Eksploracja zasobów internetowych

Wykład 1

Wprowadzenie do eksploracji zasobów internetowych

Piotr Hońko
Wydział Informatyki
Politechnika Białostocka

Czym jest eksploracja sieci WWW?

- Eksploracja sieci WWW to zastosowanie technik sztucznej inteligencji, głównie eksploracji danych, w celu odkrywania wiedzy ukrytej w strukturze, treści i sposobie wykorzystania sieci WWW.
- Główne zadania Eksploracji zasobów internetowych:
  - eksploracja struktury/połączeń sieci;
  - eksploracja zawartości sieci;
  - eksploracja sposobu wykorzystania sieci.

Czym jest eksploracja danych?

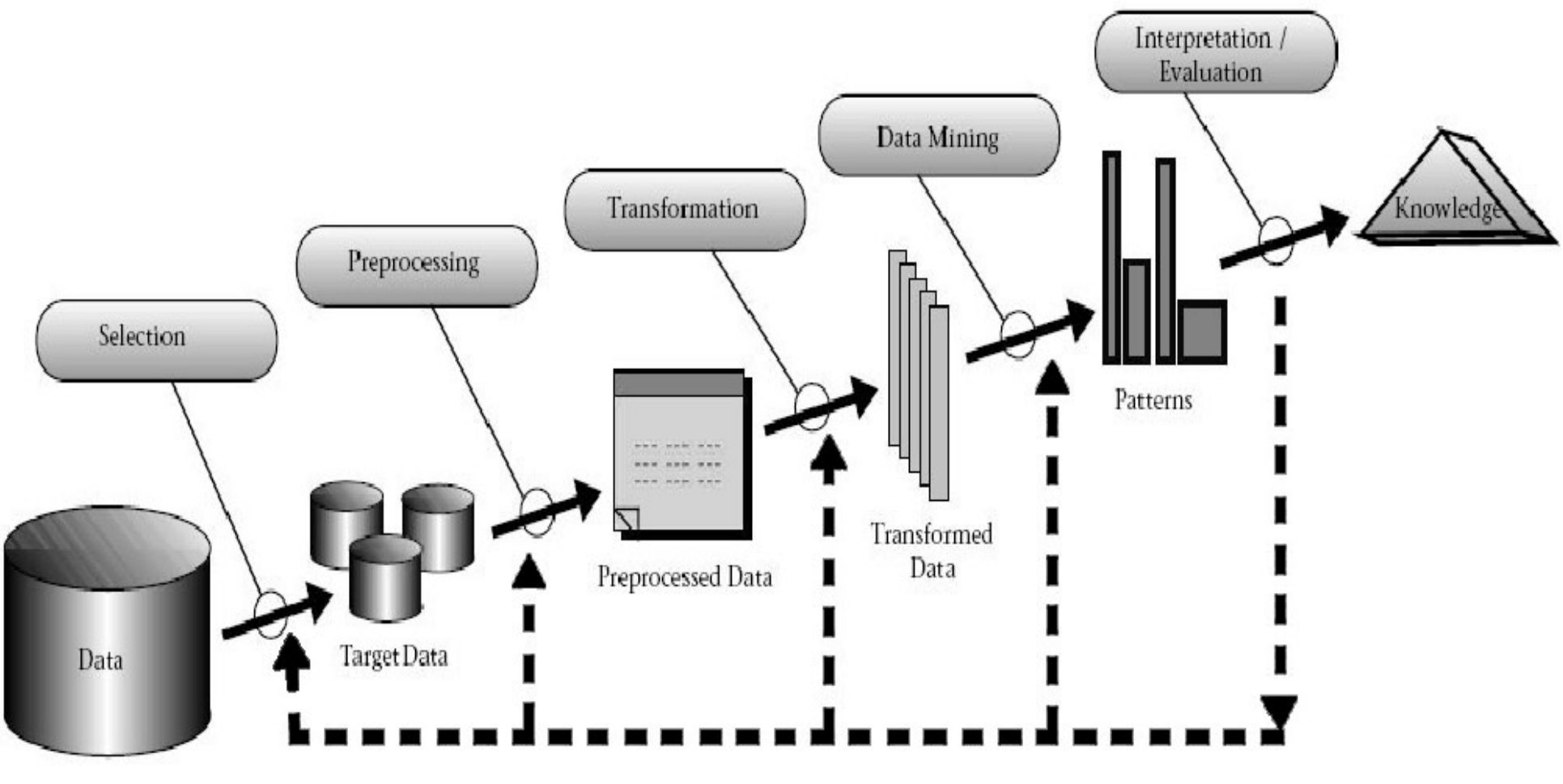
- Eksploracja danych (ang. *Data Mining*) to etap odkrywania wiedzy z baz danych polegający na zastosowaniu analizy danych oraz algorytmów, które produkują na podstawie danych zbiór wzorców lub ich model. (U. Fayyad,G. Piatetsky-Shapiro, P. Smyth)

Czym jest odkrywanie wiedzy w bazach danych?

