Wyszukiwanie i Przetwarzanie Informacji WWW

Analiza linków (1): Algorytm HITS

Marcin Sydow

PJWSTK

Plan tego wykładu

- Przypomnienie: Ranking dokumentów w wyszukiwarkach
- Podstawy racjonalne analizy linków w liczeniu rankingu
- Idea algorytmu HITS
- Sformułowanie HITS
- Analiza
- Rozszerzenia
- Wybrana literatura dodatkowa
- Znajdowanie Dokumentów Podobnych
- Zastosowania HITS w Systemach Reputacyjnych

Moduły wyszukiwarki

- Moduł zbierający (ang. Crawler)
  - podążaj po linkach i ściągaj dokumenty
- Repozytorium
  - składuj ściągnięte dokumenty - trwałość, dostęp
- Indeks
  - zapisz które słowo występuje w jakim dokumencie
- **System Rankingowy**
  - **jakie informacje dobrze pasują do zapytania użytkownika?**
  - **jakie informacje są wartościowe same w sobie?**
- Moduł prezentacji
  - znajdź dobrą formę wizualizacji wyników
- Obsługa
  - obsłuż zapytania, znajdź strony, wyświetl wyniki

Szukanie igły w stogu siana - Ranking

Przeciętne zapytanie: **tysiące zwróconych** dokumentów

Możliwości użytkownika: **kilkanaście obejrzanych** dokumentów

Jak wybrać **na początek listy te kilkanaście najlepszych** spośród tysięcy?

Rozwiązaniem jest: System Rankingowy

Systemy rankingowe istniały od lat w IR, ale nie były idealne w przypadku WWW
(rewolucja wyszukiwarkowa AD 1998)

Ranking

Najpilniej strzeżone tajemnice wyszukiwarek (decydują o **jakości wyników**)

Dokumentowi przyporządkowana jest wartość (ang. score) i wyniki są posortowane po tej wartości

Wiele składowych:

- analiza tekstu (zawartość, URL, meta, ...)
- analiza tekstu odnośników (ang. anchor text)
- **analiza struktury linków**
- analiza logów, ruchu internetowego, ...

Tekst a ranking

- statystyki (np. tf-idf)
- pozycja w tekście
- pozycja w kontekście (URL, meta, title, anchor, etc.)
- meta-znaczniki
- znaczniki prezentacji (rozmiar, pogrubienie nagłówek)

WWW - problemy z tekstem

Klasyczne, tekstowe techniki IR sprawiają problemy w przypadku WWW:

- Problem **braku samo-opisu**
  (np. zapytanie: "japoński producent samochodów")
- Problem różnorodności
- Problem nierównej jakości
- Zaszumienie, błędy, etc
- Tekst - **łatwy do spamowania**

WWW - rozwiązanie problemów IR

WWW z jednej strony **stwarza problemy** dla klasycznego IR. Z drugiej strony, **stwarza możliwości** ich obejścia dzięki istnieniu dodatkowych źródeł informacji:

- społeczny aspekt publikowania w WWW (linki)
- tekst odnośników (ang. anchor text)

To są mocne narzędzia:

- ominięcie problemu braku samo-opisu
- dokumenty nietekstowe
- dokumenty o nieznanych formatach
- dokumenty nieściągnięte

Dodatkowo: nazwa hosta, domeny, pliku, głębokość ścieżki, ilosc dokumentów na hoście, ...

Linki są użyteczną informacją

Skupmy się na wykorzystaniu analizy linków grafu WWW do automatycznego obliczania rankingu dokumentów WWW

Struktura linków w grafie WWW może zostać wykorzystana do automatycznego obliczania "ważności" (lub jakości) dokumentów, **niezależnie** od kontekstu zapytania.

Taki składnik rankingu (niezależny od zapytania) nazywamy **statycznym**

Ważną cechą linkowego składnika rankingu danego dokumentu jest to, że pochodzi **spoza** tego dokumentu.

Społeczny aspekt hiperlinków

Podstawowa obserwacja:

Zamieszczenie linku z dokumentu p do dokumentu q może być odebrane jako informacja, że podmiot tworzący dokument p uważa dokument q za **wartościowy** (skoro wybrał go do wskazania spośród miliardów innych)

W ten sposób sami twórcy dokumentów WWW są w ukryty sposób "zaprzęgnięci" do oceny dokumentów WWW.

Pojedynczy link nie jest może bardzo wartościową informacją, ale mechanizm ten zastosowany w skali masowej zaczyna działać...

