Analiza sieci - web

Eksploracja danych - wykład

Joanna Kołodziejczyk

Plan prezentacji

1. Text analytics - Analiza tekstu
   - Web - mining
2. Social Network Analysis
   - Wprowadzenie
   - Pojęcia i definicje
3. Algorytm PageRank
   - Wprowadzenie

Plan prezentacji

1. Text analytics - Analiza tekstu
   - Web - mining
2. Social Network Analysis
   - Wprowadzenie
   - Pojęcia i definicje
3. Algorytm PageRank
   - Wprowadzenie

Text analytics

Analiza tekst (tzw. text mining)

Odkrywa znaczenie i cel w zgromadzonych danych. <– Codziennie powstaje średnio 2.5 kwintyliona bajtów danych nieustrukturalizowanych: tekstów, tweetów, zdjęć i filmów.

Historia:

1. 1887 Thomas Mendenhall wykorzystał metody statystyczne do analizy krzywych rozkładu cech w artykule opublikowanym w czasopiśmie Science.
2. 1963 Claude Brinegar przeanalizował pisma Quintusa Curtiusa Snodgrassa aby udowodnić, że jest to pseudonim Marka Twaina.
3. 1963 Frederick Mosteller i David Wallace użyli statystycznego modelu Naïve Bayes aby przeanalizować, który z ojców założycieli USA, Alexander Hamilton czy James Madison napisał fragmenty Federalist Paper.
4. Lata 80 poprzedniego wieku pojawienie się pojęcia „text mining”.
5. Połow 2000 pojawia się termin „text analytics”.

Wzmożone gromadzenie danych

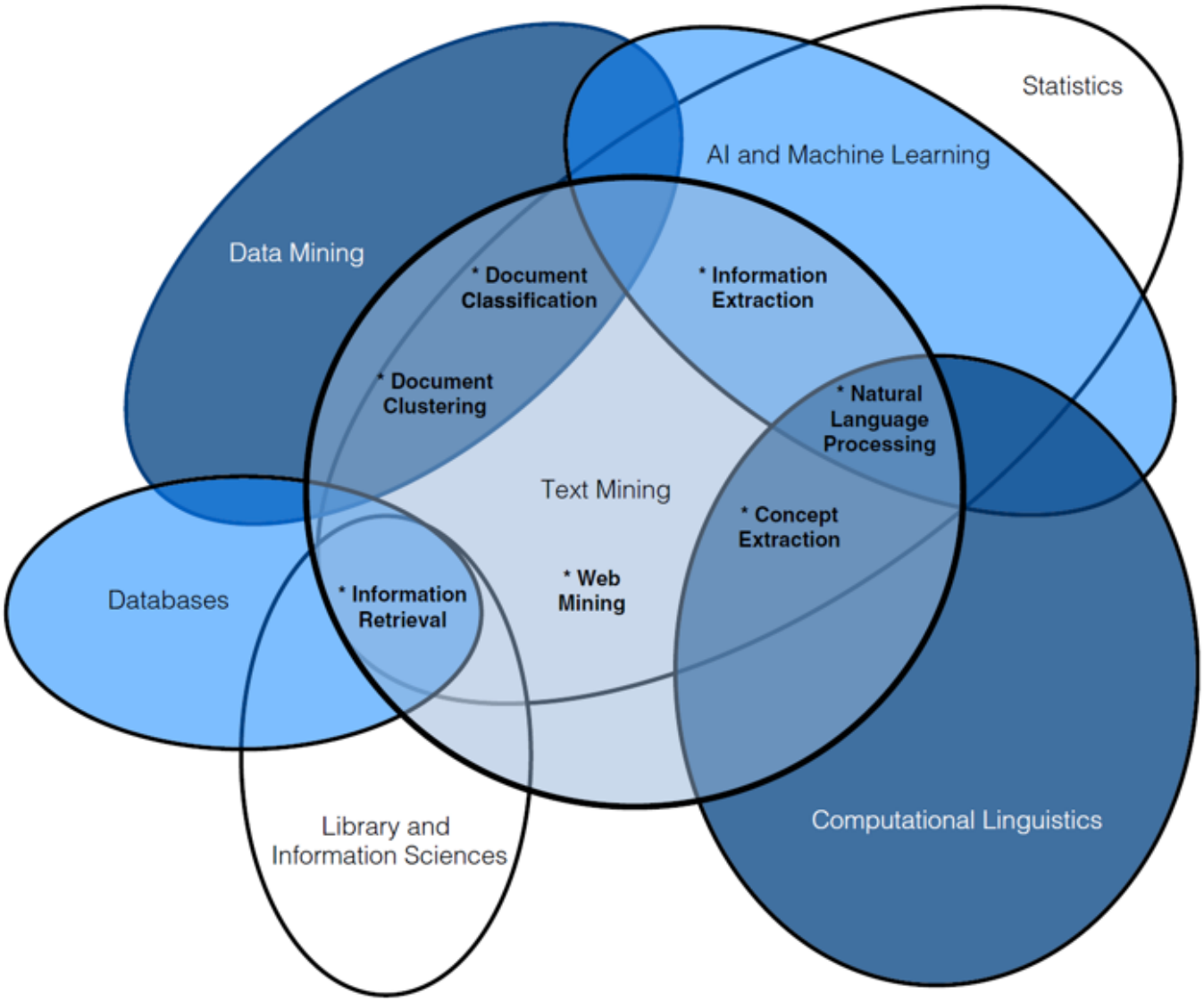
Pierwszy punkt zwrotny

Popularność analizy tekstu pokrywała się z wszechobecnością poczty elektronicznej i spamu, ponieważ była i jest używana do automatycznego wykrywania, które wiadomości są prawdopodobnie niechciane.

Drugi i trzeci punkt zwrotny

Drugi zwrot to pojawienie się smartfonów i sms-ów. Trzeci to pojawienie się mediów społecznościowych Facebook, Twitter, Instagram, YouTub napędzała generowanie ogromnych ilości danych, z których większość jest nieustrukturyzowana, a znaczna to tekst.

