Wprowadzenie do Wyszukiwania Informacji WWW

Preludium

Marcin Sydow

Web Mining Lab
PJWSTK

Co to jest wyszukiwanie informacji?

Meno zapytał Sokratesa:

“Jak zapytać o to czego nie znamy?”

(Platon: Meno, 400 p.n.e.)

A jeśli dostaniemy odpowiedź: skąd niby wiadomo, że jest prawdziwa?

Wyszukiwanie Informacji

Podstawowa aktywność człowieka dzisiaj.

Ostatnia dekada: rewolucja w podejściu człowieka do zdobywania informacji na skutek eksplozji technologii informacyjnych (w tym Internet, WWW i wyszukiwarki).

Niebezpieczne uproszczenia naszych czasów

Niektórzy ludzie są (błędnie) przekonani, że sieć zawiera obecnie całą istotną ludzką wiedzę. Oto inne naiwne a popularne uproszczenia:

- wszystko jest w sieci
- istnieje tylko to co jest w sieci
- warto szukać informacji tylko w sieci

Niebezpieczne uproszczenia naszych czasów

Niektórzy ludzie są (błędnie) przekonani, że sieć zawiera obecnie całą istotną ludzką wiedzę. Oto inne naiwne a popularne uproszczenia:

- wszystko jest w sieci
- istnieje tylko to co jest w sieci
- warto szukać informacji tylko w sieci

- to co zwróci wyszukiwarka jest prawdą
- tylko to istnieje, co zwróci wyszukiwarka
- istnieje tylko jedna dobra wyszukiwarka, etc.

Niebezpieczne uproszczenia naszych czasów

Niektórzy ludzie są (błędnie) przekonani, że sieć zawiera obecnie całą istotną ludzką wiedzę. Oto inne naiwne a popularne uproszczenia:

- wszystko jest w sieci
- istnieje tylko to co jest w sieci
- warto szukać informacji tylko w sieci

- to co zwróci wyszukiwarka jest prawdą
- tylko to istnieje, co zwróci wyszukiwarka
- istnieje tylko jedna dobra wyszukiwarka, etc.

To zadziwiające, codzienna aktywność większości społeczeństwa "zachodniego" jest totalnie uzależniona od usług zaledwie około **trzech** korporacji (!)(do tego, wszystkie mają centrale w tym samym regionie świata...)

Gdzie była i jest przechowywana wiedza ludzkości

- Tradycja ustna
  - homo erectus: 2 000 000 lat temu
  - homo sapiens sapiens: 250 000. lat temu
- Biblioteki
  - (papirus conajmniej około 4000 lat temu)
  - **Biblioteka Aleksandryjska** (ok. 300 p.n.e) **ok. 2300 lat temu**
- (ostatnio) WWW: zaledwie niecałe 19 lat temu...

Biblioteki

Rozwijane przez wieki. Słusznie uważane do niedawna za główną skarbnicę summy ludzkiej wiedzy.

- Historia:
  - biblioteka alexandryjska: najwieksza biblioteka świata starożytnego założył: Ptolemeusz I (ok. 350-283 p.n.e.), 200.000 woluminów spalona podczas inwazji Cezara na Egipt (ok. 48 p.n.e.)
- Polska:
  - najstarsza: biblioteka jagiellońska (kiedy założona?

Biblioteki

Rozwijane przez wieki. Słusznie uważane do niedawna za główną skarbnicę summy ludzkiej wiedzy.

- Historia:
  - biblioteka alexandryjska: najwieksza biblioteka świata starożytnego założył: Ptolemeusz I (ok. 350-283 p.n.e.), 200.000 woluminów spalona podczas inwazji Cezara na Egipt (ok. 48 p.n.e.)
- Polska:
  - najstarsza: biblioteka jagiellońska (kiedy założona? 1364 - zał. UJ):

Biblioteki

Rozwijane przez wieki. Słusznie uważane do niedawna za główną skarbnicę summy ludzkiej wiedzy.

- Historia:
  - biblioteka alexandryjska: najwieksza biblioteka świata starożytnego założył: Ptolemeusz I (ok. 350-283 p.n.e.), 200.000 woluminów spalona podczas inwazji Cezara na Egipt (ok. 48 p.n.e.)
- Polska:
  - najstarsza: biblioteka jagiellońska (kiedy założona? 1364 - zał. UJ): obecnie: 6.4 miliona jednostek
  - największa: biblioteka narodowa, 7.9 miliona jednostek
  - największa uczelniana: BUW 4.18M
- Świat: (obecnie największa) bibl. kongresu amerykańskiego (zał. 1800), ok 30M książek, 60M rękopisów i inne

