Eksploracja danych w Internecie

- Wykłady opracowano w oparciu o książkę Ricardo Baeza-Yates, Berthier Ribeiro-Neto, „*Modern Information Retrieval, the concepts and technology behind search*” 2nd edition, ACM Press Books, 2011
- Z tego samego źródła zaczerpnięto także różne zadania i przykłady wykorzystywane w treści wykładu.

Eksploracja danych w Internecie

Literatura:

- Ricardo Baeza-Yates, Berthier Ribeiro-Neto, „Modern Information Retrieval, the concepts and technology behind search”, 2nd edition, ACM Press Books, 2011

- Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, „An Introduction to Information Retrieval”, Online edition (c) 2009 Cambridge UP

- Anand Rajaraman and Jeffrey D. Ullman „Mining of Massive Datasets”, http://infolab.stanford.edu/~ullman/mmds/book.pdf

- „Eksploracja zasobów internetowych. Analiza struktury, zawartości i użytkowania sieci WWW.”, Zdravko Markov, Daniel T. Larose, PWN, Warszawa 2009

Eksploracja danych w Internecie

- Wprowadzenie
- Interfejsy użytkownika
- Modele dokumentów
- Ocena jakości procesu wyszukiwania
- Klasyfikacja tekstów
- Indeksowanie przy wyszukiwaniu
- Wyszukiwanie równoległe
- Problemy wyszukiwania w sieci
- Zagadnienia web crawlingu
- Wyszukiwanie danych multimedialnych

Eksploracja danych w Internecie

- Wyszukiwanie informacji (*Information Retrieval*) dotyczy reprezentacji, zapamiętywania organizacji i dostępu do poszczególnych składników informacji
  - Elementy informacyjne to dokumenty, strony sieciowe, katalogi sieciowe, uporządkowane rekordy danych i obiekty multimedialne
- Pierwotne cele wyszukiwania informacji to: indeksowanie tekstów i poszukiwanie pożądanych dokumentów w ich zbiorach
- Obecnie wyszukiwanie informacji obejmuje zagadnienia :
  - modelowanie, przeszukiwanie sieci, klasyfikację tekstów, architektury systemów, interfejsy użytkownika, filtrowanie i wizualizację danych, tłumaczenie tekstów

Eksploracja danych w Internecie

- Przez ponad 5000 lat, człowiek porządkował informację dla późniejszego jej odzyskania po odpowiednim wyszukaniu
  - Odbywało się to przez zapisywanie, archiwizowanie i indeksowanie papirusów, glinianych tabliczek, węzełków-kipu, wampumów i oczywiście książek

- W celu przechowywania nośników danych powstały specjalne budowle zwane bibliotekami
  - Najstarsza znana biblioteka powstała na wyspie Elba, pomiędzy 3000 i 2500 rokiem p.n.e.
  - Ok 300 roku p.n.e. powstała słynna Biblioteka Aleksandryjska

Eksploracja danych w Internecie

- Ponieważ ilość informacji przechowywanej w bibliotekach ciągle rośnie buduje się *indeksy* – specjalizowane struktury danych do szybkiego jej wyszukiwania

- Dawniej indeksy tworzono manualnie jako zbiory kategorii, z dodatkowymi etykietami powiązanymi z każdą kategorią

- Pojawienie się nowoczesnych komputerów pozwoliło na automatyczne budowanie dużych indeksów

- W latach 50-tych badacze tacy jak Hans Peter Luhn, Eugene Garfield,Philip Bagley i Calvin Moores wypracowali pojęcie wyszukiwania informacji - *Information Retrieval* (IR)

Eksploracja danych w Internecie

- W 1963 roku Joseph Becker i Robert Hayes opublikowali pierwszą książkę na temat IR
- W latach 60-tych badania w tym temacie prowadzili m. in. Karen Sparck Jones i Gerard Salton; doprowadziły one do utworzenia definicji *TF-IDF term weighting scheme* (Term Frequency - Inverse Document Frequency) określającego wzór wagi przypisanej poszczególnym pozycjom (pojęciom) w dokumencie

- W 1971 r. Jardine i van Rijsbergen wprowadzili hipotezę klasteringu (*cluster hypothesis*)
- W 1979 r. van Rijsbergen opublikował klasyczną pozycję książkową *Information Retrieval* omawiającą model probabilistyczny

Eksploracja danych w Internecie

- W 1983 r. Salton i McGill opublikowali książkę *Introduction to Modern Information Retrieval* omawiającą model wektorowy pozyskiwania danych