Schemat Venn'a działów analizy tekstu i ich powiązań



Analiza tekstu czym się zajmuje

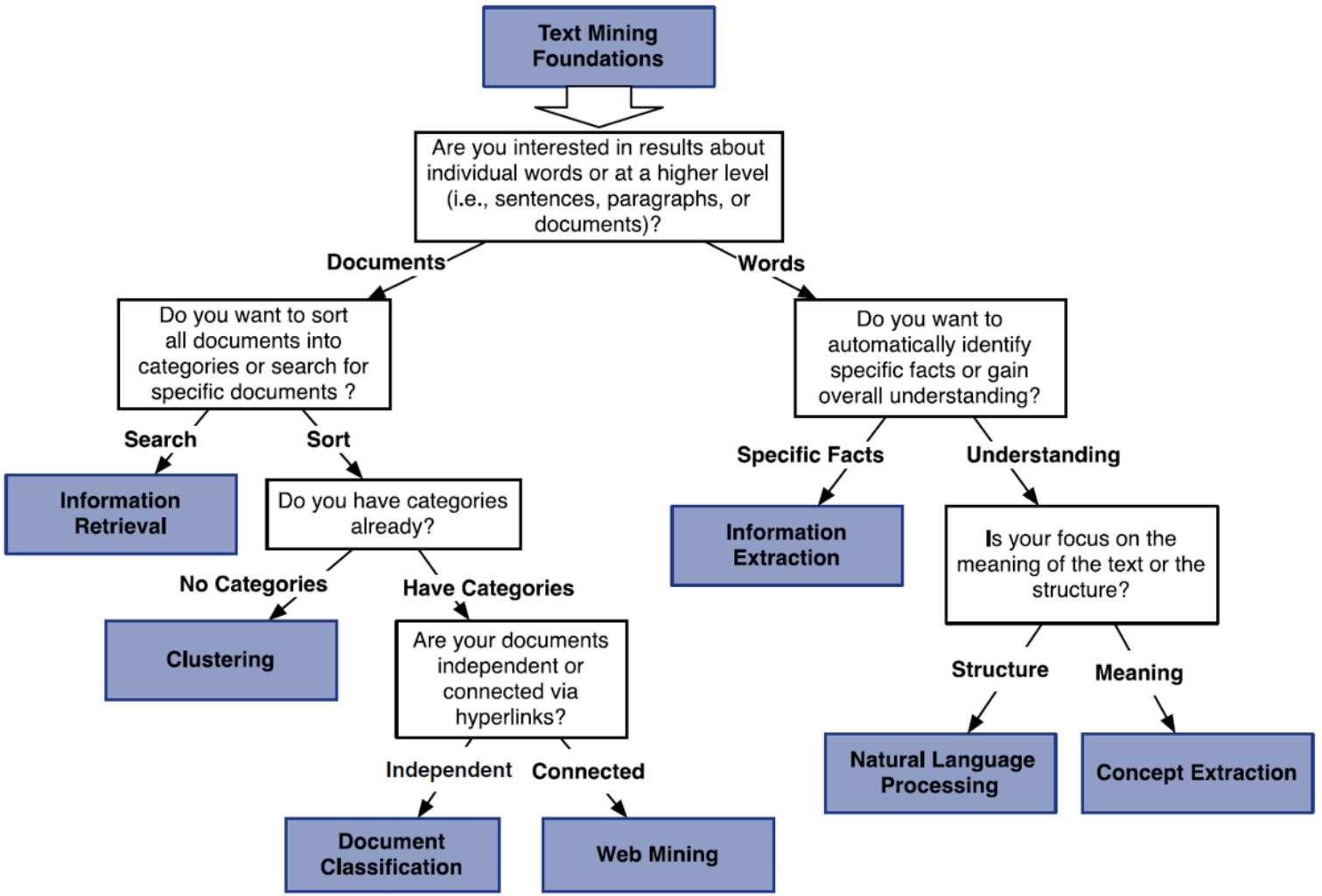
Analiza tekstu — zadania

1. zrozumieć sam tekst,
2. zidentyfikować lub skategoryzować autora(ów),
3. połączyć tekst z czymś namacalnym np. wydarzeniem.

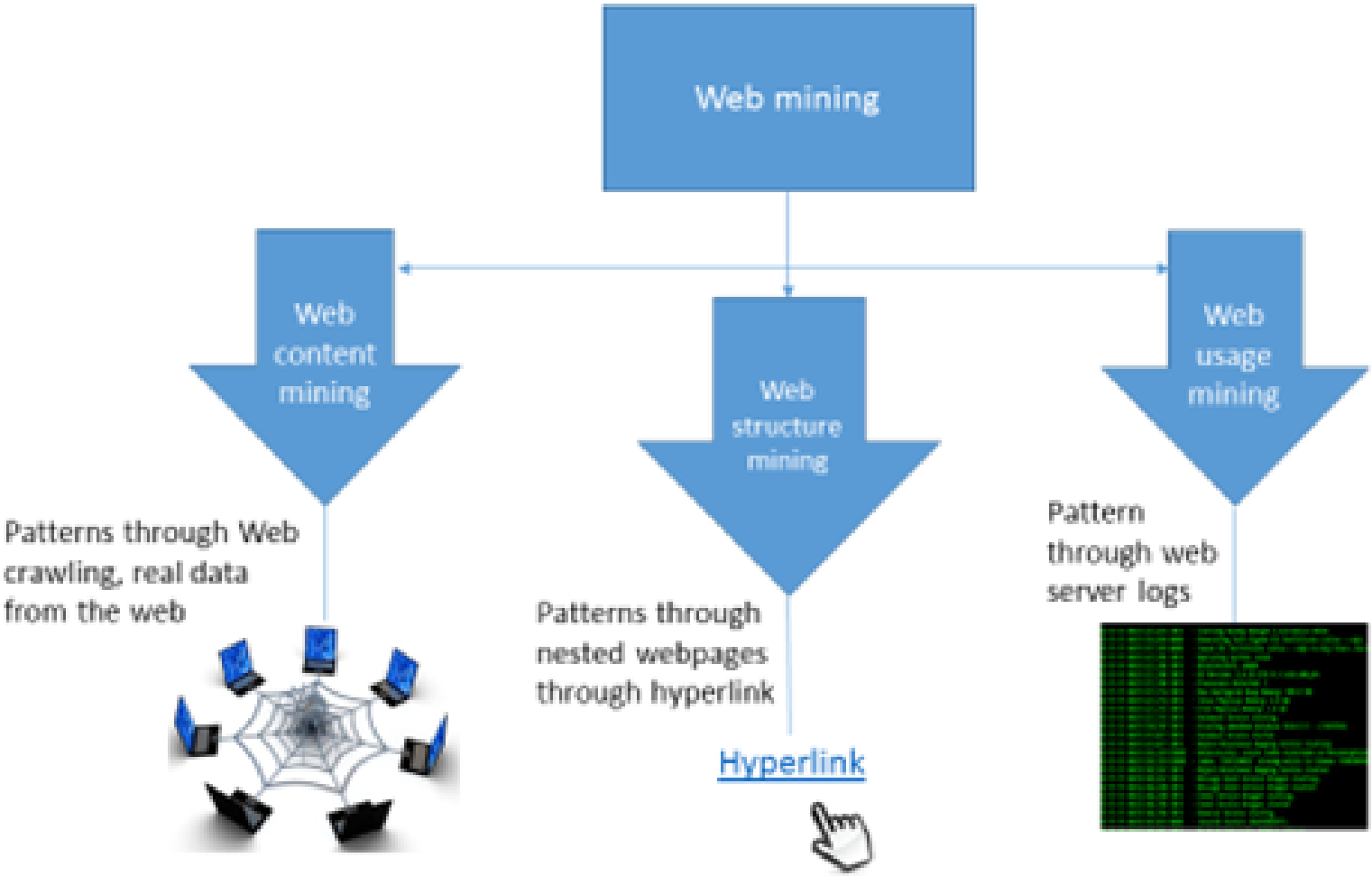
Analiza tekstu — techniki

1. computational linguistics
2. information retrieval (wyszukiwanie informacji)
3. content analysis (analizę treści)
4. natural language processing – przetwarzanie języka naturalnego

