do dokumentu C. Trzecia kolumna macierzy M zawiera elementy: 0, 1 i 0, gdyż dokument C posiada jedynie link do dokumentu B, zatem cała ważność dokumentu C jest przekazywana do dokumentu B.

Przykład (2)

- Równanie opisujące ważność dokumentów A, B, C:

$$\mathbf{v} = M\,\mathbf{v}$$

Rozwiązanie powyższego równania można znaleźć metodą iteracyjną poprzez wymnażanie macierzą M kolejnych estymat ważności stron **v**. Pierwsze 4 iteracje dają następujące oszacowania rozkładu ważności **v**:

$$\begin{bmatrix}1\\1\\1\end{bmatrix}\ \begin{bmatrix}1\\3/2\\1/2\end{bmatrix}\ \begin{bmatrix}5/4\\1\\3/4\end{bmatrix}\ \begin{bmatrix}9/8\\11/8\\1/2\end{bmatrix}\ \begin{bmatrix}5/4\\17/16\\11/16\end{bmatrix}\ \ldots\ \begin{bmatrix}6/5\\6/5\\3/5\end{bmatrix}$$

Jak już wiadomo, równanie opisujące ważność dokumentów A, B, C ma postać: v = M v. Rozwiązanie powyższego równania można znaleźć metodą relaksacyjną (iteracyjną) poprzez wymnażanie macierzą M kolejnych estymat ważności dokumentów **v**. Ponieważ stan początkowy wektora v jest nieistotny, dla uproszczenia obliczeń przyjęto, że początkowo każdy dokument ma tę samą ważność 1. Oszacowanie ważności dokumentów uzyskane po 4 pierwszych iteracjach przedstawiono na slajdzie. W granicy, rozwiązanie posiada następujące wartości a=b=6/5, c=3/5, tj. ważność dokumentów A i B jest dwukrotnie większa niż ważność dokumentu C

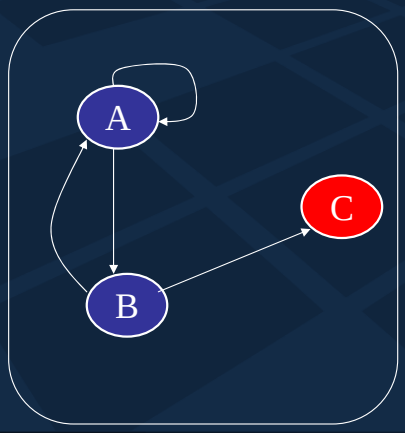
Problemy w rzeczywistych grafach

- Problemy związane z rzeczywistą strukturą grafów sieciowych
  - "pajęcza pułapka" (ang. *spider trap*), grupa dokumentów nie posiadających żadnych odnośników wychodzących, w efekcie dokonują one przejęcia całej wpływającej do nich ważności
  - "ślepa uliczka" (ang. *dead end*), dokument nie posiadający żadnych odnośników wychodzących, przez który "wycieka" cała ważność sieci

W przypadku modelowania rzeczywistych struktur sieci WWW można często natknąć się na dwa problemy, które mogą prowadzić do zniekształcenia oszacowań ważności stron. Są to problem „ślepej uliczki” i problem „pajęczej pułapki ”. „Ślepą uliczką” (ang. dead end) nazywamy dokument, który nie posiada żadnych następników, a tym samym nie ma gdzie przekazać swojej ważności. W takim przypadku ważność wszystkich stron dąży do 0 (mówimy także, że cała ważność sieci "wycieka" przez ślepą uliczkę). „Pajęczą pułapką” (ang. spider trap) nazywamy grupę dokumentów, która nie posiada odnośników wychodzących poza grupę, a tym samym, przechwytuje ważność całej sieci Web. Przykłady przedstawione na kolejnych slajdach ilustrują oba zjawiska.

Problem "ślepej uliczki"

- Sieć składa się z dokumentów A, B, i C. Poniższy graf przedstawia strukturę połączeń pomiędzy dokumentami



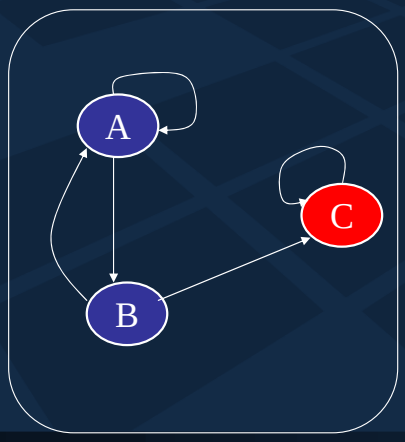
$$M = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \end{pmatrix}$$