"Nepotyzm"

Problem stanowi tzw. "nepotyzm" linków, czyli tworzenie linków wskazujących dokumenty będące pod kontrolą tego samego podmiotu, który tworzy link. Nie każdy nepotyczny link jest tworzony w złej woli, ale oczywiście takie linki powinny być inaczej (słabiej) uwzględniane

Główny problem polega na niemożliwości pewnego ustalenia czy link tworzony jest przez ten sam podmiot, który kontroluje wskazywany dokument. WWW nie zawiera mechanizmu pozwalającego to sprawdzić.

Reakcja na "nepotyzm"

Typową heurystyką jest traktowanie całego hosta (lub poddomeny) jako przestrzeni kontrolowanej przez pojedynczy podmiot (autora)

W praktyce stosuje się kilka metod uwzględniania "nepotyzmu" opartego na hostach, np:

- ważenie linków w ten sposób, że z każdym hostem związana jest ograniczona wielkość, która jest rozdzielana (np. po równo) pomiędzy wszystkie wychodzące z niego linki
- **ignorowanie** linków wewnątrz hosta (lub poddomeny) przy obliczaniu rankingu opartego na analizie linków

Geneza HITS

- Algorytm HITS (Hyperlink-induced Topic Selection) został wymyślony przez J.Kleinberga w 1998 roku
- Algorytm ma wspomagać automatyczną identyfikację wartościowych dokumentów na dany temat (w kontekście zapytania)
- Równieśnik PageRank
- Algorytm został oryginalnie przedstawiony w pracy:

  J. Kleinberg. Authoritative sources in a hyperlinked environment. In Proc. 9th Ann. ACM-SIAM Symp. Discrete Algorithms, pages 668-677, ACM Press, New York, 1998.

Idea - autorytety i koncentratory

Algorytm pracuje na specjalnie przygotowanym *grafie bazowym*, który jest podgrafem grafu WWW bogatym w potencjalnie interesujące dokumenty na dany temat.
Koncept autorytetu (ang. authority) i koncentratora (ang. hub) - wzajemnie dualnych pojęć. Pojęcia te są określone wzajemnie rekurencyjnie:

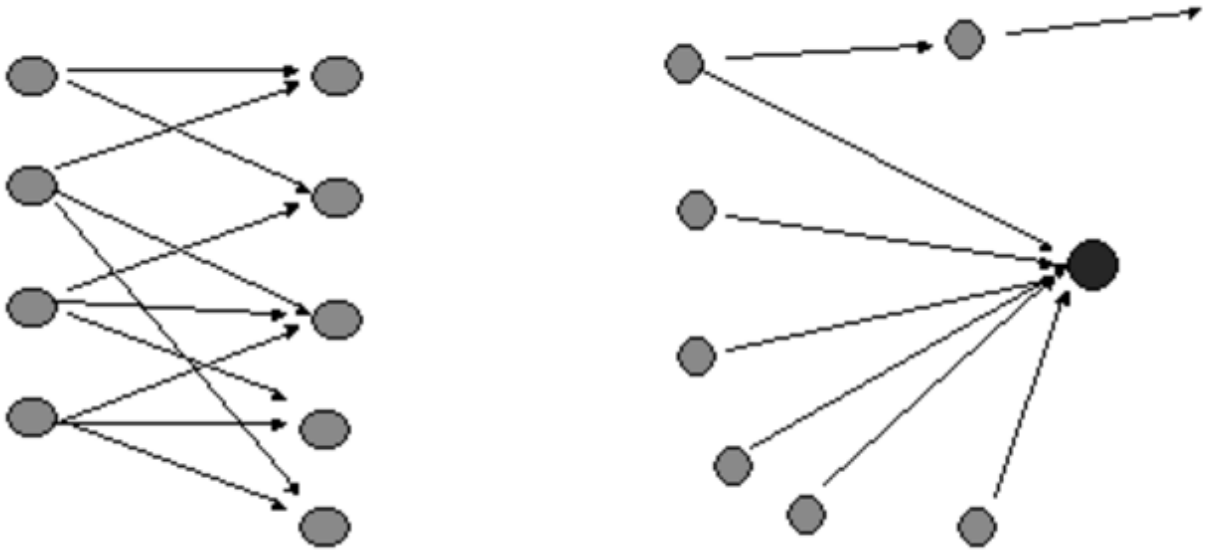
Definition

Dobry autorytet to taki dokument, który jest cytowany przez wiele dobrych koncentratorów. Analogicznie: dobry koncentrator to taki dokument, który zawiera linki do wielu dobrych autorytetów

W efekcie działania algorytmu każdemu dokumentowi przyporządkowane zostaną 2 wagi $x, y \in [0, 1]$, które określają jak dobrym jest autorytetem i koncentratorem, odpowiednio.