- Odkrywanie wiedzy w bazach danych (ang. *Knowledge Discovery in Databases*) to nietrywialny proces identyfikacji poprawnych, nowych, potencjalnie użytecznych i ostatecznie zrozumiałych wzorców w danych.

(U. Fayyad,G. Piatetsky-Shapiro, P. Smyth)

Proces odkrywania wiedzy w bazach danych

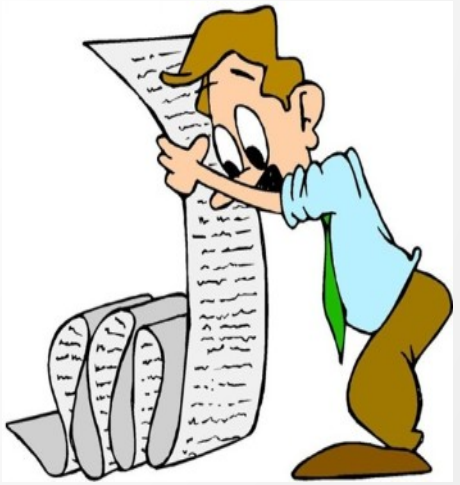


Źródło: U. Fayyad,G. Piatetsky-Shapiro, P. Smyth: From Data Mining to Knowledge Discovery in Databases.

Dwa typy zadań eksploracji danych

- Zadanie predyktywne
  - Określanie nieznanych wartości dotyczących rozpatrywanych obiektów na podstawie znanych wartości opisujących te obiekty.

- Zadanie deskryptywne
  - Znajdowanie zależności zachodzących pomiędzy rozpatrywanymi obiektami.