$$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 1/2 \\ 1/2 \end{bmatrix} \quad \begin{bmatrix} 3/4 \\ 1/2 \\ 1/4 \end{bmatrix} \quad \begin{bmatrix} 5/8 \\ 3/8 \\ 1/4 \end{bmatrix} \quad \begin{bmatrix} 1/2 \\ 5/16 \\ 3/16 \end{bmatrix} \cdots \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Rozważmy przykład przedstawiony na slajdzie. Załóżmy, że sieć składa się z dokumentów A, B, i C. Graf przedstawiony na slajdzie przedstawia strukturę połączeń pomiędzy dokumentami. Łatwo zauważyć, że dokument C nie posiada odnośników wychodzących, jest zatem typowym przykładem „ślepej uliczki”. Niech v= [a,b,c] oznacza wektor ważności dokumentów, odpowiednio, A, B, C. Macierz M tak zdefiniowanej sieci przedstawiono na slajdzie. Jak łatwo zauważyć, ponieważ dokument C nie posiada następników, ostatnia kolumna macierzy M składa się z samych zer. W konsekwencji, kolejne iteracje dają oszacowania rozkładu ważności jak przedstawiono na slajdzie. Jak widać, ważność wszystkich stron dąży do 0.

Problem "pajęczej pułapki"

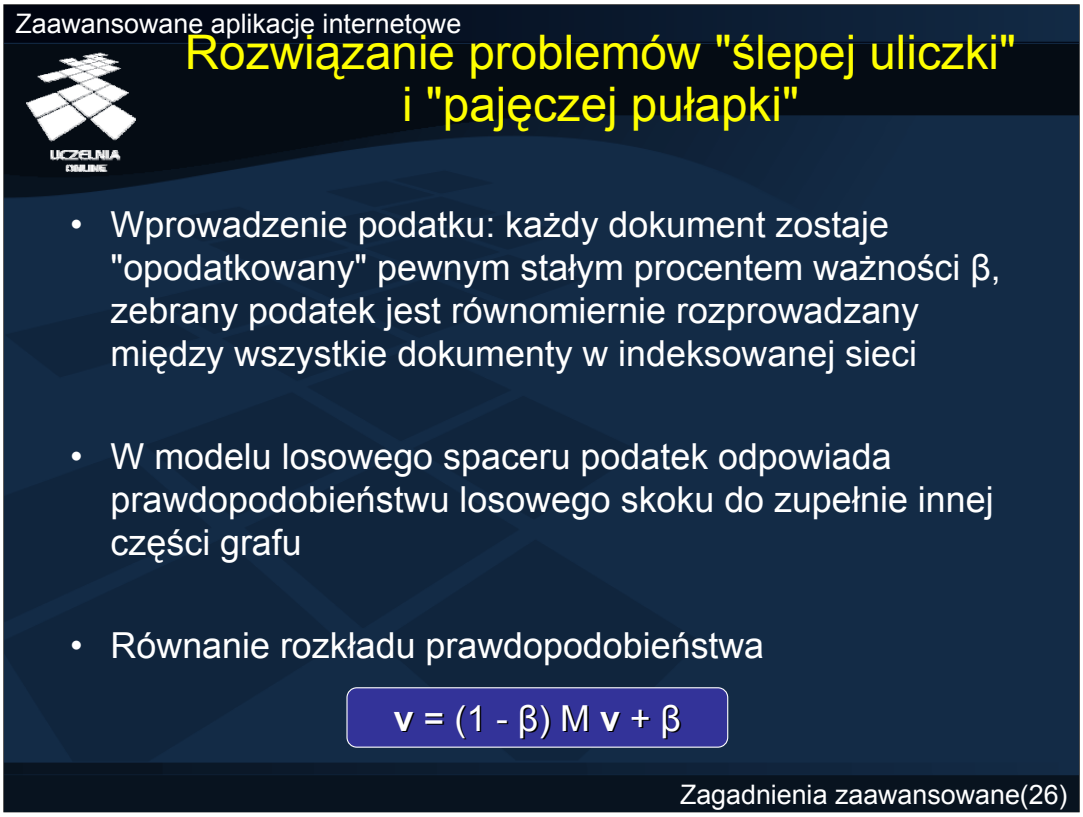
- Sieć składa się z dokumentów A, B, i C. Poniższy graf przedstawia strukturę połączeń pomiędzy dokumentami



$$M = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 1 \end{pmatrix}$$

$$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1/2 \\ 3/2 \end{bmatrix} \begin{bmatrix} 3/4 \\ 1/2 \\ 7/4 \end{bmatrix} \begin{bmatrix} 5/8 \\ 3/8 \\ 2 \end{bmatrix} \begin{bmatrix} 1/2 \\ 5/16 \\ 35/16 \end{bmatrix} \cdots \begin{bmatrix} 0 \\ 0 \\ 3 \end{bmatrix}$$