Drzewo decyzyjne



Web - mining


Web mining
Web content mining
Web structure mining
Web usage mining
Patterns through Web crawling, real data from the web
Patterns through nested webpages through hyperlink
Hyperlink
Pattern through web server logs

Web mining

Celem Web Mining

jest dostarczenie algorytmu lub techniki, by zwiększyć wydajność i wygodę dostępu do danych.

Techniki Web Mining dzieli się na:

1. **Web Content Mining (WCM)** - odnosi się do wydobywania, ekstrakcji i integracji cennych informacji (wiedzy) z zawartości stron internetowych,
2. **Web Structure Mining (WSM)** - celem jest wydobycie nieznanych wcześniej zależności pomiędzy stronami internetowymi, które należą do jednej lub wielu witryn,
3. **Web Usage Mining (WUM)** - próbuje wydobyć, odkryć i przeanalizować interesujące dostępy, transakcje, wzorce strumieni kliknięć i inne powiązane dane generowane z interakcji użytkownika z zasobami sieciowymi. Dane przechowywane w pliku tekstowego„log”, który znajduje się na serwerze sieciowym.

Plan prezentacji

1. Text analytics - Analiza tekstu
   - Web - mining
2. Social Network Analysis
   - Wprowadzenie
   - Pojęcia i definicje
3. Algorytm PageRank
   - Wprowadzenie

Problemy z wyszukiwaniem informacji

- Liczba stron internetowych gwałtownie wzrosła, a każde zapytanie w wyszukiwarce zwraca ogromną liczbę stron. Ta mnogość informacji powoduje problem z klasyfikacją, tj. jak wybrać tylko 10-30 stron i je uszeregować do przedstawienia użytkownikowi?
- Metody wyszukujące podobnych treści są łatwe do „zespamowania”. Właściciel strony może powtórzyć kilka słów kluczowych i dodać wiele luźno powiązanych haseł, aby podnieść rangę strony i wymusić podobieństwo strony do wielu zapytań.

Wykorzystanie linków (hiperłączy)

- Tradycyjnym wyszukiwanie informacji pomija w analizie linki - zajmuje się tylko analizą dokumentów tekstowych.
- WSM związana z analizą łączy/linków. Wyróżniamy dwa rodzaje hiperłączy:
  - **natywne** - wykorzystywane do uporządkowania dużej ilości informacji na tej samej stronie internetowej, a zatem wskazują tylko na podstrony.
  - **obce** wskazują na strony w innych witrynach. Hiperłącza wychodzące wskazują na domniemane przeniesienie uprawnień na strony, na które są skierowane. W związku z tym strony, na które wskazuje wiele innych stron, mogą zawierać ważne lub wysokiej jakości informacje. Takie linki powinny być zatem wykorzystywane w ocenie i rankingu w wyszukiwarkach internetowych.

Algorytmy PageRank i HITS

- **PageRank** jest algorytmem, który zasila wyszukiwarkę Google. PageRank działa poprzez liczenie liczby i jakości linków prowadzących do strony w celu określenia przybliżonego oszacowania, jak ważna jest dana witryna. Podstawowym założeniem jest to, że ważniejsze witryny prawdopodobnie otrzymają więcej linków z innych witryn.
- **Hyperlink-Induced Topic Search (HITS**) - każdej stronie internetowej przypisane są dwa wyniki, jeden nazywany jest wynikiem autorytetu, a drugi wynikiem hubów. Autorytet mają strony, które zawierają znaczące informacje o temacie zapytania, natomiast strony, które wskazują na wiele stron z autorytetem nazywane są habami i są użytecznymi zasobami w sieci.

Algorytmy PageRank i HITS

PageRank i HITS pochodzą z analizy serwisów społecznych. Oba wykorzystują strukturę hiperłączy, aby uszeregować strony zgodnie z ich poziomem „ważności”.

Ocena i ranking stron opartych na hiperłączach nie jest jedyną metodą stosowaną przez wyszukiwarki internetowe. Treść i wiele innych czynników jest również brana pod uwagę przy tworzeniu ostatecznego rankingu przedstawionego użytkownikowi.

Social Network

Sieć społeczna

– struktura społeczna złożona z węzłów, które są indywidualnymi elementami organizacji. Węzły z kolei połączone są poprzez różne rodzaje powiązań – od przypadkowych spotkań do bliskich relacji rodzinnych. Termin użyty pierwszy raz w 1954 roku przez J.A. Barnesa

Analiza sieci społecznych

To badanie podmiotów społecznych (ludzi w organizacji, zwanych aktorami) oraz ich interakcji i relacji.

Sieć społeczna - reprezentacja

Interakcje i relacje mogą być reprezentowane za pomocą sieci lub grafu, gdzie każdy wierzchołek (lub węzeł) reprezentuje aktora, a każda krawędź reprezentuje związek.

Social Network

Analiza sieci społeczna

W sieci można badać

1. właściwości jej struktury, a także rolę, pozycję i prestiż każdego aktora;
2. znaleźć różne podgrafy, np. społeczności tworzone przez grupy aktorów.

Jest przydatna w sieci Internet, ponieważ WWW jest jak *social network*, gdzie strona może być traktowana jako aktor, a hiperłącze jako relacja.

Sieć powiązań – pojęcia

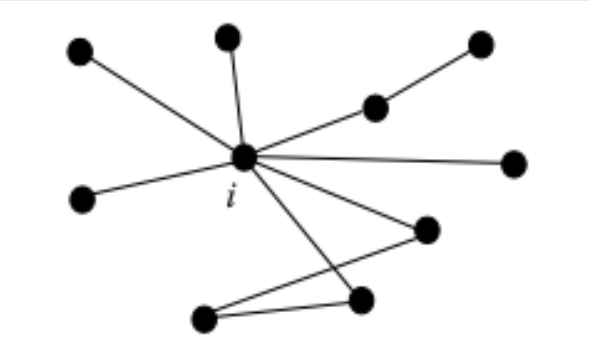
Centralizm - bycie ważnym może być widziane z różnych perspektyw:

- Ważnymi lub wyróżniającymi się aktorami są Ci, którzy są w dużym stopniu powiązani (link) z innymi aktorami.
- W kontekście organizacji za ważniejszą osobę uważa się osobę o szerokich kontaktach (powiązaniach - links) lub komunikującą się z wieloma innymi osobami w organizacji niż osobę o relatywnie małej liczbie kontaktów.
- Linki te mogą być również nazywane wiązaniami (ties). Głównym aktorem jest osoba zaangażowana w wiele powiązań.