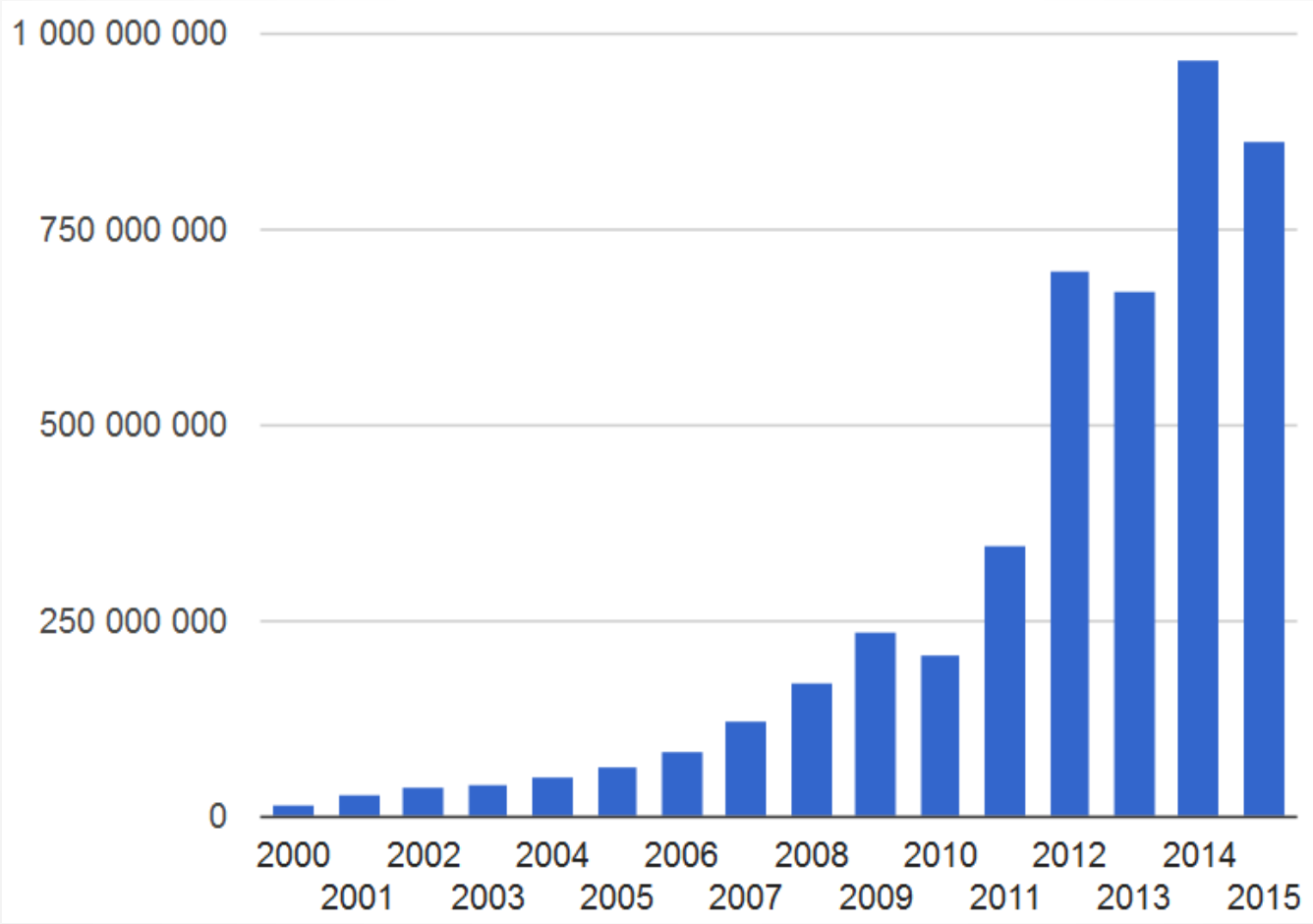


Zadania eksploracji danych

- Klasyfikacja
- Regresja
- Predykcja
- Dyskryminacja
- Grupowanie
- Odkrywanie asocjacji
- Odkrywanie sekwencji
- Odkrywanie charakterystyk

Charakterystyka sieci WWW – liczba stron

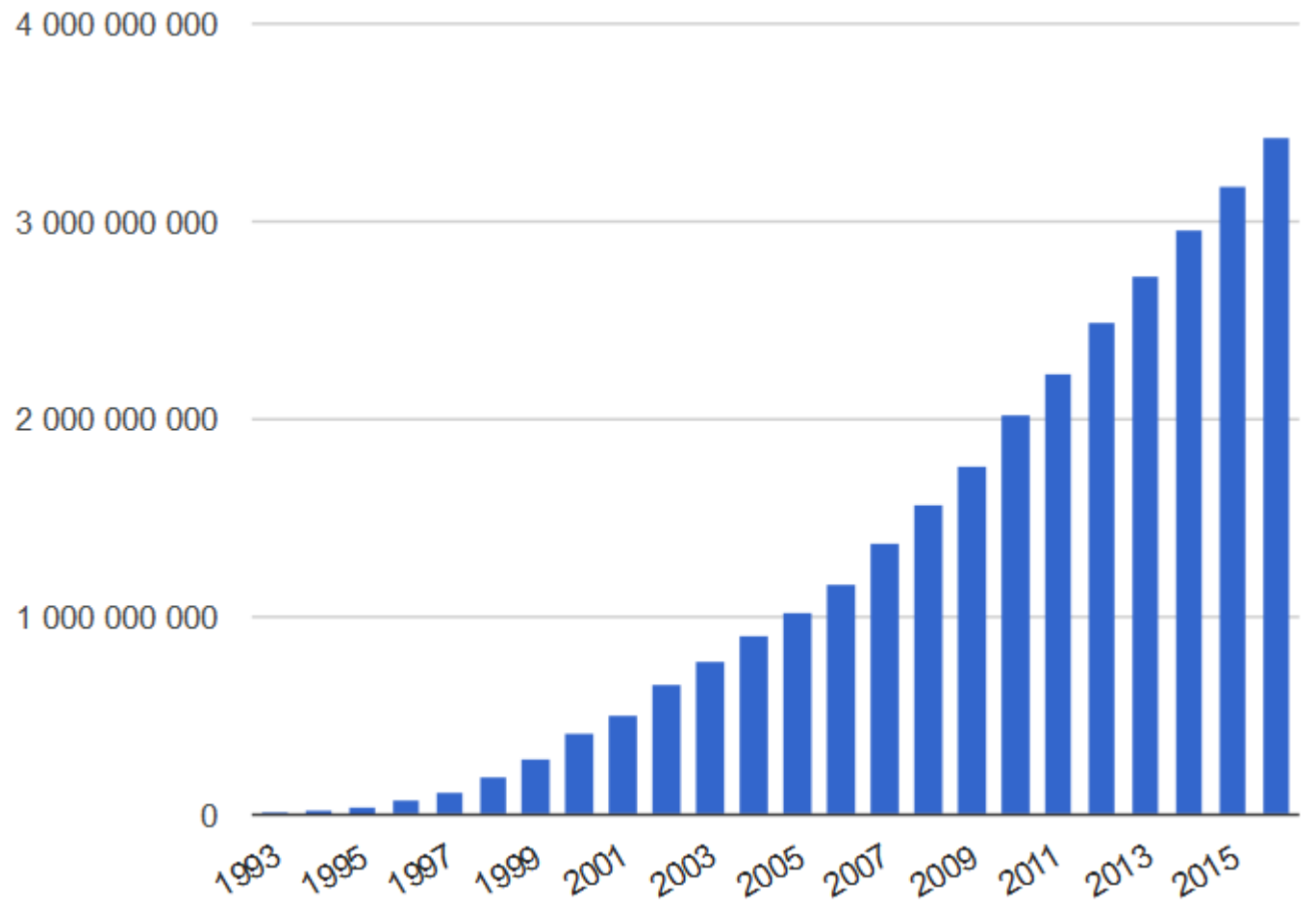
- Liczba stron internetowych (ang. website): 1 806 509 655 (stan z dnia 12.10.2020)



Źródło: http://www.internetlivestats.com/total-number-of-websites/.