Kolejny przykład ilustruje zjawisko "pajęczej pułapki". Załóżmy, podobnie jak poprzednio, że sieć składa się z dokumentów A, B, i C. Graf przedstawiony na slajdzie przedstawia strukturę połączeń pomiędzy dokumentami. Zauważmy, że dokument C tym razem posiada odnośnik wychodzący, ale jest to odnośnik zwrotny do dokumentu C. Dokument C jest zatem typowym przykładem "pajęczej pułapki", ponieważ cała ważność przekazana do dokumentu C (np. przez dokument B) pozostanie już na zawsze w dokumencie C. Niech v = [a,b,c] oznacza wektor ważności dokumentów, odpowiednio, A, B, C. Macierz M tak zdefiniowanej sieci przedstawiono na slajdzie. Jak łatwo zauważyć, ponieważ dokument C posiada tylko jeden odnośnik wychodzący do siebie, ostatnim elementem kolumny 3 macierzy M jest 1. W konsekwencji, kolejne iteracje dają oszacowania rozkładu ważności jak przedstawiono na slajdzie. Ważność dokumentu C dąży do 3, natomiast ważność dokumentów A i B wynosi a=b=0. Dokument C przechwycił ważność całej sieci.



Rozwiązanie problemów "ślepej uliczki" i "pajęczej pułapki", przyjęte przez Google, polega na "opodatkowaniu" każdego dokumentu pewnym procentem jego ważności i równomierne rozdystrybuowanie łącznego podatku pomiędzy wszystkie dokumenty. Przykładowo, wprowadzając podatek w wysokości 20%, równania ważności dokumentów z poprzedniego przykładu przyjmą następującą postać:

a = 0.8 * (1/2*a + 1/2*b +0*c) + 0.2

b = 0.8 * (1/2*a + 0*b + 0*c) + 0.2

c = 0.8 * (0*a + 1/2*b + 1*c) + 0.2

Rozwiązaniem tych równań są następujące wartości ważności dokumentów: a=7/11, b=5/11, c=21/11. Zauważmy, że obecnie, w przeciwieństwie do wartości ważności stron podanych w poprzednim przykładzie, ważność dokumentów A i B jest różna od zera.

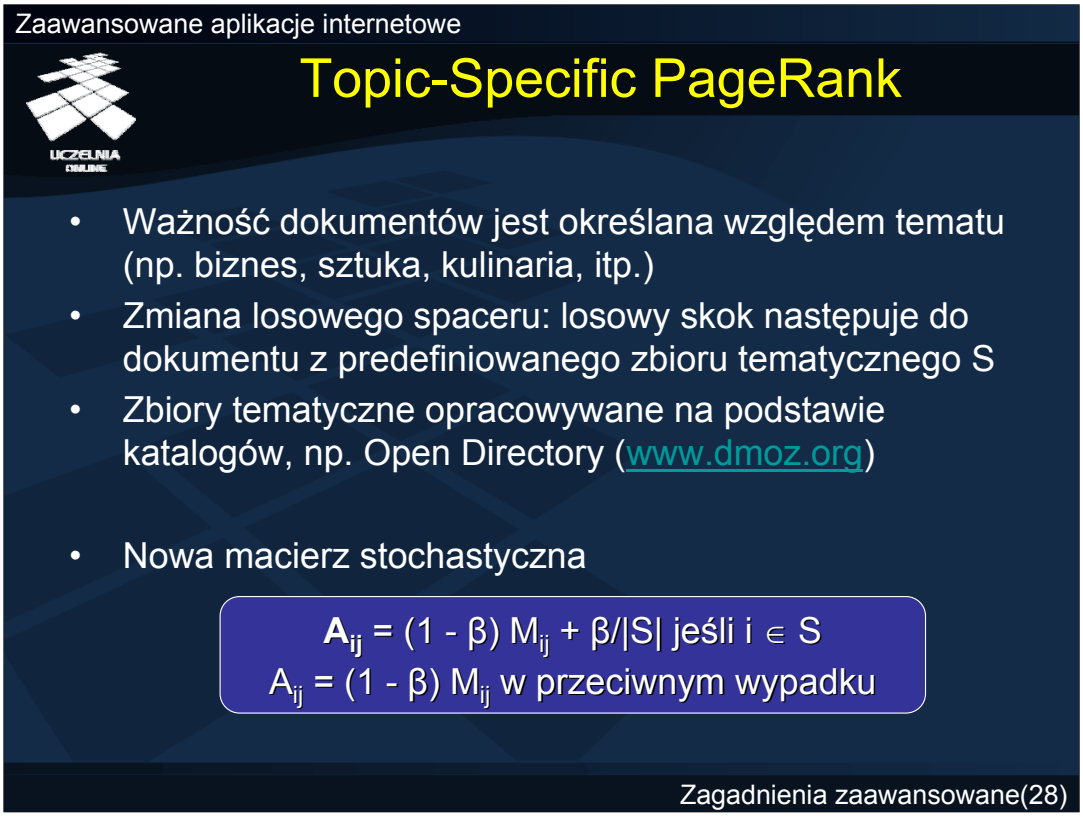
Firma Google publicznie nie ujawniła wysokości podatku, choć wyniki licznych eksperymentów wskazują, że najprawdopodobniej wysokość tego podatku jest zbliżona do 15%.

