Wyszukiwanie i Przetwarzanie Informacji WWW

Wprowadzenie

Marcin Sydow

Web Mining Lab, PJWSTK

Prowadzący

dr Marcin Sydow

Międzykatedralne Laboratorium Web Mining
oraz
Katedra Systemów Inteligentnych
PJWSTK
pokój: 311
e-mail: msyd@poljap.edu.pl
tel.: +48 22 58 44 571

Organizacja Kursu

- 15 spotkań (wykłady bez ćwiczeń)
- kolokwium ze znajomości wykładów
- sprawdzana obecność na zajęciach

Zaliczenie - system punktowy (razem max. 55 p.):

- pisemny **sprawdzian** (max. 30)
- około 10 kartkówek na pocz. zajęć (10 x 2 = 20)
- obecność/aktywność (ok. 5)
- (opcjonalnie - dla bardzo chętnych) projekt (?)

Ocena wynikowa dana jest wzorem: $\lfloor \frac{score}{10} \rfloor$
(wersja dla purystów: $min(5, max(2, \lfloor \frac{score}{10} \rfloor)))$

Wymagania

Na pozytywne zaliczenie wymagana jest:

1. całość materiału wykładów: **ogólna orientacja**
2. **wybrane** 1-3 wykłady: **dobra** znajomość

Wykłady będą na bardzo różne tematy i o zróżnicowanym charakterze:

- poglądowe (większość)
- techniczno-inżynierskie
- techniczno-algorytmiczne

Nie ma obowiązku zgłębiania wszystkich szczegółów - pozostawiony jest wybór

Założenia

Przydatna znajomość następujących zagadnień:

- względne obycie z WWW
- umiejętność korzystania z wyszukiwarek
- rozumienie podstaw html, http (TIN)
- elementarna wiedza z zakresu informatyki

Możliwie mały nacisk na szczegóły techniczne i matematykę

Jakich dziedzin dotyczy ten kurs?

1. wyszukiwanie informacji w korpusach dokumentów tekstowych (ang. Information Retrieval, IR)
2. wyszukiwarki internetowe (ang. search engines, rownież: WIR od ang. Web Information Retrieval)
3. eksploracja danych w sieci WWW (ang. Web Mining WM)
4. wybrane zagadnienia ekonomiczne i społeczne dotyczące WWW

Co celowo pominięto

Niektóre zagadnienia zaliczają się do tematyki Web Mining ale pominięto je ze względu na ograniczenia czasowe i fakt, że wymagają odrębnego kursu (lub taki kurs już istnieje)

Należą do nich m.in.

- Przetwarzanie Języka Naturalnego (ang. NLP)
- Uczenie Maszynowe i Analiza Danych

Czego kurs **nie** dotyczy bezpośrednio?

- tzw. technologii internetowych (html, PHP, JavaScript, Flash, CGI, CMS, Web Services, ...)
- budowy portali internetowych
- programowania (w tym sieciowego) i IO
- protokołów (HTTP, TCP/IP)
- zagadnień związanych z Internetem (DNS, etc.)
- technologii XML, RDF, XPath, ...
- mechanizmów działania sieci P2P
- pozycjonowania stron

(choć większość powyższych zagadnień ma duży związek z niniejszym kursem)

Plan Kursu

- Wprowadzenie
- Podstawy wyszukiwania informacji (ang. IR) (indeks, zapytania, interfejs)
- Globalne własności WWW i specyfika wyszukiwania w WWW (ang. WIR)
- Wyszukiwarki internetowe dużej skali (z "lotu ptaka")
- Systemy zbierania dokumentów WWW (ang. crawler)
- Repozytoria
- Przykłady konkretnych rozwiązań architektury wielkich wyszukiwarek
- Analiza struktury grafu hyperlinków WWW
- Algorytm PageRank, jego właściwości i warianty
- HITS, inne algorytmy i zastosowania w sieciach społecznych
- Ekonomiczne podstawy wyszukiwarek: reklamy
- Wybrane społeczne aspekty wyszukiwarek: zjawisko spamu

Wyszukiwanie Informacji w ujęciu klasycznym (ang. Information Retrieval)

- **wiedza** - reprezentowana przez: **korpus** dokumentów
- **potrzeba informacyjna** - reprezentowana przez: **zapytanie**

system ma **zwrócić** dokumenty, które odpowiadają potrzebie informacyjnej
Jest bardzo wiele wariantów tego systemu.

Dotyczy środowisk o **słabej, zaszumionej lub niejednorodnej strukturze**, takich jak WWW

Wyszukiwanie w bazach danych (gdzie jest dobrze zdefiniowana struktura) **nie zalicza się** do tego rodzaju.

Rola Wyszukiwarek

Zadanie wyszukiwania w WWW spełniają dzisiaj głównie wyszukiwarki internetowe - należące do **najczęściej używanych narzędzi** przez ludzi (81% gobalnej populacji Internetu użyło przynajmniej raz wyszukiwarki w grudniu 2006 w Wielkiej Brytanii, wg. Nielsen/NetRatings)

Wyszukiwarki WWW wywodzą się z "klasycznych" systemów IR (rozwijanych od lat 60 XX. wieku) pracujących na kontrolowanych kolekcjach dokumentów tekstowych w korporacjach, etc.