Wyjaśnienie koncepcji

Koncentratory są pojęciem pomocniczym wprowadzonym po to aby: odróżnić strony autorytatywne od po prostu *popularnych*



Rysunek: Różnica pomiędzy autorytetami na jakiś temat (ko-cytowanymi przez podobne dokumenty) a stronami popularnymi (często cytowanymi przez niezwiązane ze sobą dokumenty)

Obliczanie, Faza 1.1 - przygotowanie zbioru pierwotnego

Dane jest zapytanie q
Najpierw przygotowujemy dla q *zbiór bazowy* B_q (ang. base set)
W oryg. pracy miał on spełniać 3 warunki:

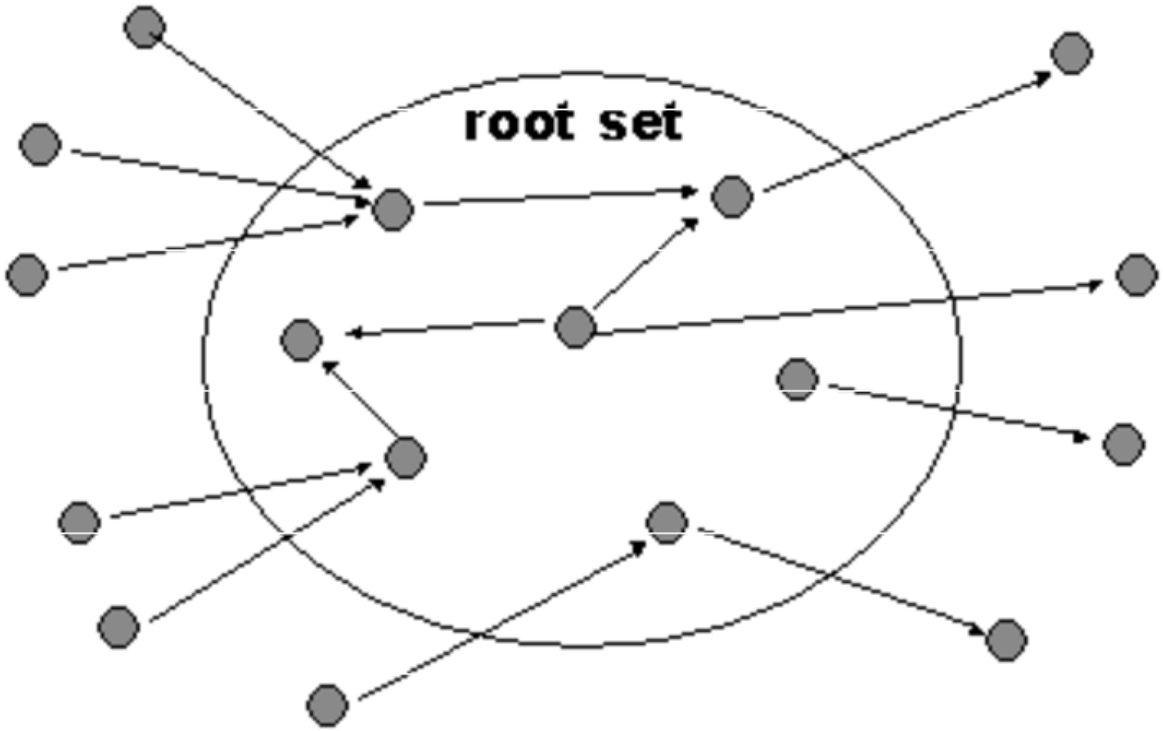
1. bogaty w dokumenty związane z q,
2. zawierający dużo autorytetów,
3. stosunkowo niewielki

Wg. Kleinberga wykorzystujemy do tego celu wyszukiwarkę internetową i pobieramy k najlepszych (wg. rankingu) dokumentów zwróconych w odpowiedzi na zapytanie q, gdzie k jest parametrem. Tak powstaje pomocniczy *zbiór pierwotny* (ang. root set) R_q, który spełnia 1 warunek.