PageRank - usprawnienia

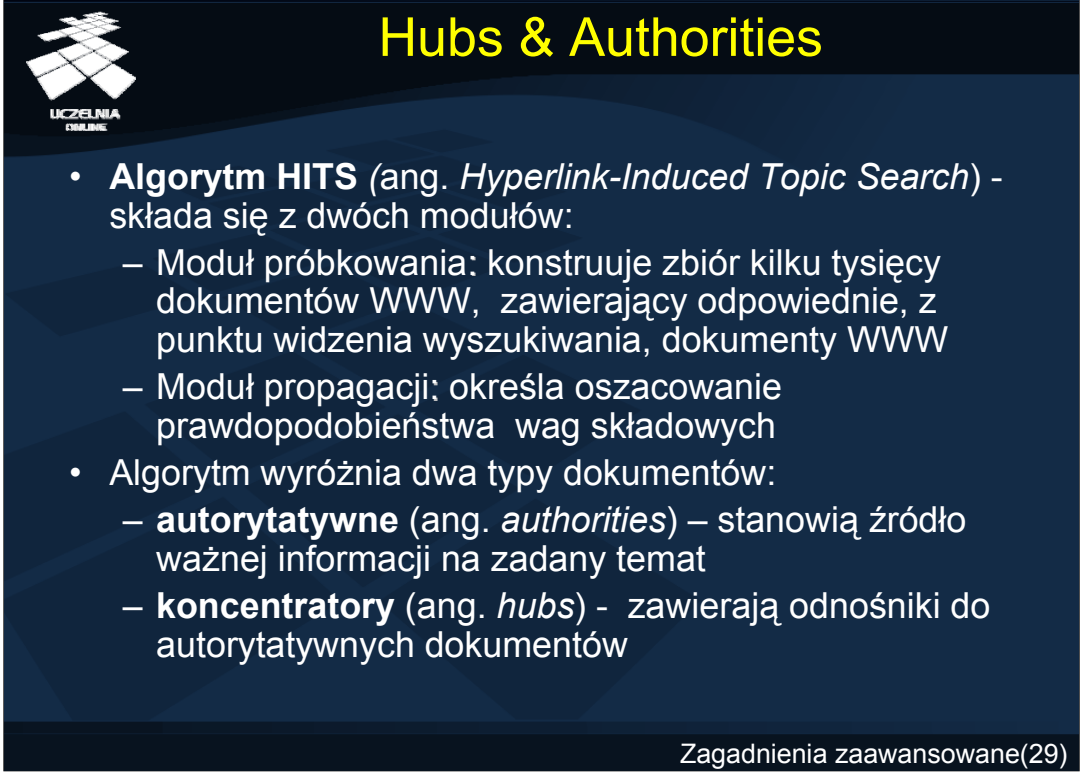
- Predykcja wartości granicznych w wektorze własnym
- Wykorzystanie cech lokalnych: wyliczenie lokalnej wartości PageRank w ramach serwera/domeny
- Symulacja wielu losowych surferów poruszających się równolegle po sieci

- Problemy
  - podatny na spamming
  - nadmierna wartość niektórych dokumentów
  - wieloznaczność słów kluczowych