Charakterystyka sieci WWW – popularność

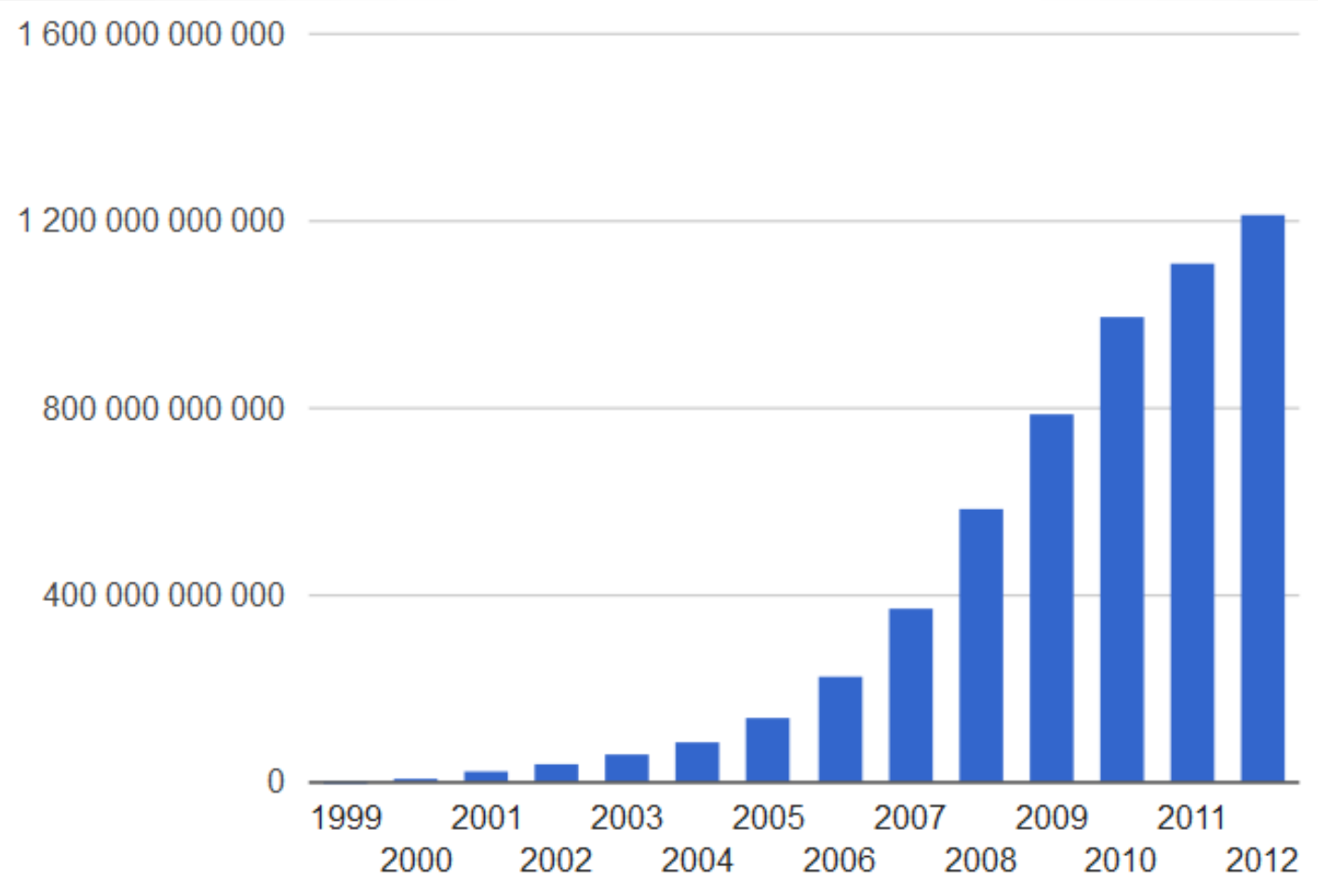
- Liczba użytkowników: 4 703 939 757 (stan z dnia 12.10.2020)



Źródło: http://www.internetlivestats.com/total-number-of-websites/.

Charakterystyka sieci WWW – obciążenie

- Liczba zapytań rocznie (Google):
  2 132 053 388 202 (stan z dnia 12.10.2020)



Źródło: http://www.internetlivestats.com/total-number-of-websites/.