Sieć powiązań – pojęcia

Example

Przykład z wykorzystaniem grafu nieskierowanego



- Każdy węzeł w sieci jest aktorem a krawędź wskazuje, że aktorzy na końcach połączeń komunikują się ze sobą.
- Aktor *i* jest najbardziej *centralny*, ponieważ może komunikować się z większością innych aktorów.

Centrality - part 1

Aktorzy środkowi (centralni) są najbardziej aktywni i mają najwięcej powiązań z innymi aktorami.

Degree Centrality - stopień powiązań w grafie nieskierowanym

$C_D(i)$ — *Degree centrality* i-tego aktora znormalizowanym liczbą połączeń $(n-1)$, to:

$$C_D(i) = \frac{d(i)}{n-1}$$

gdzie:

- n całkowita liczba aktorów w sieci
- $d(i)$ stopień połączeń aktora liczony liczbą krawędzi

Centrality - part 2

Należy rozróżnić linki do aktora i (krawędzie wskazujące na i) oraz linki od (linki wskazujące z i).

Degree Centrality - na podstawie połączeń wychodzących

$C'_D(i)$ — *Degree centrality* typu output i-tego aktora definiowane tylko na podstawie połączeń :

$$C'_D(i) = \frac{d_o(i)}{n-1}$$

gdzie:

- n całkowita liczba aktorów w sieci
- $d_o(i)$ stopień połączeń aktora liczony liczbą krawędzi **wychodzących**

Closeness Centrality - part 1

- Miara oparta jest na bliskości którą można mierzyć jako odległość.
- Aktor *i*-ty jest *centralny*, jeśli może łatwo wchodzić w interakcję z wszystkimi innymi aktorami.
- Ideą miary jest to, że dystans aktora *i*-tego do wszystkich innych aktorów jest mały.

Closeness Centrality - w grafie nieskierowanym

$C_C(i)$ — *Closeness Centrality* $i-$ *tego* aktora mierzona jako liczba linków w najkrótszej ścieżce to:

$$C_C(i) = \frac{n-1}{\sum_{j=1}^{n} d(i,j)}$$

gdzie:

- n całkowita liczba aktorów w sieci
- $d(i,j)$ najkrótsza odległość od aktora i do aktora j

Closeness Centrality - part 2

Closeness Centrality:

- mieści się w przedziale od 0 do 1, ponieważ $n - 1$ jest minimalną wartością mianownika, który jest sumą najmniejszych odległości od i do wszystkich innych aktorów.
- Równanie można stosować tylko dla grafu.
- To samo równanie może być użyte dla grafu **skierowanego**. Przy obliczaniu odległości należy uwzględnić kierunki połączeń.

Betweenness Centrality – part 1

- Jeśli dwaj nie sąsiadujący ze sobą aktorzy j i k chcą wejść w interakcję, a aktor i jest pomiędzy j i k, to i może mieć kontrolę nad ich połączeniem.
- BC - mierzy kontrolę i nad innymi parami aktorów. Tak więc, jeśli i węzeł jest na ścieżce wielu takich interakcji, to jestem ważnym aktorem.

Betweenness Centrality

$C_B(i)$ — *Betweenness centrality* aktora i jest zdefiniowana jako:

$$C_B(i) = \sum_{j<k} \frac{p_{jk}(i)}{p_{jk}}$$

gdzie:

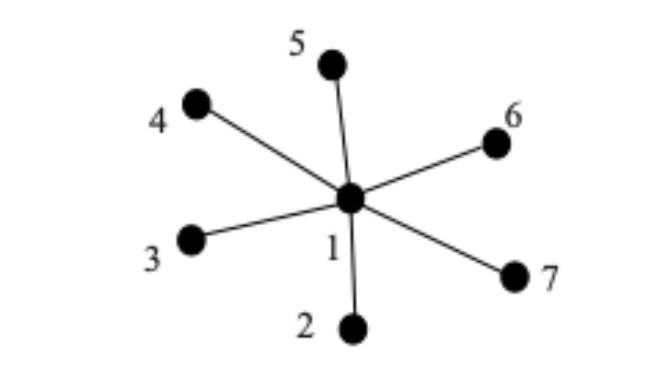
- $p_{jk}(i)$ liczba najkrótszych ścieżek, które przechodzą przez i i $j \neq i$ i $k \neq i$
- p_{jk} będzie całkowitą liczbę najkrótszych ścieżek wszystkich par aktorów j i k nie włączając i

Betweenness Centrality – part 2

- Między aktorem j a aktorem k może istnieć wiele ścieżek. Niektóre z nich przechodzą i, a inne nie.
- Zakładamy, że wszystkie ścieżki są jednakowo prawdopodobne.
- Gdy i nie znajduje się na najkrótszej ścieżce, to $C_B(i) = 0$ która jest wartością minimalną.
- Maksymalna wartość $C_B(i)$ to $(n-1)(n-2)/2$ czyli liczba par aktorów bez i (aktor jest wykluczony).

Betweenness Centrality – part 3

Example



- Aktor 1 jest najbardziej centralnym aktorem mierzony $C_B(i)$.
- Aktor 1 leży na wszystkich 15 najkrótszych ścieżkach łączących pozostałych 6 aktorów.
- $C_B(1)$ ma maksymalną wartość 15.
- $C_B(2) = C_B(3) = C_B(4) = C_B(5) = C_B(6) = C_B(7) = 0$.

Prestige

- Prestiż jest bardziej wyrafinowaną miarą wyeksponowania aktora niż centralizm.
- Odróżnia linki wychodzące i przychodzące.
- Prestiżowy aktor to ten, który jako odbiorca jest obiektem rozbudowanych więzi.
- Aby obliczyć prestiż aktora, wystarczy spojrzeć na linki skierowane lub wskazane do aktora.
- Relacja zatem wymaga grafu skierowanego.
- Główna różnica pomiędzy pojęciami centralizmu i prestiżu: centralizm skupia się na linkach wychodzących, a prestiż na przychodzących.
- Definiujemy trzy miary prestiżu. Miara *rank prestige* stanowi podstawę większości algorytmów analizy linków stron internetowych, w tym PageRank i HITS.