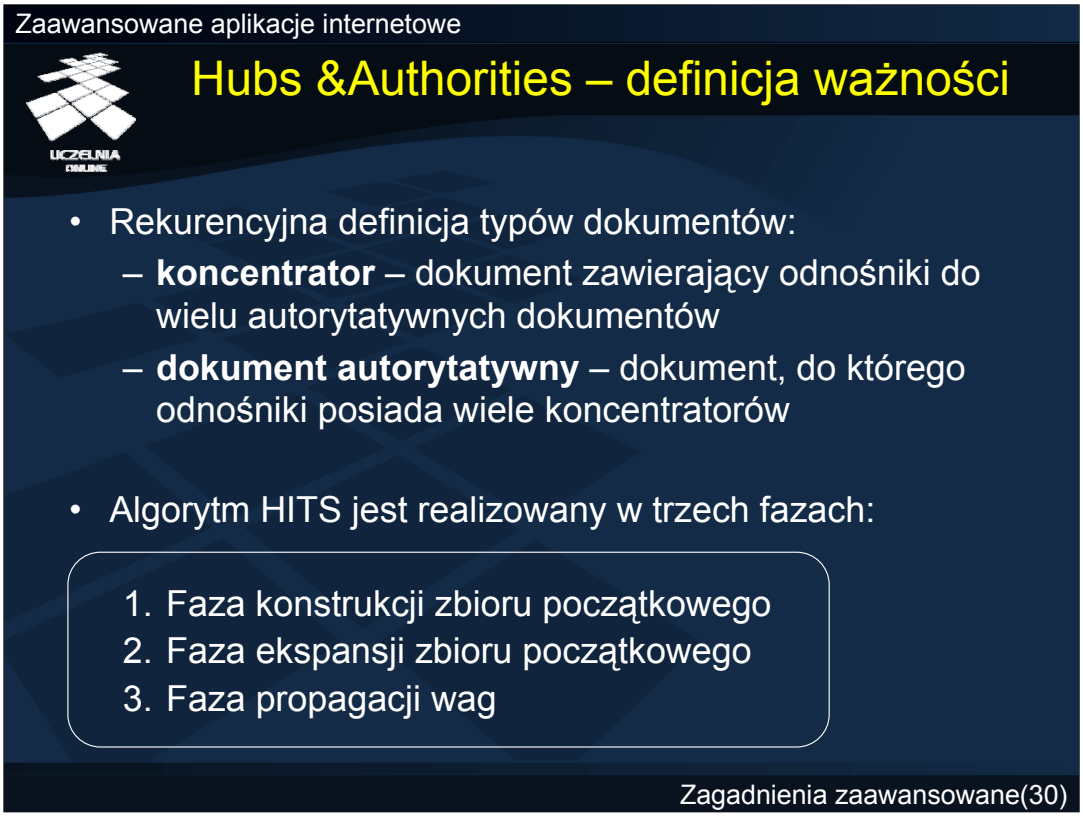
Istnieje wiele możliwości ulepszania oryginalnego algorytmu PageRank. W wielu przypadkach zamiast kosztownych obliczeń można posłużyć się predykcją prawdopodobnych wartości granicznych w wektorze własnym. Takie rozwiązanie jest szczególnie korzystne w momencie, w którym kolejne iteracje przeliczania wektora ważności dokumentów wyraźnie wskazują na niewielkie wahania asymptotyczne wokół wartości granicznych. Inną propozycją jest wykorzystanie cech lokalności odwołań. Wiadomo skądinąd, że dokumenty umieszczone na jednym serwerze/domenie są ze sobą w naturalny sposób silniej powiązane. Zamiast pracochłonnego wyliczania wektora własnego macierzy reprezentującej wszystkie dokumenty, można wcześniej pogrupować dokumenty (np. według serwera), wyliczyć lokalną wartość PageRank dla każdego dokumentu, na tej podstawie wyliczyć uśrednioną wartość PageRank dla grupy dokumentów i kontynuować wyliczenia dla grup dokumentów. Takie rozwiązanie bardzo wydatnie przyspiesza obliczenia. Wreszcie, można pokusić się o zamodelowanie wielu równolegle poruszających się surferów (szczegółowy opis tego usprawnienia wykracza jednak poza ramy tego wykładu). Mimo ogromnej popularności i praktycznej użyteczności algorytmu PageRank, doczekał się on wielu słów krytyki. Algorytm jest dość podatny na tzw. link spamming, czyli tworzenie klik tysięcy kooperujących ze sobą serwisów, które wykorzystują powiązania między sobą do skoordynowanego podniesienia rankingu wybranych stron. Zaobserwowano również tendencję PageRank do nadmiernego podnoszenia wartości wybranych typów serwerów, w szczególności serwerów bardzo silnie zorientowanych tematycznie. Bodaj największym problemem związanym z algorytmem PageRank jest jednak fakt, że ranking jest bezkontekstowy, tzn. niezależnie od aktualnie przetwarzanego zapytania ranking dokumentów pozostaje niezmienny. W przypadku gdy zapytanie jest wieloznaczne (np. "jaguar"), w odpowiedzi wyszukiwarka zwróci zarówno adres serwisu producenta samochodów, jak i adres dokumentu w encyklopedii opisującego drapieżnika południowoamerykańskiego.