Charakterystyka sieci WWW – rozmiar danych

- Rozmiar danych sieciowych (2020):
  44 ZB = 44 000 EB = 44 000 000 PB =
  44 000 000 000 TB,
- Rozmiar danych sieciowych (2025 – prognozowany):
  175 ZB.
  - Czas pobierania wszystkich danych internetowych:
    1 800 000 000 lat!
  - Długość półki z płytami DVD, zawierającymi wszystkie dane internetowe: 222 obwody ziemi!

Źródło: https://www.nodegraph.se/how-much-data-is-on-the-internet/

Eksploracja danych sieciowych – Eksploracja dużych danych (ang. Big Data mining)

- Rozmiar (ang. **V**olume) – duża ilość danych produkowana każdego dnia;
- Szybkość (ang. **V**elocity) – ogromne tempo produkcji danych;
- Różnorodność (ang. **V**ariety) – zróżnicowanie danych pod względem typów i źródeł;
- Wiarygodność (ang. **V**eracity) – dane o różnym stopniu wiarygodności;
- Wartość (ang. **V**alue) – dane, z których można wydobyć istotną wiedzę.

Eksploracja struktury sieci (ang. *Web linkage mining*)

- Eksploracja struktury fragmentu sieci WWW (np. serwis internetowy) z wykorzystaniem strategii przeszukiwania w celu np.
  - wydobycia informacji o jej organizacji,
  - pobrania danych lub informacji o odnośnikach zawartych w poszczególnych dokumentach struktury,
  - zbudowania rankingu stron wchodzących w skład struktury.

Struktura przeznaczona do reprezentacji sieci WWW

- Sieć WWW można postrzegać jako graf:
  - wierzchołki – dokumenty WWW,
  - krawędzie – odnośniki między dokumentami.
- Właściwości grafu:
  - skierowany,
  - cykliczny,
  - gęsty.

Rodzaje danych dostępnych w sieci WWW

- Dane strukturalne (*ang. structured data*)
  - Dane zorganizowane za pomocą połączeń typu klucz – wartość, np. bazy danych.
- Dane niestrukturalne (*ang. unstructured data*)
  - Dane, które nie są zorganizowane według jakichkolwiek reguł (dane zrozumiałe jedynie dla człowieka), np. czysty tekst.