Degree Prestige

Degree Prestig

$P_D(i))$ — *Degree Prestig* jest najprostszą miarą prestiżu aktora i liczona jako jego stopień połączeń przychodzących (in-linków lub nominacji):

$$P_D(i) = \frac{d_I(i)}{n-1}$$

gdzie:

- $d_I(i)$ jest stopniem (liczbą połączeń przychodzących do i)
- n całkowita liczba aktorów w sieci

Podobnie jak w przypadku centalizacji, podzielenie przez $n-1$ normalizuje wartość prestiżu do zakresu od 0 do 1. Maksymalna wartość prestiżu wynosi 1, gdy każdy inny aktor łączy się lub wybiera aktora i.

Proximity Prestige – założenia

- *Degree Prestig* aktora *i* bierze pod uwagę tylko aktorów, którzy sąsiadują z *i*.
- *Proximity Prestige* uogólnia go poprzez uwzględnienie zarówno aktorów bezpośrednio, jak i pośrednio powiązanych z aktorem *i*. Oznacza to, że bierzemy pod uwagę każdego aktora *j*, który może dotrzeć do *i*, tj. istnieje w grafie ścieżka od *j* do *i*.
- Proximity = Bliskość jest definiowana jako odległość innych aktorów do *i*.

Proximity Prestige - średnia odległość w grafie

Niech:

- I_i to domena wpływu aktora i to zespół aktorów, który może dotrzeć do aktora i
- $d(j, i)$ to najkrótsza droga od aktora j do aktora i. Każde połączenie ma długość w pewnej określonej jednostce odległości.
- średnia odległość węzłów j od i to

$$\frac{\sum_{j \in I_i} d(j, i)}{|I_i|},$$

gdzie: $|I_i|$ jest rozmiarem zbioru I_i

Proximity Prestige - definicja

$P_P(i)$ — Proximity Prestige patrzy na proporcję aktorów, którzy mogą osiągnąć i do średniej odległości, jaką ci aktorzy mają od i:

Proximity Prestige

$$P_P(i) = \frac{\frac{|I_i|}{(n-1)}}{\frac{\sum_{j \in I_i} d(j,i)}{|I_i|}}$$

gdzie:

- $\frac{|I_i|}{(n-1)}$ jest proporcją aktorów, którzy mogą dotrzeć do aktora i
- n całkowita liczba aktorów w sieci

Proximity Prestige - własności

- $P_P(i)$ ma zakres wartości [0, 1].
- Jeżeli każdy aktor może dotrzeć do aktora i, to $\frac{|I_i|}{(n-1)} = 1$.
- Mianownik jest równy 1 jeśli każdy aktor przylega do i.
- $P_P(i) = 1$ jeżeli zachodzą dwa powyższe przypadki.
- Jeżeli żaden aktor nie może dotrzeć do aktora i, to $\frac{|I_i|}{(n-1)} = 0$ i $P_P(i) = 0$.

Rank Prestige – intuicja

1. Dwa wcześniejsze mierniki prestiżu opierają się na stopniach i dystansach.
2. Ważnym czynnikiem, który nie został uwzględniony, jest znaczenie poszczególnych aktorów, którzy dokonują *głosowania* lub *wyboru*. W świecie rzeczywistym osoba *i*, którą wybrano przez osobę ważną, jest bardziej prestiżowa, niż gdyby była wybrana przez osobę mniej ważną.
3. Prestiż zależy więc od rangi lub statusu zaangażowanych aktorów.

Example

Dyrektor generalny firmy głosujący na daną osobę jest znacznie ważniejszy niż inny pracownik głosujący na tę osobę. Jeżeli w kręgu wpływów jest pełno prestiżowych aktorów, to prestiż własny również jest wysoki.

Rank Prestige – definicja

Prestiż rangi aktora jest funkcją rangi aktorów, którzy głosują lub wybierają aktora. $P_R(i)$ — *Rank Predtige* definiuje się jako liniowe połączenie powiązań, które wskazują na i:

Rank Prestige

$$P_R(i) = A_{1i}P_R(1) + A_{2i}P_R(2) + \ldots + A_{ni}P_R(n)$$

gdzie:

- $A_{ij} = 1$ jeśli j wskazuje na i, a 0 w przeciwnym wypadku
- n całkowita liczba aktorów w sieci

Rank Prestige – definicja

Ponieważ istnieje n równań dla n aktorów, można użyć notacji macierzowej.

Rank Prestige - macierzowo

$$\boldsymbol{P} = \boldsymbol{A}^T \boldsymbol{P},$$

gdzie:

- $\boldsymbol{P} = (P_R(1), P_R(2), \ldots, P_R(n))^T$
- A_{ij} jest macierzą zawierającą 0 lub 1.

Plan prezentacji

1. Text analytics - Analiza tekstu
   - Web - mining
2. Social Network Analysis
   - Wprowadzenie
   - Pojęcia i definicje
3. Algorytm PageRank
   - Wprowadzenie

Wprowadzenie

- PageRank został zaprezentowany przez Sergeya Brina i Larry'ego Page'a na siódmej międzynarodowej konferencji World Wide Web (WWW7) w kwietniu 1998 roku.
- W oparciu o algorytm zbudowali oni wyszukiwarkę **Google**.
- PageRank stał się dominującym modelem analizy linków w wyszukiwarkach internetowych, dzięki
  - niezależnej od zapytań ocenie stron internetowych
  - zdolności do zwalczania spamu (nieuczciwego wpływu autorów stron na rankingi)
  - częściowo dzięki sukcesowi biznesowemu Google.

PageRank – Założenia

Algorytm PageRank:

- wykorzystuje demokratyczny charakter sieci, wykorzystując jej szeroką strukturę linków jako wskaźnik jakości poszczególnych stron.
- interpretuje hiperłącze od strony *x* do strony *y* jako niezależy głos oddany przez stronę *x*, dla strony *y*.
- analizuje również stronę, która oddała głos. Głosy oddane przez strony, które same są *ważne*, ważą więcej i pomagają zwiększyć *ważniejszymi*. A to jest właśnie idea *Rank Prestige* w social network.
- PageRank jest statycznym rankingiem stron internetowych a wartość PageRank jest obliczana dla każdej strony w trybie off-line i nie zależy od zapytań wyszukiwania.

Główne pojęcia w kontekście WWW

In-link – linki przychodzące

Hiperłącza, które wskazują na stronę *i* z innych stron WWW. Zazwyczaj, hiperłącza wewnętrzne z tej samej strony nie są brane pod uwagę.

Out-links – linki wychodzące

Są to hiperłącza, które wskazują na inne strony ze strony *i*. Zazwyczaj, linki do stron tej samej witryny nie są brane pod uwagę.s

PageRank - elementy

Aby wyprowadzić PageRank niezbędne jest:

- Hiperłącze ze strony wskazującej na inną stronę jest wewnętrznym przekazaniem istotności do strony docelowej. Tak więc, im więcej in-linków otrzymuj strona *i*, tym większy ma prestiż.
- Strony, które wskazują na stronę *i* mają również swój własny prestiż. Strona z wyższym prestiżem wskazuje na *i* jest ważniejsza niż strona z niższym prestiżem wskazująca na *i*. Innymi słowy, strona jest ważna, jeśli jest wskazywana przez inne ważne strony.