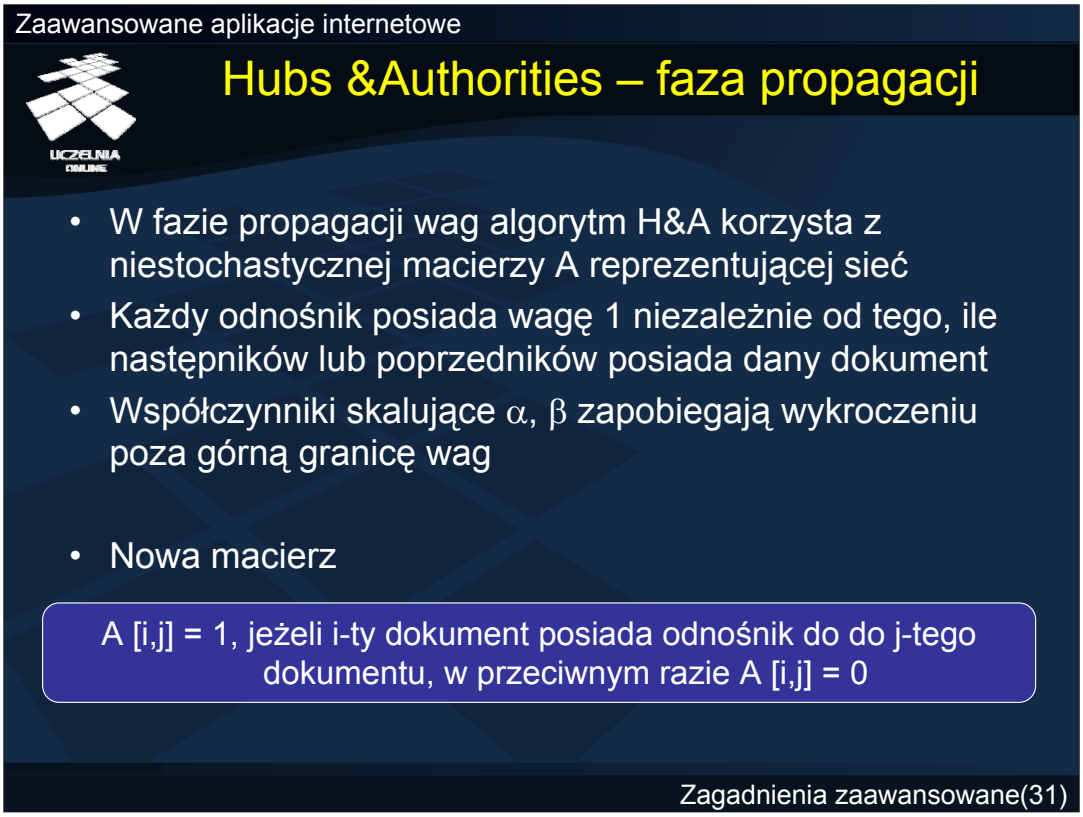
Ciekawym rozwiązaniem jest zaproponowany przez T.Haveliwalę algorytm Topic-Specific PageRank. Jest to prosta modyfikacja oryginalnego algorytmu PageRank, polegająca na wprowadzeniu do algorytmu zbioru kontekstów. Konteksty odpowiadają grupom tematycznym. Każda grupa tematyczna posiada swój zbiór źródłowy dokumentów S. W trakcie losowego spaceru, jeśli surfer decyduje się na skok do innej części sieci, skok musi nastąpić do jednego z dokumentów ze zbioru S. W ten sposób dokumenty źródłowe uzyskują dużo wyższy ranking, a co za tym idzie, wszystkie dokumenty położone stosunkowo blisko dokumentów źródłowych również pną się do góry w rankingu. Zbiory dokumentów źródłowych mogą być tworzone ręcznie lub automatycznie, np. przy użyciu dostępnych katalogów (przykładem publicznego katalogu tematycznego jest Open Directory). Cała reszta algorytmu Topic-Specific PageRank jest identyczna jak w przypadku oryginalnego algorytmu PageRank, jedyna różnica to sformułowanie elementów macierzy stochastycznej: wprowadza się do niej preferencję skoku do dokumentów źródłowych, jak pokazano na slajdzie. Przeprowadzone eksperymenty sugerują, że ranking zwrócony przez Topic-Specific PageRank jest wyższej jakości niż ranking zwracany przez tradycyjny PageRank. W przypadku algorytmu Topic-Specific PageRank największym wyzwaniem jest skalowalność algorytmu względem liczby obsługiwanych grup tematycznych. I w tej dziedzinie pojawiły się niedawno ciekawe rozwiązania, warto np. wspomnieć o możliwości redukcji złożoności obliczeniowej przez zastąpienie wyliczania wektorów własnych wyliczaniem wektorów częściowych.