- Biblioteki były pierwszymi instytucjami wykorzystującymi systemy IR dla pozyskiwania informacji w formie przeszukiwania kart katalogowych

- Następnie tę prostą funkcjonalność rozszerzono o:
  - analizę nagłówków, szukanie wg. słów kluczowych, specjalizowane operatory zapytań

- Obecnie rozwój systemów IR koncentruje się na doskonaleniu interfejsów graficznych, wspomaganiu sprzętowym i cechach hipertekstowych dokumentów

Eksploracja danych w Internecie

- Przed epoką Internetu, wydobywaniem informacji interesowali się najczęściej bibliotekarze i eksperci od przetwarzania danych

- Internet stał się obecnie największym archiwum wiedzy w historii ludzkości

- Znalezienie pożądanej informacji w Internecie, z powodu gigantycznych rozmiarów tego repozytorium wymaga użycia nowej technologii wyszukiwania informacji

Problem wyszukiwania informacji

- Pełny opis wymaganej informacji podany przez użytkownika nie zawsze jest trafnym zapytaniem do systemu IR

- Użytkownik sieci może także sformułować swoje wymagania w formie zapytania

- Najistotniejszy dla wyszukiwania jest zbiór słów kluczowych (*keywords*) lub terminów indeksujących (*index terms*).

- Celem systemów IR jest jak najtrafniejsze wydobycie informacji istotnej dla użytkownika na podstawie podanego zapytania

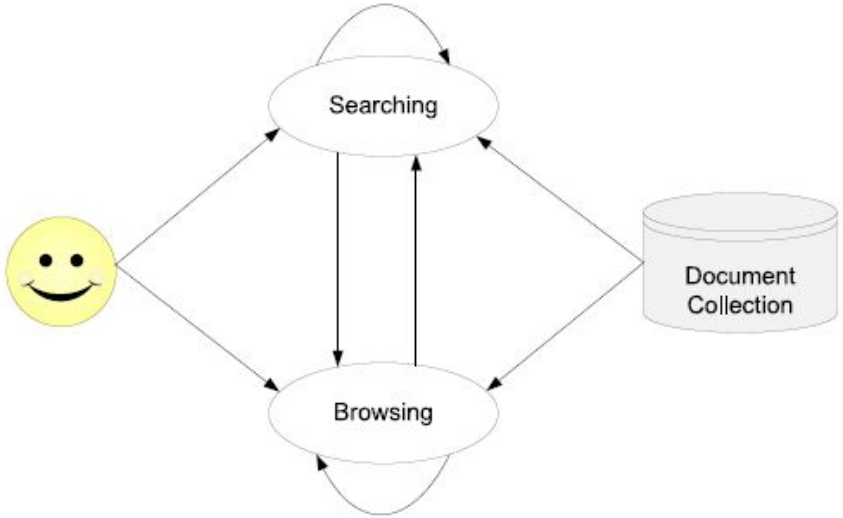
Problem wyszukiwania informacji

- System IR powinien uszeregować elementy informacji według ich istotności w zapytaniu użytkownika

- Celem systemu IR jest wydobycie wszystkich elementów istotnych dla zapytania użytkownika i jak najmniejszej ilości elementów nieistotnych

- Pojęcie istotności informacji w systemach IR jest kluczowe

Problem wyszukiwania informacji



- Jeśli użytkownik precyzuje konkretny temat, często w formie zapytania, to mówi się że poszukuje, wyławia (*sarching*) lub pyta o informację (*querying*)

- Jeżeli użytkownik formułuje wymagania szeroko lub nieprecyzyjnie to mówi się o żeglowaniu (*navigating*) lub przeglądaniu (*browsing*) dokumentów w Internecie