Kurs m.in. **wyjaśnia podstawowe zasady działania** zarówno klasycznych systemów jak i nowoczesnych wyszukiwarek WWW.

Oprócz zagadnień technicznych wspominane są ważne aspekty socjologiczno-ekonomiczne wyszukiwania w WWW.

Eksploracja Danych WWW (ang. Web Mining)

Skrzyżowanie starszej dziedziny: Eksploracji Danych (Data Mining) i zagadnień specyficznych dla sieci WWW.

Dotyczy wyszukiwania wzorców i automatycznego odkrywania użytecznej wiedzy z sieci WWW poprzez zastosowanie technik typowych dla "klasycznej" analizy danych wzbogaconych o **techniki specyficzne** dla WWW.

Czyli w wielkim skrócie:

$$WebMining = DataMining + WWW \tag{1}$$

Web Mining

Cechy WWW:

- ogromne bogactwo danych zawartych w WWW
- wyjątkowa dynamika (ciągły wykładniczy wzrost)
- wysoka różnorodność i "zaszumienie"
- uczestnictwo setek milionów wzajemnie powiązanych procesów (sterowanych zarówno przez ludzi jak i maszyny)
- ogromne (i wciąż rosnące) zaangażowanie ekonomiczne, polityczne i społeczne milionów "agentów" (o często sprzecznych interesach)

1. Web należy do najciekawszych obecnie pól zastosowań Data Mining
2. Web Mining ciągle stawia niezwykłe wyzwania koncepcyjne i technologiczne, z których wiele wciąż czeka na rozwiązanie

Web Mining

W Web Mining - tradycyjny podział na 3 główne działy:

1. Eksploracja Zawartości WWW (ang. Content Mining)
   (dawniejszy text mining + eksploracja struktury + NLP + ...)
2. Eksploracja Struktury WWW (ang. Link Analysis)
   (grafy, grafy losowe, algebra, procesy stochastyczne, kombinatoryka, ...)
3. Analiza Użytkowników WWW (ang. Web Usage Mining)
   (eksploracja danych, analiza logów, analiza danych temporalnych, modelowanie użytkowników, ...)

Można uznać, że WIR (Web Information Retrieval, czyli Wyszukiwanie Informacji w WWW) jest również poddziedziną Web Mining

Przykłady

- Ekstrakcja Informacji na zadany temat z WWW
- Automatyczne porównywanie cen wybranych produktów
- Identyfikacja Grup Użytkowników o określonych zainteresowaniach lub aktywności
- Systemy demaskowania plagiatów (np. plagiat.pl)
- Automatyczne generowanie wiedzy z zasobów WWW
- Odnajdywanie osób
- Automatyczne śledzenie opinii publicznej na dany temat
- Wyszukiwarka multimediów (filmy, muzyka, etc.)
- Wykrywanie i Zwalczanie Chłamu Wyszukiwarkowego (ang. Spam)
- Wykrywanie nadużyć i przestępstw (finanse, terroryzm, etc.)
- Identyfikacja grup klientów
- Optymalizacja przestrzeni reklamowej

Dostęp do informacji WWW

Obecne "paradygmaty" organizacji dostępu do informacji w WWW:

1. nawigacja "ręczna" po dokumentach (pierwotny, obecnie w zaniku)
2. katalogi tematyczne dokumentów (w defensywie?)
3. wyszukiwarki "boolowskie" (obecnie dominuje)

Wyszukiwarki **zmieniły** proces rozwoju WWW.

Dostęp do informacji WWW

Obecne "paradygmaty" organizacji dostępu do informacji w WWW:

1. nawigacja "ręczna" po dokumentach (pierwotny, obecnie w zaniku)
2. katalogi tematyczne dokumentów (w defensywie?)
3. wyszukiwarki "boolowskie" (obecnie dominuje)

Wyszukiwarki **zmieniły** proces rozwoju WWW.

Co dalej?

- QA ("odpowiadarki" na pytania)
- nawigacja "inteligentna" (semantyczna)
- ...

(Pre)historia WIR w skrócie

- 1611: prototyp indeksu (Strong's Exhaustive Concordance of Bible)
- 1945: Memex - "prototyp" WWW (V.Bush "As we may think")
- 1960: SMART Information Retrieval System (G.Salton, Cornell Univ.)
- 1965: Xanadu - *hypertext* (Ted Nelson)
- 1980: system do nawigacji po dokumentach (T.Berners-Lee)
- 1990: narodziny WWW (Tim Berners-Lee, CERN)
- 1993-95: pierwsze przeglądarki (Mosaic/Netscape)
- 1994: Lycos - pierwsza wyszukiwarka
- 1994: WebCrawler, 4K hostów (Brian Pinkerton)
- 1994: "Jerry's Guide to the World Wide Web" (później: Yahoo)
- 1995: AltaVista, Excite, InfoSeek, Inktomi
- 1996: Yahoo wchodzi na giełdę
- 1996-1998: początki Google

Co wypada wiedzieć po tym wykładzie:

1. Jakie są reguły zaliczenia :)
2. Co to jest Web Information Retrieval
3. Czym zajmuje się Web Mining
4. Działy Web Mining (3-4)
5. Przykłady zastosowań (ze 3)
6. Orientacyjne liczby dotyczące WWW
7. Rola wyszukiwarek
8. Podstawowa wiedza historyczna (co? kiedy?)

Dziękuję za uwagę