Specjalizacja Web Mining

dr Marcin Sydow

Laboratorium Web Mining

Katedra Systemów Inteligentnych
PJWSTK

15.11.07

Co to jest Web Mining?

Jest to połączenie 2 dziedzin:

1. Web Information Retrieval
2. Data Mining

Co to jest Web Mining?

Jest to połączenie 2 dziedzin:

1. Web Information Retrieval
2. Data Mining

Czyli, w skrócie:

Zastosowanie metod analizy danych i sztucznej inteligencji do przetwarzania i wyszukiwania informacji WWW.

Wyszukiwarki Internetowe

Dzisiaj jest ok. 24 000 000 000 (24 miliardy!) dokumentów WWW.
Wyszukiwarki zapewniają do nich dostęp - używamy ich codziennie.

5 najpopularniejszych stron WWW na świecie?

Wyszukiwarki Internetowe

Dzisiaj jest ok. 24 000 000 000 (24 miliardy!) dokumentów WWW.
Wyszukiwarki zapewniają do nich dostęp - używamy ich codziennie.

5 najpopularniejszych stron WWW na świecie?
(Yahoo, Google, YouTube, Live, MSN) - prawie same wyszukiwarki

Wyszukiwarki Internetowe

Dzisiaj jest ok. 24 000 000 000 (24 miliardy!) dokumentów WWW.
Wyszukiwarki zapewniają do nich dostęp - używamy ich codziennie.

5 najpopularniejszych stron WWW na świecie?
(Yahoo, Google, YouTube, Live, MSN) - prawie same wyszukiwarki

Wyszukiwarki obsługują ponad 500 milionów zapytań dziennie.
Potrzeba do tego dziesiątek tysięcy maszyn i specjalnych algorytmów

Wyszukiwarki Internetowe

Dzisiaj jest ok. 24 000 000 000 (24 miliardy!) dokumentów WWW.
Wyszukiwarki zapewniają do nich dostęp - używamy ich codziennie.

5 najpopularniejszych stron WWW na świecie?
(Yahoo, Google, YouTube, Live, MSN) - prawie same wyszukiwarki

Wyszukiwarki obsługują ponad 500 milionów zapytań dziennie.
Potrzeba do tego dziesiątek tysięcy maszyn i specjalnych algorytmów

Czy wiesz co jest najtrudniejszym problemem w wyszukiwaniu?

Wyszukiwarki Internetowe

Dzisiaj jest ok. 24 000 000 000 (24 miliardy!) dokumentów WWW.
Wyszukiwarki zapewniają do nich dostęp - używamy ich codziennie.

5 najpopularniejszych stron WWW na świecie?
(Yahoo, Google, YouTube, Live, MSN) - prawie same wyszukiwarki

Wyszukiwarki obsługują ponad 500 milionów zapytań dziennie.
Potrzeba do tego dziesiątek tysięcy maszyn i specjalnych algorytmów

Czy wiesz co jest najtrudniejszym problemem w wyszukiwaniu?

Czy wiesz ile wyniósł roczny zysk z reklam wyszukiwarkowych w 2006 w USA?

Wyszukiwarki Internetowe

Dzisiaj jest ok. 24 000 000 000 (24 miliardy!) dokumentów WWW.
Wyszukiwarki zapewniają do nich dostęp - używamy ich codziennie.

5 najpopularniejszych stron WWW na świecie?
(Yahoo, Google, YouTube, Live, MSN) - prawie same wyszukiwarki

Wyszukiwarki obsługują ponad 500 milionów zapytań dziennie.
Potrzeba do tego dziesiątek tysięcy maszyn i specjalnych algorytmów

Czy wiesz co jest najtrudniejszym problemem w wyszukiwaniu?

Czy wiesz ile wyniósł roczny zysk z reklam wyszukiwarkowych w 2006 w USA? **6.75 miliarda dolarów (!!)**

Przykładowe projekty

Jeśli interesuje Cię:

- Wykrywanie i zwalczanie spamu wyszukiwarkowego
- Przewidywanie zachowania użytkowników WWW w oparciu o rzeczywiste dane o ruchu internetowym
- Automatyczne zbieranie i analiza dziesiątek milionów dokumentów WWW
- Budowa klastra zwykłych PC do potężnych obliczeń rozproszonych w modelu Map/Reduce

W laboratorium Web Mining właśnie aktualnie to robimy.

Przydatne cechy kandydata

- Zainteresowanie wyszukiwarkami i WWW “od kuchni”
- Zainteresowanie sztuczną inteligencją i analizą danych
- Programowanie (Java, C++, języki skryptowe)
- Zainteresowanie Algorytmiką i Matematyką
- Znajomość środowiska GNU/Linux i narzędzi open source
- Gotowość poznania nowych narzędzi (np. Weka, R, Latex)
- Pasja badawcza

Laboratorium Web Mining PJWSTK

Zespół Laboratorium Web Mining utrzymuje kontakty z czołowymi ośrodkami zagranicznymi i krajowymi, naukowymi i biznesowymi: obecnie są to:

- Yahoo! Research Barcelona
- Joint Research Center of European Union, Ispra, Włochy
- IPI PAN, Warszawa
- Politechnika Poznańska, Wydział Informatyki
- Netsprint Sp. z o.o.
- Gemius S.A.

Przykładowa tematyka prac dyplomowych

- automatyczne zbieranie i analiza milionów dokumentów WWW
- budowa komponentów wyszukiwarki eksperymentalnej
- uczenie maszynowe i sztuczna inteligencja w Web Mining
- nowe algorytmy pozyskiwania i porządkowania informacji
- budowa infrastruktury do rozproszonych obliczeń Web Mining
- dopasowanie reklam internetowych do stron WWW/zapytań

Współpraca dydaktyczno-badawcza z Katedrami PJWSTK

- Systemów Inteligentnych
- Metod Programowania
- Algorytmiki
- Sieci Komputerowych
- Matematyki i Statystycznej Analizy Danych
- Systemów Wieloagentowych i Robotyki

Bardzo interdyscyplinarny charakter.

Gdzie pracują eksperci od Web Mining

- wyszukiwarki internetowe
- duże portale internetowe
- agendy rządowe i Unii Europejskiej
- firmy konsultingowe i instytuty badań rynku/opinii
- instytucje związane z bezpieczeństwem
- własna działalność

Kontakt

Chętnie udzielamy wszelkich dodatkowych informacji.

`http://www.pjwstk.edu.pl/~msyd/webmining.html`

Osoba kontaktowa:

dr Marcin Sydow

konsultacje: poniedziałek 15:00 - 16:30, p. 311

msyd@pjwstk.edu.pl, telefon: +48 22 58 44 571