Kolejnym prezentowanym algorytmem rankingu stron jest algorytm H&A (oryginalna nazwa algorytmu to algorytm HITS, ang. Hyperlink-Induced Topic Search). Algorytm H&A składa się z dwóch modułów: modułu próbkowania oraz modułu propagacji. Moduł próbkowania konstruuje zbiór kilku tysięcy dokumentów WWW, zawierający odpowiednie, z punktu widzenia wyszukiwania, dokumenty WWW. Zbiór ten może zostać skonstruowany na podstawie wyniku zwróconego przez tradycyjną wyszukiwarkę internetową. Z kolei moduł propagacji określa oszacowanie ważności dokumentów. Algorytm wyróżnia dwa typy dokumentów: dokumenty autorytatywne (ang. authorities) – stanowią one źródło ważnej informacji na zadany temat, oraz koncentratory (ang. hubs) – są to dokumenty zawierające odnośniki do autorytatywnych dokumentów.



Definicja typów dokumentów w algorytmie H&A ma charakter rekurencyjny (podobnie jak definicja ważnego dokumentu w algorytmie PageRank). Koncentrator to dokument zawierający odnośniki do wielu autorytatywnych dokumentów, natomiast dokument autorytatywny to dokument, do którego odnośniki posiada wiele koncentratorów. Algorytm H&A jest realizowany w trzech fazach: fazie konstrukcji zbioru początkowego, fazie ekspansji oraz fazie propagacji wag. Do konstrukcji zbioru początkowego wykorzystuje się indeks wyszukiwarki, który, w oparciu o zbiór słów kluczowych, znajduje początkowy zbiór ważnych dokumentów (zarówno autorytatywnych jak i koncentratorów). Następnie, w fazie ekspansji, początkowy zbiór dokumentów jest rozszerzony do tzw. zbioru bazowego (ang. *base set*) poprzez włączenie do zbioru początkowego wszystkich dokumentów, do których dokumenty zbioru początkowego zawierają odnośniki, oraz dokumentów, które zawierają odnośniki do dokumentów ze zbioru początkowego. Warunkiem stopu procesu ekspansji jest osiągnięcie określonej liczby dokumentów (zazwyczaj kilka tysięcy). Wreszcie, w fazie propagacji wag moduł propagacji oblicza iteracyjnie wartości oszacowania prawdopodobieństwa, że dany dokument jest autorytatywny lub że jest koncentratorem. Odnośniki pomiędzy dokumentami z tej samej domeny, najczęściej, służą do celów nawigacyjnych, stąd, odnośniki te są wyłączane z analizy.



W fazie propagacji wag, algorytm H&A korzysta z macierzowego opisu sieci Web, podobnie jak algorytm PageRank. Różnica polega na tym, że w przypadku algorytmu H&A macierz A modelująca indeksowaną sieć nie jest macierzą stochastyczną. W macierzy A każdy odnośnik posiada wagę 1, niezależnie od tego, ile następników lub poprzedników posiada dany dokument. Ze względu na brak ograniczenia dotyczącego stochastyczności macierzy A, algorytm H&A wprowadza dwa współczynniki skalujące, alfa i beta, tak, aby wartości wag nie przekroczyły górnego ograniczenia wartości wag. Definicja macierzy A została przedstawiona na slajdzie.

Hubs &Authorities – definicje wektorów

- Wektory **a** i **h** oznaczają, odpowiednio, wektory autorytatywności i koncentratywności, których i-ty element odpowiada wartości stopnia autorytatywności i koncentratywności i-tego dokumentu

h = α A **a**. Koncentratywność danego dokumentu jest sumą autorytatywności wszystkich dokumentów, do których dany dokument posiada odnośniki, pomnożoną przez współczynnik α

a = β A^T **h**. Autorytatywność danego dokumentu jest sumą koncentratywności wszystkich dokumentów, które posiadają odnośniki do danego dokumentu, pomnożoną przez współczynnik β

Niech wektory a i h oznaczają, odpowiednio, wektory autorytatywności i koncentratywności dokumentów. I-ty element każdego wektora odpowiada wartości stopnia autorytatywności i koncentratywności i-tego dokumentu. Niech alfa i beta oznaczają odpowiednie współczynniki skalujące. Z definicji typów dokumentów, przedstawionych poprzednio, otrzymujemy, że h = alfa*A*a. Innymi słowy, koncentratywność danego dokumentu jest sumą autorytatywności wszystkich dokumentów, do których dany dokument posiada odnośniki, pomnożoną przez współczynnik skalujący alfa. Podobnie, definiujemy autorytatywność stron: a = beta*A(transponowane)*h. Autorytatywność danego dokumentu jest sumą koncentratywności wszystkich dokumentów, które posiadają odnośniki do danego dokumentu, pomnożoną przez współczynnik skalujący beta.