WWW

Zaledwie 19 lat, ale...

w 2005 roku: 11.5 miliarda “sensownych” stron: ok 40TB tekstu

tymczasem ogromna Biblioteka Kongresu:

WWW

Zaledwie 19 lat, ale...

w 2005 roku: 11.5 miliarda "sensownych" stron: ok 40TB tekstu

tymczasem ogromna Biblioteka Kongresu: ok. 30TB tekstu (szacunkowo)

WWW

Zaledwie 19 lat, ale...

w 2005 roku: 11.5 miliarda "sensownych" stron: ok 40TB tekstu

tymczasem ogromna Biblioteka Kongresu: ok. 30TB tekstu (szacunkowo)

nasza biblioteka PJWSTK spokojnie zmieści się na iPodzie (jeśli ograniczymy do tekstu).

Digitalizacja naszego dziedzictwa kulturowego

- Projekt Gutenberg (1971, M.Hart, cel: digitalizacja 10.000 tekstów z "public domain" w ciągu 30 lat. Ukończony w 2003, następny cel: 1M. Początkowo wpisywane przez ochotników, potem skanowane i sprawdzane przez ochotników)
- Million Book Project (2001, Carnegie Mellon University (CMU). Wolny dostęp, pełny indeks tekstowy. Technologia: OCR, etc. Książki są pakowane i wysyłane do Indii i Chin, gdzie są tanio skanowane. Słaba jakość w porównaniu do Gutenberga)
- Internet Archive (1996 - cel: zapisywać większość WWW w kolejnych "migawkach"(obecnie około 150.000.000.000 stron ! w tym "wayback machine"). Obecnie przechowuje także "Million Book". 2002: partnerstwo z "Bibliotheca Alexandrina" (odbudowywana). Książki "niekomercyjne" i "osierocone" pod względem praw autorskich)
- Amazon (ograniczony dostęp do dużej ilości dzieł (ponad 100.000) chronionych prawem autorskim. System wyszukiwania i rekomendacji warunkowany komercyjnie.

Digitalizacja naszego dziedzictwa kulturowego, cd.

- Google Books (2004 - początkowo: współpraca z 5 bibliotekami uniwersyteckimi: Harvard, Michigan, NY, Oxford, Stanford. Częściowo otwarty dostęp. Skanowanie i przeszukiwanie dużej liczby książek (masowa technologia: ok. 1000 stron/h), w tym chronionych prawem autorskim. Ostatnio inicjatywa wzbudziła wielkie kontrowersje dotyczące rozszerzenia oferty o bezpośrednią sprzedaż. Grudzień 2009: Francja zatrzymała proces masowego skanowania francuskiej literatury (jako "pogwałcenie praw autorskich")

Digitalizacja naszego dziedzictwa kulturowego, cd.

- Google Books (2004 - początkowo: współpraca z 5 bibliotekami uniwersyteckimi: Harvard, Michigan, NY, Oxford, Stanford. Częściowo otwarty dostęp. Skanowanie i przeszukiwanie dużej liczby książek (masowa technologia: ok. 1000 stron/h), w tym chronionych prawem autorskim. Ostatnio inicjatywa wzbudziła wielkie kontrowersje dotyczące rozszerzenia oferty o bezpośrednią sprzedaż. Grudzień 2009: Francja zatrzymała proces masowego skanowania francuskiej literatury (jako "pogwałcenie praw autorskich")
- Open Content Alliance (2005 - w odpowiedzi na zamknięty i (w sumie) komercyjny projekt Google. Otwarty, kooperacja wielu uniwesytetów i firm (np. MSN, Yahoo!, HP, Adobe, Internet Archive, etc.) z postanowieniem pełnego respektowania praw autorskich.

Digitalizacja naszego dziedzictwa kulturowego, cd.

- Google Books (2004 - początkowo: współpraca z 5 bibliotekami uniwersyteckimi: Harvard, Michigan, NY, Oxford, Stanford. Częściowo otwarty dostęp. Skanowanie i przeszukiwanie dużej liczby książek (masowa technologia: ok. 1000 stron/h), w tym chronionych prawem autorskim. Ostatnio inicjatywa wzbudziła wielkie kontrowersje dotyczące rozszerzenia oferty o bezpośrednią sprzedaż. Grudzień 2009: Francja zatrzymała proces masowego skanowania francuskiej literatury (jako "pogwałcenie praw autorskich")
- Open Content Alliance (2005 - w odpowiedzi na zamknięty i (w sumie) komercyjny projekt Google. Otwarty, kooperacja wielu uniwesytetów i firm (np. MSN, Yahoo!, HP, Adobe, Internet Archive, etc.) z postanowieniem pełnego respektowania praw autorskich.
- Nowy model dostępu do książek (e-book. Książki w wersji elektronicznej. Z jednej strony szansa na czytanie niedostępnych fizycznie książek. Z drugiej liczne **potencjalne** zagrożenia: nowi właściciele mogą zastosować środki prawne do manipulowania dostępem do dzieł. Np.: blokada tylko do określonego urządzenia (koniec z pożyczaniem książek od przyjaciół!), niemożliwość odsprzedaży (koniec z rynkiem drugiego obiegu!), czasowy limit istnienia "zakupionej" książki (koniec z trwałą kolekcją – po upływie terminu, książki po prostu znikną!)