Sieć WWW jako graf skierowany

W zależności od rangi prestiżu w social network, o znaczeniu strony i (PageRank score dla i) jest obliczany przez zsumowanie wyników PageRank wszystkich stron, które wskazują na stronę i. Ponieważ strona może wskazywać na wiele innych stron, jej prestiż powinien być dzielony pomiędzy wszystkie strony, na które wskazuje. Zwróć uwagę na różnicę w stosunku do prestige score, gdzie score nie jest dzielony.

PageRank score

Aby sprostać powyższym założeniom, traktujemy Sieć jako skierowany wykres $G = (V, E)$, gdzie V to zbiór wierzchołków lub węzłów, czyli zbiór wszystkich stron, a E to zbiór skierowanych krawędzi, czyli hiperłączy. Niech całkowita liczba stron w sieci to n (tzn. $n = |V|$). PageRank strony i można wyrazić przez:

$$P(i) = \sum_{(j,i) \in E} \frac{P(j)}{O_j} \tag{1}$$

gdzie O_j jest liczbą linków zewnętrznych strony j.

PageRank score

Rozwijając równanie 1 mamy układ n równań liniowych z n niewiadomymi. Możemy użyć macierzy do reprezentacji wszystkich równań. Niech P będzie n-wymiarowym wektorem zawierającym w kolumnach wartości PageRank,

$$\boldsymbol{P} = (P(1), P(2), \ldots, P(n))^T$$

Niech A będzie macierzą pomocniczą:

$$A_{ij} = \begin{cases} \frac{1}{O_i} & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Zatem:

$$\boldsymbol{P} = \boldsymbol{A}^T \boldsymbol{P} \tag{3}$$

Interpretacja P

Jest to układ równań gdzie rozwiązanie P jest wektorem własnym o odpowiadającej mu wartości własnej 1. Ponieważ jest to definicja zagnieżdżona, do jej rozwiązania stosuje się algorytm iteracyjny. Jeżeli spełnione są pewne warunki, to 1 jest największą wartością własną, a wektor PageRank P jest głównym wektorem własnym.

Problem polega jednak na tym, że równanie 3 nie jest wystarczające, ponieważ graf WWW nie spełnia warunków. Aby uwzględnić wszystkie warunki i rozwinąć równanie 3, wyprowadzimy to samo równanie oparte na łańcuchu Markowa.

PegeRank - łańcuchy Markowa

W modelu łańcucha Markowa, każda strona WWW jest traktowana jako stan. Hiperłącze to przejście, które z pewnym prawdopodobieństwem prowadzi z jednego stanu do drugiego. Takie podejście modeluje surfowanie w sieci jako proces stochastyczny. Modeluje on surfowanie w sieci jako losowe przejście przez stany w łańcuchu Markowa.

Jeżeli O_i to liczba out-linków węzła i, to prawdopodobieństwo przejścia jest równe $1/O_i$ przy założeniu, że surfer będzie klikał hiperłącza na stronie i z rozkładem jednostajnym. Przycisk „wstecz” nie jest używany i nie wpisuje adresu URL.

PegeRank - łańcuchy Markowa cd

Niech A będzie macierzą prawdopodobieństwa przejścia pomiędzy stanami w łańcuchu Markowa:

$$\boldsymbol{A} = \begin{pmatrix} A_{11} & A_{12} & . & . & . & A_{1n} \\ A_{21} & A_{22} & . & . & . & A_{2n} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ A_{n1} & A_{n2} & . & . & A_{nn} & \end{pmatrix}$$

A_{ij} przedstawia prawdopodobieństwo przejścia surfera w stanie i (strona i) do stanu j (strona j). A_{ij} jest zdefiniowane dokładnie tak, jak w równaniu 2.

PegeRank - łańcuchy Markowa cd

Biorąc pod uwagę początkowy wektor rozkładu prawdopodobieństwa, że surfer jest w każdym stanie (lub stronie)
$\boldsymbol{p}_0 = (p_0(1), p_0(2), ..., p_0(n))^T$ (wektor wektorów) oraz macierz $n \times n$ prawdopodobieństwa przejścia A:

$$\sum_{i=1}^{n} p_0(i) = 1 \tag{4}$$

$$\sum_{j=1}^{n} A_{ij} = 1 \tag{5}$$

Równanie 5 nie jest do końca prawdziwe dla niektórych stron internetowych, ponieważ nie mają one linków zewnętrznych. Jeśli macierz A spełnia równanie 2, to mówimy, że A jest stochastyczną macierzą łańcucha Markowa. Załóżmy, że A spełnia równanie 5.

PegeRank - łańcuchy Markowa cd

Biorąc pod uwagę początkowy rozkład prawdopodobieństwa p_0, jakie jest prawdopodobieństwo, że m kroków/przejść później będziemy w każdym stanie j? Chcemy określić prawdopodobieństwo, że system (lub przypadkowy surfer) znajduje się w stanie j po 1 kroku przy pomocy poniższego rozumowania:

$$p_1(j) = \sum_{i=1}^{n} A_{ij}(1) p_0(i)$$

gdzie $A_{ij}(1)$ jest prawdopodobieństwem przejścia z i do j w 1 przejściu, a $A_{ij}(1) = A_{ij}$. Możemy to napisać:

$$\boldsymbol{p}_1 = \boldsymbol{A}^T \boldsymbol{p}_0 \tag{6}$$

Uogólniając:

$$\boldsymbol{p}_k = \boldsymbol{A}^T \boldsymbol{p}_{k-1} \tag{7}$$

PegeRank - łańcuchy Markowa cd

Według Ergodycznego twierdzenia, skończony łańcuch Markowa zdefiniowany przez stochastyczną macierz przejść A ma unikalny stacjonarny rozkład prawdopodobieństwa, jeśli A jest nieredukowalny i aperiodyczny.
Stacjonarny rozkład prawdopodobieństwa oznacza, że po serii przejść p_k zbiegną do wektora prawdopodobieństwa stanu ustalonego niezależnie od wyboru początkowego wektora prawdopodobieństwa p_0, tzn,

$$\lim_{k \to \infty} \boldsymbol{p_k} = \pi$$

Kiedy dotrzemy do stanu ustalonego, to $\boldsymbol{p_k} = \boldsymbol{p_{k+1}} = \pi$, a więc $\pi = \boldsymbol{A^T}\pi$. π jest wektorem własnym $\boldsymbol{A}^T$ o wartości własnej 1. W PageRank, jest on używany jako wektor PageRank P. Tak więc, otrzymujemy równanie 3:

$$\boldsymbol{P} = \boldsymbol{A}^T \boldsymbol{P} \tag{8}$$

PegeRank - łańcuchy Markowa cd

Użycie stacjonarnego rozkładu prawdopodobieństwa jako wektora PageRank jest rozsądne i dość intuicyjne, ponieważ odzwierciedla długookresowe prawdopodobieństwa, że przypadkowy surfer odwiedzi wszystkie strony. Strona ma wysoki prestiż, jeśli prawdopodobieństwo jej odwiedzenia jest wysokie.