Hubs &Authorities - obliczenia

- Z powyższych definicji wynika, że:

$$\mathbf{h} = \alpha\beta A A^T \mathbf{h}$$

$$\mathbf{a} = \alpha\beta A^T A \mathbf{a}$$

- Rozwiązanie powyższych równań można znaleźć metodą iteracyjną, zakładając, że początkowe wartości autorytatywności i koncentratywności każdego dokumentu wynoszą 1

- Problem wyboru wartości współczynników α i β
- Problem przetwarzania on-line

Podstawiając do wzoru na koncentratywność dokumentu definicję wektora a, oraz, podstawiając do wzoru na autorytatywność dokumentu definicję wektora h, otrzymujemy następujące wzory:

h = beta * alfa* A*A(transponowane)***h**

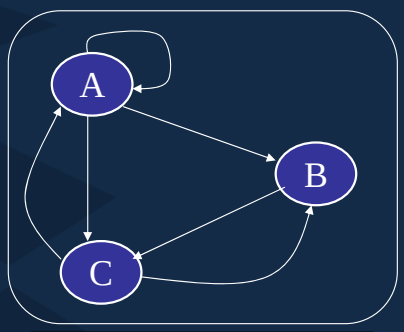
a = beta * alfa* A(transponowane)*A***a**

Rozwiązanie powyższych równań można znaleźć metodą iteracyjną, zakładając, że początkowe wartości autorytatywności i koncentratywności każdego dokumentu wynoszą 1. Oczywiście, pozostaje jeszcze problem wyboru odpowiednich wartości współczynników skalujących alfa i beta. Wartości tych współczynników dobiera się doświadczalnie. Celem tych współczynników jest zagwarantowanie, że wartości autorytatywności i koncentratywności dokumentów nie przekroczą górnych ograniczeń przyjętych dla tych wartości. Oprócz tego obecność fazy konstrukcji zbioru początkowego powoduje, że zastosowanie algorytmu H&A w praktyce staje się trudne.

Przykład (1)

- Sieć WWW składa się z dokumentów A, B, i C

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



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

$$AA^T = \begin{pmatrix} 3 & 1 & 2 \\ 1 & 1 & 0 \\ 2 & 0 & 2 \end{pmatrix}$$

$$A^T A = \begin{pmatrix} 2 & 2 & 1 \\ 2 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

Dla ilustracji działania algorytmu H&A rozważmy prosty przykład przedstawiony na slajdzie. Załóżmy, że sieć składa się z dokumentów A, B, i C. Graf przedstawiony na slajdzie pokazuje strukturę połączeń pomiędzy dokumentami. Macierz A tak skonstruowanej sieci WWW przedstawiono na slajdzie. Przykładowo, wiersz 1 macierzy A zawiera następujące elementy: 1, 1 i 1. Elementy A [1,1] = A [1,2] = A [1,3] = 1, gdyż dokument A (któremu odpowiada numer 1) posiada odnośniki wychodzące do dokumentów A, B (z numerem 2) i C (z numerem 3). Wiersz 2 macierzy A zawiera następujące elementy: 0, 0, i 1. Elementy A [2,1] = A [2,2] = 0, gdyż dokument B nie posiada odnośników wychodzących do dokumentów A i B, natomiast element A [2,3] = 1, gdyż dokument B posiada odnośnik wychodzący do dokumentu C. Wreszcie, wiersz 3 macierzy A zawiera elementy: 1, 1, i 0. Elementy A [3,1] = A [3,2] = 1, gdyż dokument C posiada odnośniki wychodzące do dokumentów A i B, natomiast element A [3,3] = 0, gdyż dokument C nie posiada odnośnika zwrotnego do siebie. Na slajdzie przedstawiono również macierze: A(transponowane) oraz iloczyny macierzy A * A(transponowane) oraz A(transponowane)* A.

Przykład (2)

- Przyjmując $\alpha = \beta = 1$ i zakładając, że początkowe wartości autorytatywności i koncentratywności każdego dokumentu wynoszą 1, **h** = [ha=1, hb=1, hc=1] i **a** = [aa=1, ab=1, ac=1], po trzech pierwszych iteracjach otrzymujemy następujące wartości:

aa = 1	5	24	114
ab = 1	5	24	114
ac = 1	4	18	84
ha = 1	6	28	132
hb = 1	2	8	36
hc = 1	4	20	96

Przyjmując, że współczynniki skalujące alfa = beta = 1 i zakładając, że początkowe wartości autorytatywności i koncentratywności każdego dokumentu wynoszą 1, h = [ha=1, hb=1, hc=1] i a = aa=1, ab=1, ac=1], po trzech pierwszych iteracjach otrzymujemy wartości autorytatywności i koncentratywności dokumentów jak przedstawiono na slajdzie. Łatwo zauważyć, że dokument A jest typowym koncentratorem, natomiast dokument B jest dokumentem autorytatywnym.