Pozytywne aspekty rewolucji informacyjnej

Duży temat – conajmniej na doktorat z filozofii nauki lub socjologii, etc.

Niewątpliwie rewolucja informacyjna przynosi wiele pozytywnych aspektów.

Zwykle to o nich się mówi.

przykłady?

Ciemne strony rewolucji informacyjnej

Zróbmy więc ćwiczenie i pokuśmy się o dostrzeżenie również negatywnych aspektów:

Ciemne strony rewolucji informacyjnej

Zróbmy więc ćwiczenie i pokuśmy się o dostrzeżenie również negatywnych aspektów:

np.

- potencjalny oligopol kilku monstrualnych korporacji kontrolujących prawie cały dostęp ludzkości do informacji i komunikacji

Ciemne strony rewolucji informacyjnej

Zróbmy więc ćwiczenie i pokuśmy się o dostrzeżenie również negatywnych aspektów:

np.

- potencjalny oligopol kilku monstrualnych korporacji kontrolujących prawie cały dostęp ludzkości do informacji i komunikacji
- teoretyczne możliwości manipulowania informacją na skalę masową

Ciemne strony rewolucji informacyjnej

Zróbmy więc ćwiczenie i pokuśmy się o dostrzeżenie również negatywnych aspektów:

np.

- potencjalny oligopol kilku monstrualnych korporacji kontrolujących prawie cały dostęp ludzkości do informacji i komunikacji
- teoretyczne możliwości manipulowania informacją na skalę masową
- niespotykane wcześniej możliwości "kontrolowania myśli" prawie całego społeczeństwa (a'la Orwell "1984") (logi wyszukiwania i kliknięcia zdradzają nasze aktualne myśli. Jest to w formie idealnej do automatycznej analizy na masową skalę)

Ciemne strony rewolucji informacyjnej

Zróbmy więc ćwiczenie i pokuśmy się o dostrzeżenie również negatywnych aspektów:

np.

- potencjalny oligopol kilku monstrualnych korporacji kontrolujących prawie cały dostęp ludzkości do informacji i komunikacji
- teoretyczne możliwości manipulowania informacją na skalę masową
- niespotykane wcześniej możliwości "kontrolowania myśli" prawie całego społeczeństwa (a'la Orwell "1984") (logi wyszukiwania i kliknięcia zdradzają nasze aktualne myśli. Jest to w formie idealnej do automatycznej analizy na masową skalę)
- (potencjalne) odcięcie ludzkości od trwałych nośników kultury (digitalizacja) i możliwość późniejszego kontrolowania dostępu (np. e-booki)

Ciemne strony rewolucji informacyjnej

Zróbmy więc ćwiczenie i pokuśmy się o dostrzeżenie również negatywnych aspektów:

np.

- potencjalny oligopol kilku monstrualnych korporacji kontrolujących prawie cały dostęp ludzkości do informacji i komunikacji
- teoretyczne możliwości manipulowania informacją na skalę masową
- niespotykane wcześniej możliwości "kontrolowania myśli" prawie całego społeczeństwa (a'la Orwell "1984") (logi wyszukiwania i kliknięcia zdradzają nasze aktualne myśli. Jest to w formie idealnej do automatycznej analizy na masową skalę)
- (potencjalne) odcięcie ludzkości od trwałych nośników kultury (digitalizacja) i możliwość późniejszego kontrolowania dostępu (np. e-booki)
- totalne uzależnienie ludzi od technologii i elektronicznych gadżetów (nasza pamięć staje się coraz słabsza...)

Co wypada wiedzieć po tym wykładzie:

1. Podstawowa wiedza historyczna (co? kiedy?)
2. Wymień i krótko opisz co najmniej 3 z największych ostatnio przedsięwzięć digitalizacji dziedzictwa kulturowego ludzkości
3. Wymień co najmniej 3 (najlepiej dostrzeż samodzielnie) zagrożenia związane z obecnymi rewolucyjnymi zmianami w organizacji informacji

Dziękuję za uwagę