Wróćmy teraz do prawdziwego kontekstu WWW i zobaczmy, czy powyższe warunki są spełnione, tzn. czy $\boldsymbol{A}$ jest nieredukowalna i aperiodyczna. W rzeczywistości, żaden z warunków nie jest spełniony. Dlatego musimy rozszerzyć równanie 8, aby stworzyć „rzeczywisty model PageRank”. Rozpatrzmy każdy z warunków.

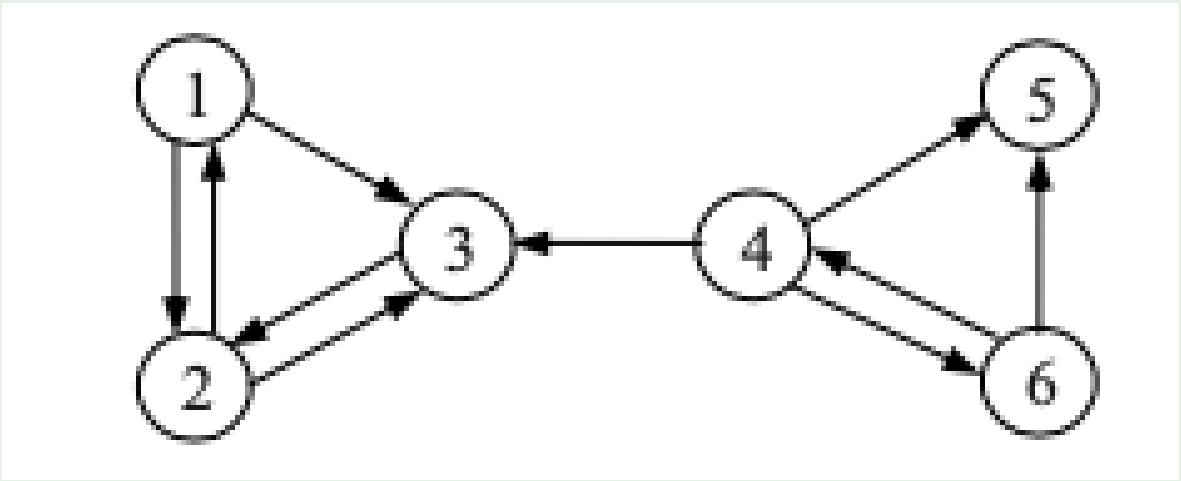
PegeRank - łańcuchy Markowa cd

Po pierwsze, A nie jest macierzą stochastyczną (przejść). Macierz przejść dla skończonego łańcucha Markowa, gdzie w każdym wierszu wartości są nieujemnymi liczbami rzeczywistymi i sumują się do 1. Wymaga to, aby każda strona internetowa posiadała przynajmniej jedno łącze zewnętrzne. W sieci nie jest to spełnione, ponieważ wiele stron nie ma odnośników zewnętrznych, co powoduje, że w macierzy A wiersze wypełnione są 0. Takie strony są nazywane dangling.

PegeRank - łańcuchy Markowa cd

Example

Rozważmy przykład grafu linków.



PegeRank - łańcuchy Markowa cd

Example

Jeśli założymy, że internauta będzie klikał hiperłącza na stronie jednolicie losowo, mamy następującą macierz prawdopodobieństwa przejścia:

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

$A_{12} = A_{13} = 1/2$, ponieważ węzeł 1 ma dwa połączenia zewnętrzne. Widzimy, że A nie jest macierzą stochastyczną, ponieważ piąty wiersz to tylko 0, tzn. strona 5 jest dangling.

PegeRank - łańcuchy Markowa cd

Możemy rozwiązać ten problem na kilka sposobów. Dwa z nich to:

1. Usuń strony bez linków zewnętrznych z systemu podczas obliczania PageRank, ponieważ strony te nie mają bezpośredniego wpływu na ranking innych stron. Out-linki z innych stron wskazujących na te strony są również usuwane. Po obliczeniu PageRanks, strony te i hiperłącza do nich wskazujące mogą zostać dodane. PageRanks są łatwe do obliczenia w oparciu o równanie 8. Zauważ, że prawdopodobieństwo przejścia tych stron z usuniętymi linkami będzie nieznacznie zmienione, ale nie znacząco.
2. Dodaj pełny zestaw linków wychodzących z każdej takiej strony i do wszystkich stron w sieci. Stąd prawdopodobieństwo przejścia z i na każdą stronę wynosi $1/n$ przy założeniu równomiernego rozkładu prawdopodobieństwa. Oznacza to, że każdy wiersz zawierający same 0 zamieniamy na e/n, gdzie e jest wektorem n-wymiarowym samych 1.

PegeRank - łańcuchy Markowa cd

Example

Jeśli użyjemy drugiej metody do stworzenia macierzy stochastycznej poprzez dodanie linku od strony 5 do każdej strony, otrzymamy:

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

$A_{12} = A_{13} = 1/2$, ponieważ węzeł 1 ma dwa połączenia zewnętrzne. Widzimy, że A nie jest macierzą stochastyczną, ponieważ piąty wiersz to tylko 0, tzn. strona 5 jest dangling.

PegeRank - łańcuchy Markowa cd

Po drugie, A A nie jest nieredukowalne. Nieredukowalny oznacza, że graf G jest silnie powiązany.

Graf silnie powiązany

Graf skierowany $G = (V, E)$ jest silnie powiązany, wtedy i tylko wtedy, gdy dla każdej pary węzłów $u, v \in V$, istnieje ścieżka od u do v.

Graf WWW reprezentowany przez A nie jest nieredukowalny, ponieważ dla niektórych par węzłów u i v nie ma ścieżki od u do v.

Example

Na przykład w poprzednim przykładnie nie ma ścieżki od węzła 3 do węzła 4. Korekta dotycząca dangling node nie jest wystarczająca, aby zapewnić nieredukowalność, nadal nie ma połączenia od węzła 3 do węzła 4.

PegeRank - łańcuchy Markow cd

Po trzecie, A nie jest aperiodyczna. Stan i w łańcuchu Markowa jest okresowy oznacza, że istnieje cykl, który łańcuch musi przejść.