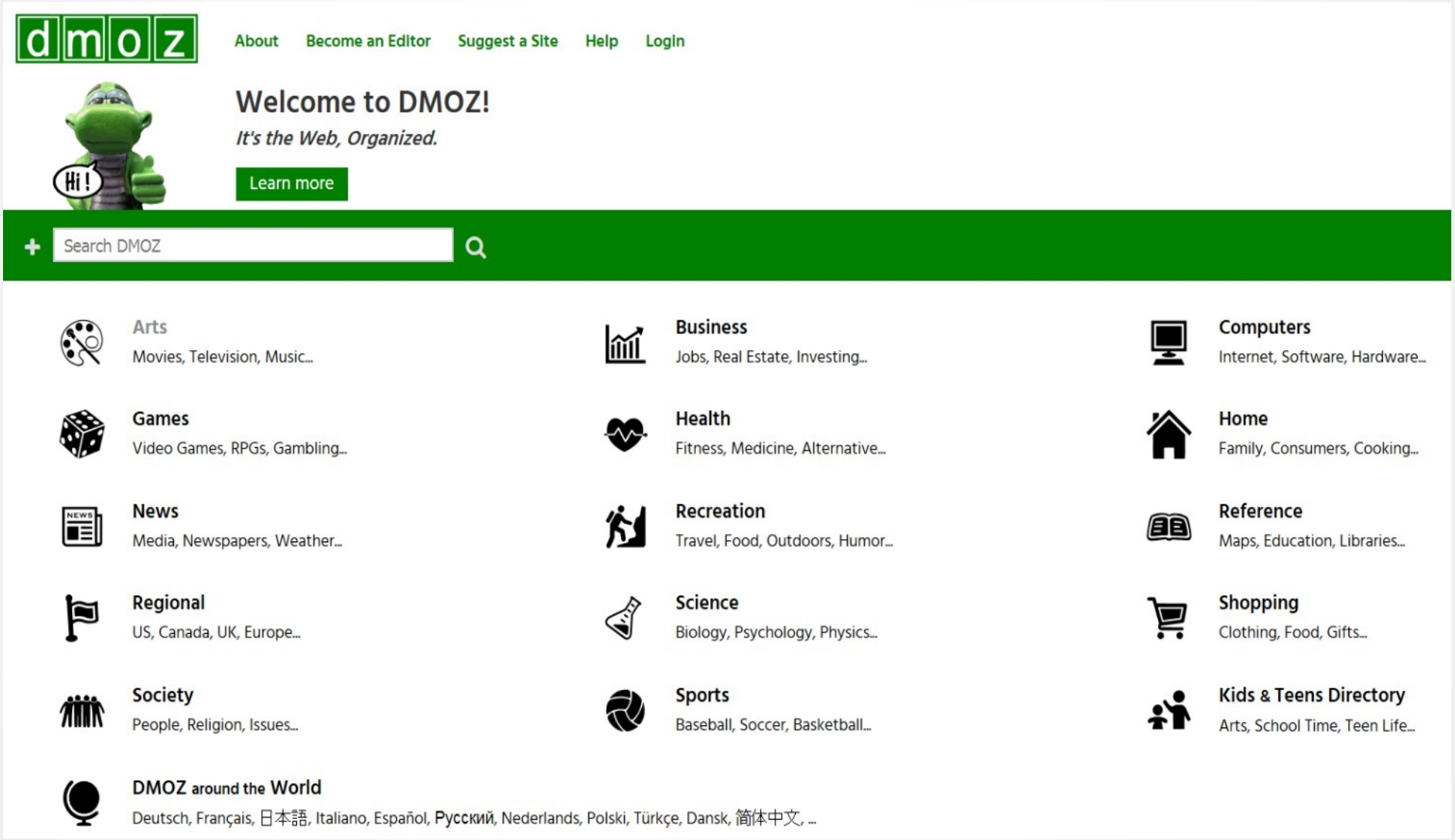
Sieć WWW jako źródło wiedzy

- Pozyskiwanie wiedzy z sieci WWW wymaga zarówno odpowiedniej struktury organizacji zawartości treści, jak i mechanizmów przeszukiwania sieci:
  - Topic Directories,
  - Semantic Web,
  - silniki przeszukiwania sieci.

Topic Directories

- Struktury zawierające strony WWW pogrupowane zgodnie z ich zawartością.
- Nie wymaga przebudowy dokumentów sieciowych.
- Struktura hierarchiczna tworzona ręcznie.