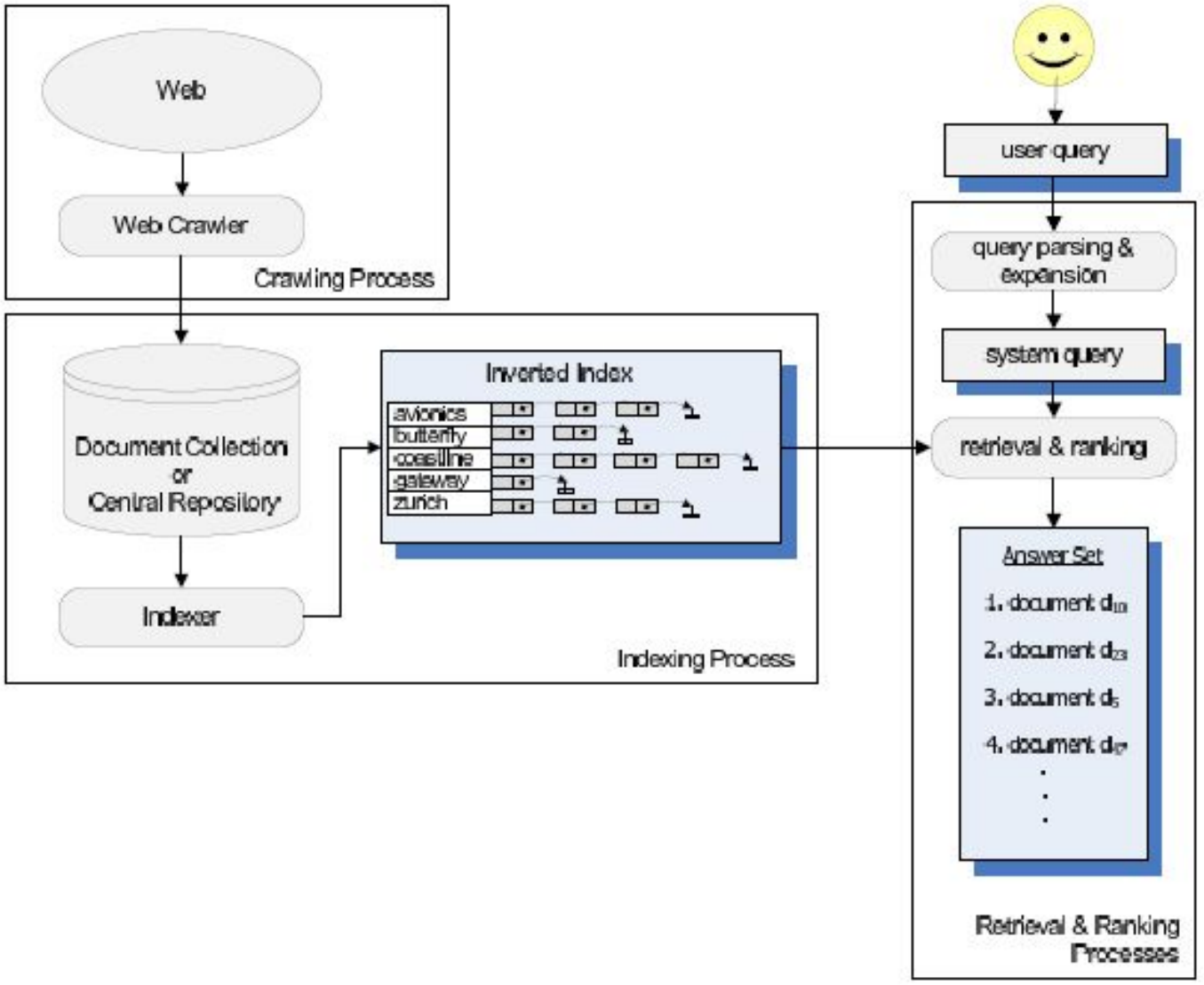
Wyszukiwanie danych i informacji

- Wyszukiwanie danych (*Data retrieval*): zadanie definiujące ściśle, które dokumenty ze zbioru zawierają słowa kluczowe z zapytania