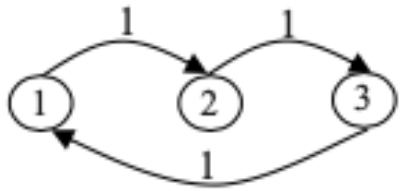
Graf aperiodyczny

Stan i jest okresowy z okresem $k > 1$ jeśli k jest najmniejszą liczbą taką, że wszystkie ścieżki prowadzące ze stanu i z powrotem do stanu i mają długość będącą wielokrotnością k. Jeśli stan nie jest okresowy (tj. $k = 1$), to jest aperiodyczny. Łańcuch Markowa jest aperiodyczny, jeśli wszystkie stany są aperiodyczne.

PegeRank - łańcuchy Markowa cd

Example

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



Rysunek przedstawia okresy łańcuch Markowa z $k = 3$. Macierz przejścia jest podana po lewej stronie. Każdy stan w tym łańcuchu ma okres 3. Na przykład, jeżeli zaczniemy od stanu 1, to aby powrócić do stanu 1, jedyną ścieżką jest 1-2-3-1 przez pewną liczbę razy, powiedzmy h. Tak więc każdy powrót do stanu 1 będzie wymagał $3h$ przejść. W sieci WWW może być wiele takich przypadków.

PegeRank - łańcuchy Markow cd

Problem nieredukowalności i aperiodyczności można rozwiązać pojedynczą strategią:

System naprawczy

Dodajemy link z każdej strony do każdej strony i nadajemy każdemu linkowi małe prawdopodobieństwo przejścia kontrolowane przez parametr d.

Macierz naprawiona staje się nieredukowalna, ponieważ jest silnie powiązana. Jest również aperiodyczna, ponieważ sytuacja z przykładu przestaje istnieć, ponieważ mamy teraz ścieżki o każdej możliwej długości od stanu i z powrotem do stanu i. Oznacza to, że przypadkowy surfer nie musi przechodzić przez stały cykl dla dowolnego stanu.

PegeRank - łańcuchy Markowaa cd

Po tej naprawie otrzymujemy ulepszony model PageRank. W modelu tym, na stronie, losowy surfer ma dwie opcje:

1. Z prawdopodobieństwem d, losowo wybiera out-link do przejścia.
2. Z prawdopodobieństwem $1 - d$, przeskakuje na losową stronę nie korzystając z odnośnika.

Równanie ulepszonego modelu:

$$\boldsymbol{P} = \left((1-d)\frac{1}{n}\boldsymbol{E} + d\boldsymbol{A}^T \right) \boldsymbol{P} \tag{9}$$

gdzie $\boldsymbol{E}$ jest macierzą kwadratową $n \times n$ wypełnioną 1. $1/n$ jest prawdopodobieństwem skoku na daną stronę, gdzie n jest całkowitą liczbą węzłów na grafie internetowym. Zakłada się, że $\boldsymbol{A}$ jest daną macierzą prawdopodobieństw przejść.

PegeRank - łańcuchy Markowa cd

Example

Rozważając poprzedni przykład i wykorzystując $\bar{\boldsymbol{A}}$ jako $\boldsymbol{A}$ oraz $d = 0.9$ we wzorze (9) uzyskujemy:

$$(1-d)\frac{\boldsymbol{E}}{n} + d\boldsymbol{A}^T = \begin{pmatrix} 1/60 & 7/15 & 1/60 & 1/60 & 1/6 & 1/60 \\ 7/15 & 1/60 & 11/12 & 1/60 & 1/6 & 1/60 \\ 7/15 & 7/15 & 1/60 & 19/60 & 1/6 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 1/6 & 7/15 \\ 1/60 & 1/60 & 1/60 & 19/60 & 1/6 & 7/15 \\ 1/60 & 1/60 & 1/60 & 19/60 & 1/6 & 1/60 \end{pmatrix}$$

PegeRank - łańcuchy Markowa cd

Jeżeli przeskalujemy równanie (9) stosując $\boldsymbol{e}^T \boldsymbol{P} = n$ uzyskamy:

$$\boldsymbol{P} = (1-d)\boldsymbol{e} + d\boldsymbol{A}^T \boldsymbol{P} \tag{10}$$

To daje formułę na PageRank dla każdej strony i:

$$P(i) = (1-d) + d \sum_{j=1}^{n} A_{ji} P(j) \tag{11}$$

co jest równoważne z formułą podaną w literaturze:

$$P(i) = (1-d) + d \sum_{(i,j)\in E} \frac{P(j)}{O_j} \tag{12}$$

Parametr d jest nazywany współczynnikiem tłumienia, który można ustawić w zakresie od 0 do 1, najczęściej $d = 0,85$.

PegeRank – algorytm

Obliczanie wartości PageRank dla stron internetowych może być wykonane przy użyciu metody „power iteration method”, która wyliczy wektor wartości własnych (1). Algorytm jest stosunkowo prosty:

PageRank-Iterate(G)
$\boldsymbol{P}_0 \leftarrow \boldsymbol{e}/n$
$k \leftarrow 1$
repeat
$\boldsymbol{P}_k \leftarrow (1-d)\boldsymbol{e} + d\boldsymbol{A}^T \boldsymbol{P}_{k-1}$;
$k \leftarrow k+1$;
until $||\boldsymbol{P}_k - \boldsymbol{P}_{k-1}||_1 < \varepsilon$
return $\boldsymbol{P}_k$

Można go rozpocząć od początkowego przypisania wartości PageRank. Iteracja kończy się, gdy wartości PageRank nie zmieniają się zbytnio lub są zbieżne. Praktycznie jeżeli suma wszystkich składowych jest mniejsza od zadanego ϵ

PegeRank – zalety

1. Zdolność do walki ze spamem – Strona jest ważna, jeśli strony na nią wskazujące są ważne. Ponieważ nie jest łatwo właścicielowi strony dodać in-linki do swojej strony z innych ważnych stron, nie jest łatwo wpłynąć na PageRank.
2. Jest to miara globalna i niezależna od zapytań – wartości PageRank dla wszystkich stron w sieci są obliczane i zapisywane w trybie off-line, a nie w czasie zapytania. W czasie składania zapytań, tylko wyszukiwanie jest potrzebne, aby znaleźć wartość, która ma być zintegrowana z innymi strategiami rankingu. Skutkuje dużą wydajnością.

PegeRank – wady

1. Niezależność PageRank od zapytań – Nie odróżnia stron, które są ważne w ogóle od stron, które są ważne w temacie zapytania. Google ma sposoby na radzenie sobie z tym problemem.
2. Ranking oparty na linkach nie jest jedyną strategią stosowaną w wyszukiwarce. Stosowanych jest również wiele innych metod wyszukiwania informacji, heurystyki i parametry empiryczne.

Źródła

1. Eric Luellen, „An Updated Text Analytics Primer: Key Factors in a Text Analytics Strategy”, 2019
2. Data Mining vs Web Mining
   `https://www.educba.com/data-mining-vs-web-mining/`
3. Bing Liu, „Web Data Mining”, Second Edition 2011