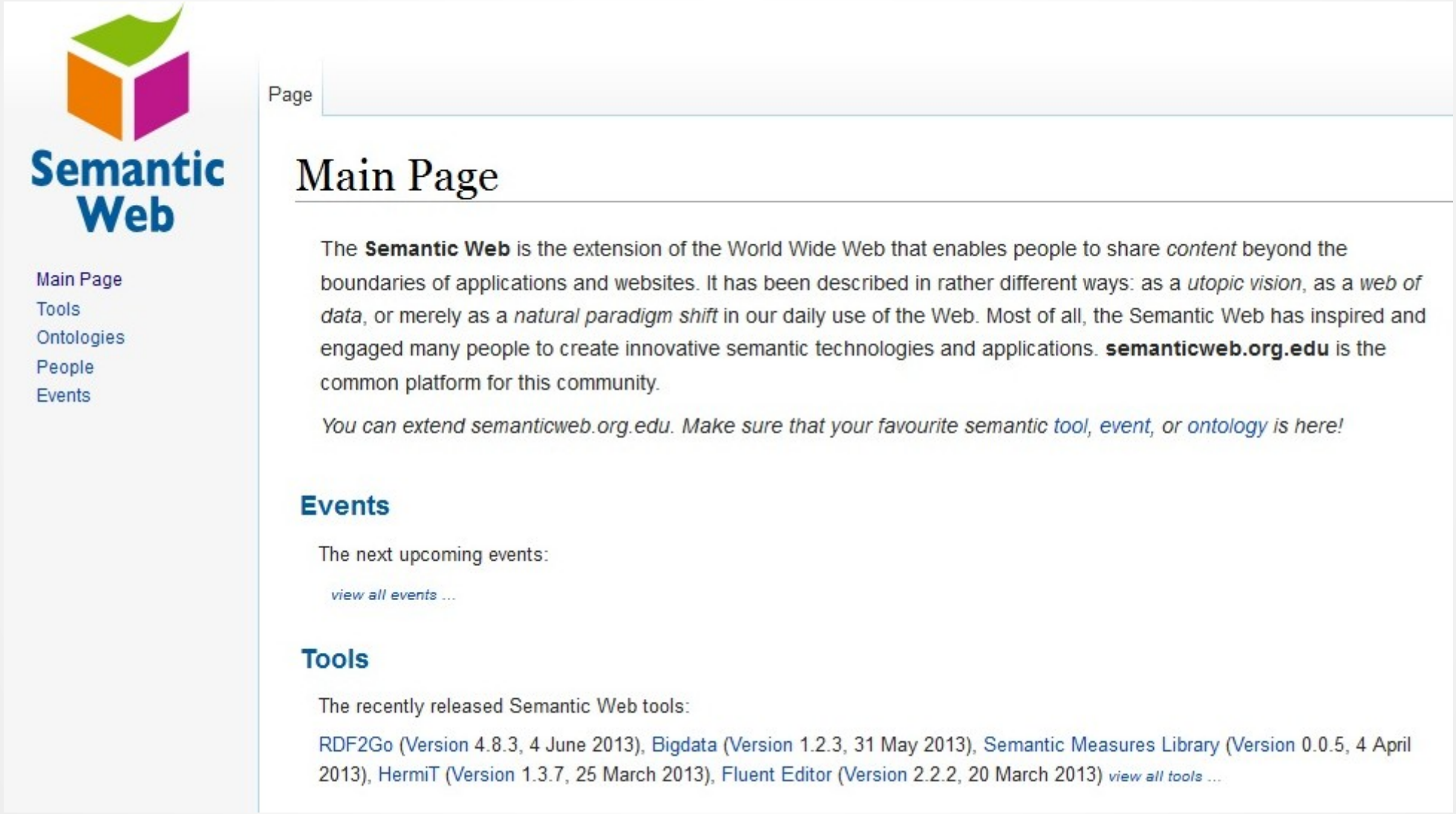
Topic Directories – dmoz.org



Semantic Web

- Projekt mający na celu stworzenie standardów opisu treści sieciowej, umożliwiający automatyczne przetwarzanie informacji uwzględniając jej znaczenie.
- Opisy ułatwiają klasyfikację treści dokumentów.
- Technika dodawania metadanych opisujących zawartość dokumentu do poszczególnych stron WWW.
- Metadane są niewidoczne dla człowieka.
- Wymaga przebudowy dokumentów WWW.

Semantic Web – semanticweb.org



Silnik wyszukiwania

- Pozwala na zwracanie kolekcji dokumentów zgodnych z zapytaniem złożonym ze słów kluczowych.
- Nie wymaga przebudowy dokumentów sieci WWW.
- Pozwala na zwracanie wyników z uwzględnieniem istotności poszczególnych dokumentów.

Silniki wyszukiwania – kategoryzacja

Odkrywanie struktury sieci WWW – robot sieciowy

- Celem robota sieciowego jest analiza struktury sieci WWW i budowanie na jej podstawie grafu oraz pobranie zawartości poszczególnych dokumentów.
- Podstawowe elementami powiązane z robotem sieciowym:
  - parsera języka HTML,
  - algorytmy analizy danych,
  - indeksera odwiedzonych stron.

Wyszukiwanie informacji – Information Retrieval (IR)

- Ideą IR jest wykorzystanie prostej zależności boolowskiej – występowanie lub brak występowania słowa lub słów kluczowych podanych przez użytkownika w dokumencie.
- W IR istotne jest sortowanie wyników zapytań względem ich trafności.
- Silniki wyszukiwania opierają się głównie na IR.

Struktury danych do przechowywania treści pobranych ze stron WWW

- Jedną z często wykorzystywanych struktur jest macierz term-dokument, która pokazuje informację o występowaniu termów w dokumentach.
- Typy macierzy:
  - boolowski,
  - ilościowy,
  - pozycyjny.

Model wektorowy

- Definiowanie dokumentów jako wektorów w wielowymiarowej przestrzeni, gdzie osie są reprezentowane przez termy.
- Typy reprezentacji wektorowej:
  - Term-Frequency (TF),
  - Inverse Document Frequency (IDF),
  - Term Frequency – Inverse Document Frequency (TFIDF).

Normy w ocenie wyników zapytania

- Zwracane wyniki można posortować pod względem istotności, bazującej na termach poprzez obliczenie odległości pomiędzy wektorem reprezentującym zapytanie a pozostałymi wektorami reprezentującymi dokumenty.
- Typowe miary:
  - norma Euklidesowa,
  - podobieństwo cosinusowe.

Ranking wyników na bazie linków

- Idea rozwiązania opiera się na przydzielaniu każdej stronie WWW pewnego współczynnika określającego jej istotność.
- Grafowy charakter sieci WWW umożliwia wzbogacenie sortowania wyników o zależności krawędziowe pomiędzy wierzchołkami grafu.

Współczynnik prestiżu

- Dla każdej zindeksowanej strony WWW przypisywany jest współczynnik prestiżu.
- Jest on obliczany na podstawie liczby linków wskazujących na daną stronę, a także współczynników prestiżu stron linkujących.

Eksploracja treści (ang. *Web content mining*)

- Konstrukcja dedykowanej bazy danych na podstawie treści dokumentów WWW.
- Ewentualna transformacja danych w celu dopasowania ich formatu do wymogów narzędzi eksploracji danych.
- Zastosowanie wybranego narzędzia eksploracji danych.
- Wykorzystanie pozyskanej wiedzy z danych.

Konstrukcja bazy danych

- Zapis danych w postaci:
  - pojedynczej tabeli, np. macierz typu term-dokument,
  - relacyjnej bazy danych, np. dodatkowe tabele mogą pokazywać powiązania pomiędzy różnego typu dokumentami.
- Zwykle jeden rekord pojedynczej tabeli (lub głównej tabeli relacyjnej bazy danych) odpowiada jednemu dokumentowi.