Obliczanie, Faza 1.2 - przygotowanie zbioru bazowego

Następnie, aby nie pominąć żadnych dobrych autorytetów i koncentratorów, dołączamy do zbioru pierwotnego dokumenty wskazujące i wskazywane przez zbiór pierwotny
Dokładniej, dla każdego $d \in R_q$ dodajemy do R_q co najwyżej t dokumentów wskazujących i wskazywanych przez d (gdzie t jest parametrem - wg. Kleinberga np. 50). (tutaj można by nieuwzględniać tzw. *nepotycznych* linków - w obrębie tego samego hosta, itp.)
Zastosowanie ograniczenia t wynika z warunku 3 i natury grafu WWW (istnieją np. dobre strony o setkach tysięcy linków wchodzących - szczególnie wśród najlepszych na dany temat).
Wynikowy zbiór nazywamy zbiorem bazowym B_q. Powinien on spełniać warunki 1-3.

Konstruowanie zbioru bazowego



Rysunek: Konstruowanie zbioru bazowego z pierwotnego

Widoczne wady tego podejścia

Takie sformułowanie zbioru wejściowego algorytmu HITS sprawia, że ma on następujące wady:

- zależy od zewnętrznej wyszukiwarki, więc średnio nadaje się jako algorytm rankingowy (przynajmniej w oryginalnym sformułowaniu)
- wymaga wiedzy jakie dokumenty wskazują na zbiór pierwotny. Jest to trudne do zrealizowania w praktyce jeśli dysponujemy tylko zbiorem pierwotnym (connectivity server?)

Obliczanie wag (faza 2) - opis koncepcji

Mając obliczony zbiór bazowy iteracyjnie obliczamy wagi x(p) i y(p) dla każdej strony p.

1. Inicjalizujemy wszystkie wagi x i y wartością 1
2. Wykonujemy na przemian dwie operacje I oraz O
3. Operacja I (input): uaktualniamy autorytatywność każdej strony q sumując miarę bycia dobrym koncentratorem po wszystkich stronach cytujących q
4. Operacja O (output): uaktualniamy dla strony p miarę bycia dobrym koncentratorem sumując autorytatywność wszystkich stron wskazywanych przez p
5. Po każdej parze iteracji wagi normalizujemy
6. jeśli wagi zbiegły (z pożądaną dokładnością): stop
   else: goto 2

Obliczanie wag (faza 2) - wzory

- Inicjalizujemy wagi wartością 1
- Operacja I (od ang. input) uaktualnia wagi x odpowiadające konceptowi autorytetu:

$$x_q := \sum_{p|(p,q)\in E} y_p \tag{1}$$

- Analogicznie, operacja O (ang. output) uaktualnia wagi odpowiadające pojęciu koncentratora:

$$y_p := \sum_{q|(p,q)\in E} x_q \tag{2}$$

- Po każdej parze I oraz O występuje normalizacja wag tak, aby:

$$\sum_{p\in V} x_p^2 = \sum_{p\in V} y_p^2 = 1 \tag{3}$$

Zbieżność

Niech A oznacza macierz sąsiedztwa grafu G(V,E) odpowiadającego zbiorowi bazowemu B_q
W języku macierzowym operacje I oraz O wyrażają się bardzo prosto:

$$I: \ x := A^T y \tag{4}$$

$$O: \ y := Ax \tag{5}$$

W ten sposób wektor x po k parach iteracji wyraża się wzorem:

$$x^{(k)} = (A^T A)^{k-1} A^T z, \tag{6}$$

gdzie z to wektor początkowy. Analogicznie, wektor y po k parach iteracji jest opisany przez:

$$y^{(k)} = (AA^T)^k z \tag{7}$$

Macierze $A^T A$ i AA^T

$$x^{(k)} = (A^T A)^{k-1} A^T z, \quad y^{(k)} = (AA^T)^k z \tag{8}$$

- Macierze $A^T A$ i AA^T nazywamy macierzami ko-referencji i ko-cytowania, odpowiednio. (ang. co-reference, co-citation)
- Te pojęcia istnieją od dawna w *analizie bibliograficznej* — dziedzinie wiedzy, która rozwijała się w latach 60-tych 20. wieku.
- Zauważmy, że obliczanie wektorów x i y to *metoda potęgowa*.
- W tym przypadku obie macierze są kwadratowe i symetryczne. Dzięki tym własnościom, metoda potęgowa zbiega do głównych wektorów własnych macierzy ko-referencji i ko-cytowania [Golub and Van Loan "Matrix Computations"].