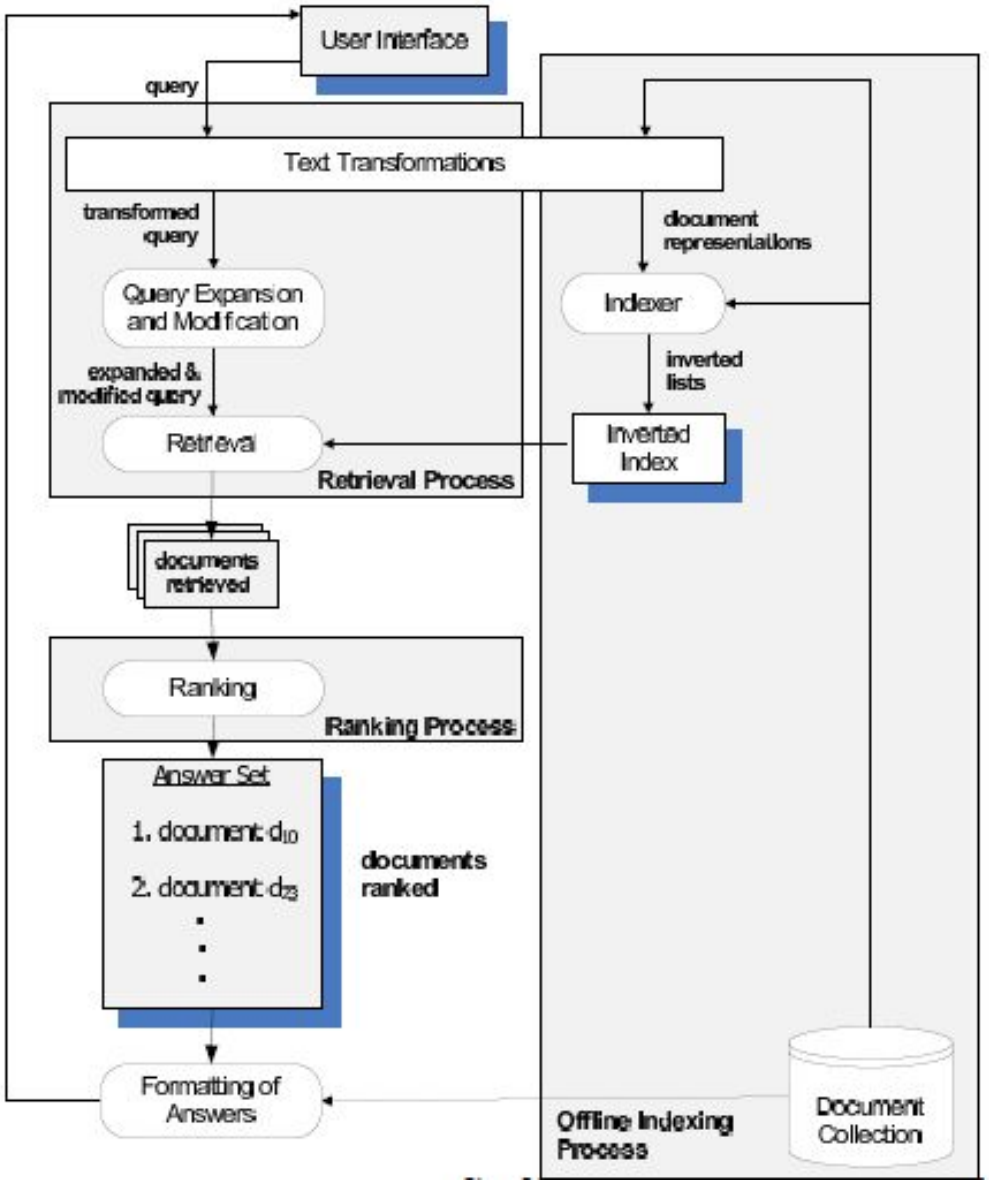
- System wyszukiwania danych (*Data retrieval system*) to np. relacyjna baza z danymi o dokładnie określonej strukturze i semantyce. Zasadniczo nie dopuszcza żadnych błędów w wynikach wyszukiwania.

- Wyszukiwanie danych nie rozwiązuje problemu wydobywania informacji na dany temat

Architektura systemu IR



Proces wyszukiwania i szeregowania



Sieć WWW - historia

- W 1990 roku Berners-Lee
  - opracował protokół HTTP,
  - zdefiniował język HTML,
  - napisał pierwszą przeglądarkę, którą nazwał World Wide Web,
  - opracował pierwszy serwer sieciowy.
- W 1991, udostępnił serwer i przeglądarkę w Internecie
- Tak narodziła się sieć

Wyszukiwanie w sieci

- Wyszukiwarki sieciowe są najpopularniejszymi aplikacjami stosującymi technologię IR wraz z jej zasadniczymi elementami: szeregowaniem i indeksowaniem.

- Sieć narzuca specyficzną charakterystykę wyszukiwania zbioru dokumentów – strony rozproszone w milionach witryn połączonych hiperlinkami; rozproszone dokumenty o pożądanych cechach są wydobywane i kopiowane w jedno miejsce przed ich indeksowaniem. Taki sposób wyszukiwania stron w procesie IR nazywa się „pełzaniem (po stronach)” (*crawling*).

Wyszukiwanie w sieci

- Drugi wpływ sieci na wyszukiwanie to krytyczne znaczenie jakości i skalowalności procesu IR
- Wobec przeszukiwania dużych zbiorów w sieci przwidywanie istotności danych staje się bardzo istotne
- Sieć stanowi także medium do prowadzenia biznesu; strony zawierają linki do ładowania programów, adresy, numery telefonów instytucji itp.
- Przy wyszukiwaniu w sieci należy eliminować spamy
- Zapewnienie bezpieczeństwa, prywatności, praw autorskich i patentowych
- Skanowanie i rozpoznawanie pisma przy wyszukiwaniu w różnych językach