Transformacja danych

- Baza po utworzeniu może zostać przekształ-cona, stosując techniki takie, jak
  - filtrowanie danych,
  - selekcja cech,
  - dyskretyzacja,
  - uzupełnianie danych niekompletnych.

Narzędzia eksploracji danych

- Zastosowanie algorytmu, którego celem jest wydobycie wiedzy dotyczącej danych sieciowych.
- Wiedza ta jest wyrażana zwykle za pomocą wzorców:
  - reguły decyzyjne/regresyjne,
  - drzewa decyzyjne/regresyjne,
  - wzorce częste,
  - reguły asocjacyjne.

Zadania eksploracji danych sieciowych – grupowanie

- Idea grupowania
  - Podział danych obiektów na grupy, w taki sposób, że obiekty należące do jednej grupy są do siebie zbliżone, a obiekty z różnych grup są od siebie oddalone.
- Zastosowanie grupowania
  - automatyczne tworzenie Topic Directories,
  - grupowanie wyników zwracanych przez wyszukiwarki internetowe w celu zwiększenia trafności wyszukiwania.

Zadania eksploracji danych sieciowych – klasyfikacja

- Idea klasyfikacji
  - Konstrukcja mechanizmu umożliwiającego automatyczne przypisywanie danego dokumentu na podstawie jego cech do jednej z predefiniowanych grup.
- Zastosowanie klasyfikacji
  - automatyczne rozbudowywanie Topic Directories,
  - określanie profilu użytkownika na podstawie danych umieszczanych przez niego w sieci.
  - wykrywanie spamu w wiadomościach elektronicznych.

Zadania eksploracji danych sieciowych – odkrywanie asocjacji

- Idea odkrywanie asocjacji
  - Wykrywanie cech wspólnych dla odpowiednio dużej grupy obiektów (wzorce częste).
  - Odkrywanie często występujących zależności określonych za pomocą cech obiektów (reguły asocjacyjne).
- Zastosowanie asocjacji
  - automatyczne uzupełnianie tekstów w wyszukiwarkach na podstawie słów często występujących razem,
  - rekomendacja produktów internetowych.

Wykorzystanie wiedzy pozyskanej z danych

- Wzbogacenie opisu badanych przypadków poprzez wykrycie nowych zależności, np. znalezienie nowych wspólnych cech dokumentów należących do jednej kategorii.
- Budowa systemu wykorzystującego w automatyczny sposób nową wiedzę w procesie analizy nowych przypadków, np. system decyzyjny określających kategorię nowego dokumentu WWW.

Eksploracja sposobu wykorzystania sieci (ang. *Web usage mining*)

- Konstrukcja dedykowanej bazy danych na podstawie sposobu wykorzystani sieci przez użytkowników.
- Ewentualna transformacja danych w celu dopasowania ich formatu do wymogów narzędzi eksploracji danych.
- Zastosowanie wybranego narzędzia eksploracji danych.
- Wykorzystanie pozyskanej wiedzy z danych.

Konstrukcja bazy danych

- Zapis danych w postaci:
  - pojedynczej tabeli, która zawiera cechy dotyczące sposobu użytkowania sieci, utworzone np. za pomocą narzędzi statystycznych.
  - relacyjnej bazy danych, pokazująca zależności pomiędzy użytkownikami lub pomiędzy użytkownikiem a dokumentami, które przeglądał.
- Zwykle jeden rekord pojedynczej tabeli (lub głównej tabeli relacyjnej bazy danych) odpowiada jednemu użytkownikowi.

Mechanizmy wykorzystywane do utworzenia bazy

- Analiza strumienia stron, czyli ciągu stron odwiedzanych przez użytkownika.
- Oczyszczanie i filtrowanie danych.
- Identyfikacja użytkownika i jego sesji użytkownika.
- Kompletowanie ścieżki odwiedzin.

Zadania eksploracji sposobu wykorzystania sieci

- Grupowanie – podział użytkowników na grupy w zależności od sposobu wykorzystania przez nich sieci.
- Klasyfikacja – identyfikacji istotnych cech dotyczących sposobów korzystania z danego serwisu.
- Analiza asocjacyjna – określenie, jakie cechy dotyczące danego sposobu wykorzystania sieci
  - występują często razem (wzorce częste),
  - są od siebie zależne (reguły asocjacyjne).
- Odkrywanie sekwencji – określanie często występujących ciągów odwiedzin stron.

Zaliczenie wykładu

- Egzamin pisemny (wersja elektroniczna):
  - Zadania zamknięte (test wielokrotnego wyboru) – sprawdzenie znajomości/ rozumienia podstawowych pojęć EZI;
  - Zadania otwarte (obliczeniowe) – sprawdzenie znajomości podstawowych mechanizmów stosowanych w EZI.