Wady HITS

Wady HITS

- związane z przygotowaniem danych (wymienione wcześniej)
- dodatkowo: wysoka podatność na manipulacje (spam)
- w HITS wynik zdominowany jest przez główną wartość własną. Odpowiada to dominującemu grafowi dwudzielnemu ("dominating bibartite community"). Pozostałe są ignorowane.

Wartość HITS

Z powyższych względów HITS mniej nadaje się jako algorytm rankingowy w wyszukiwarkach internetowych.
Mimo to można stosować go np. w kontrolowanych kolekcjach (np. intranety).
Wartość HITS:

- Jest to ważny, z punktu widzenia rozwoju analizy linków, algorytm, który równolegle z PageRank zapoczątkował rozwój tego typu technik.
- HITS i PageRank posłużyły i służą za podstawę wielu innym nowym algorytmom rankingowym (np. “Salsa”, czy “Unified Framework”).

Przykładowe rozszerzenia - PHITS

PHITS (Probabilistic HITS) Ulepszenie HITS (wada 3). Wprowadza ukrytą zmienną, która modeluje "temat" dokumentu.
Niweluje poważny problem dominacji wyniku przez główną wartość własną.
Cohn, D. and H.Chang, "Learning to Probabilistically Identify Authoritative Documents", Proceedings of the 17th International Conference on Machine Learning, 2000

Przykładowe rozszerzenia - Salsa

Próba połączenia modelu losowego internauty z koncepcją HITS.
W efekcie jest matematycznie równoważny zliczaniu stopni wejściowych (sic), co jest starannie udowodnione w pracy :)
Lempel, R. and S.Moran, "The Stochastic Approach for Link-Structure Analysis (SALSA) and the TKC Effect", in Proceedings of the 9th International WWW Conference, 2000

"Unified Framework"

Ciekawe uogólnienie i zarazem połączenie PageRank i HITS w jeden ogólny, parametryzowalny schemat.
PageRank i HITS stanowią dwa przeciwległe bieguny w tym schemacie.
Analizuje się też kilka pośrednich algorytmów.
Ding, C. and X.He and P.Husbands and H.Zha and H.Simon, "PageRank, Hits and a Unified Framework for Link Analysis", Lawrence Berkeley National Laboratory Technical Report 49372, 2001

Więcej odnośników literaturowych...

- S. Chakrabarti, B.E. Dom, S.R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg,"Mining the web's link structure", Computer, 32(8), pp. 60-67, 1999
- Brian Amento, Loren Terveen, Will Hill, "Does Authority Mean Quality? Predicting Expert Quality Ratings of Web Documents", Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2000
- A. Borodin, G.O. Roberts, J.S. Rosenthal, and P. Tsaparas, "Finding authorities and hubs from link structures on the world wide web", In Tenth International World Wide Web Conference, 2001
- R. Lempel and A. Soffer, "Picashow: Pictorial authority search by hyperlinks on the Web", Acm Transactions On Information Systems, 20(1), pp.1-24, 2002

Automatyczne znajdowanie stron podobnych

Koncepcja zbliżona do HITS, ale stosuje się wagi w celu m.in. zmniejszenia "nepotyzmu" (wagi dla każdego hosta lub dokumentu sumują się do 1 - podobnie jak w PageRank).

- Bharat, K. and M.Henzinger, "Improved Algorithms for Topic Distillation in Hyperlinked Environments", Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), pp. 104-111, 1998
- Dean, J. and M.Henzinger, "Finding Related Pages in the World Wide Web", Proceedings of the 8th International WWW Conference, 1999

Liczenie reputacji uczestników w aukcjach on-line

Stosunkowo niedawno zauważono, że w aukcjach internetowych (np. eBay, Allegro) kupujący i sprzedający w naturalny sposób są kandydatami do zastosowania na nich HITS i jego wariantów (jako potencjalne koncentratory i autorytety, odpowiednio).
Ma to bardzo ważne zastosowania w automatycznym obliczaniu tzw. *reputacji* kupujących i sprzedających na aukcjach internetowych. Jest to stosunkowo nowa dziedzina zastosowań dla pochodnych HITS.

Na zaliczenie tego wykładu:

- Podstawy racjonalne analizy linków w liczeniu rankingu
- Idea algorytmu HITS
- Sformułowanie HITS
- Analiza
- Rozszerzenia
- Znajdowanie Dokumentów Podobnych
- Zastosowania HITS w Systemach Reputacyjnych

Dziękuję za uwagę