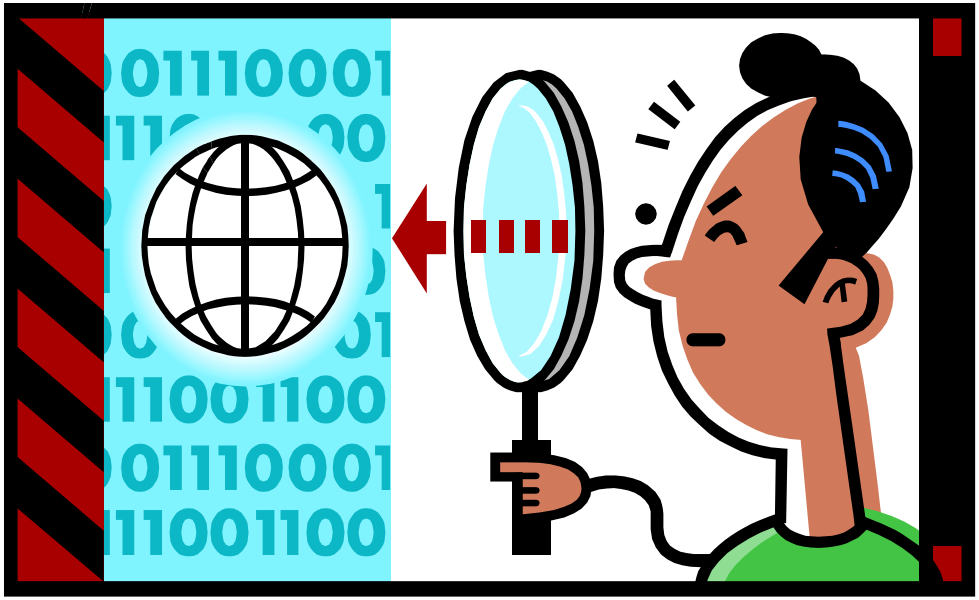
Wyszukiwanie informacji w internecie



Nguyen Hung Son

Jak znaleźć informację w internecie?

- Wyszukiwarki internetowe:
  - Potężne machiny wykorzystujące najnowsze metody z różnych dziedzin
- Architektura: trzy główne moduły
  - Zarządzanie pająków;
  - Serwer indeksowania;
  - Interfejs użytkownika
- Wyniki wyszukiwania:
  - Lista rankingowa

Architektura

Menedżer indeksów

Serwer indeksowania

Wstępne przetwarzanie i tworzenie indeksów

Wyszukiwarka Internetowa

Menedżer pająków

Adresy url

Zawartość strony

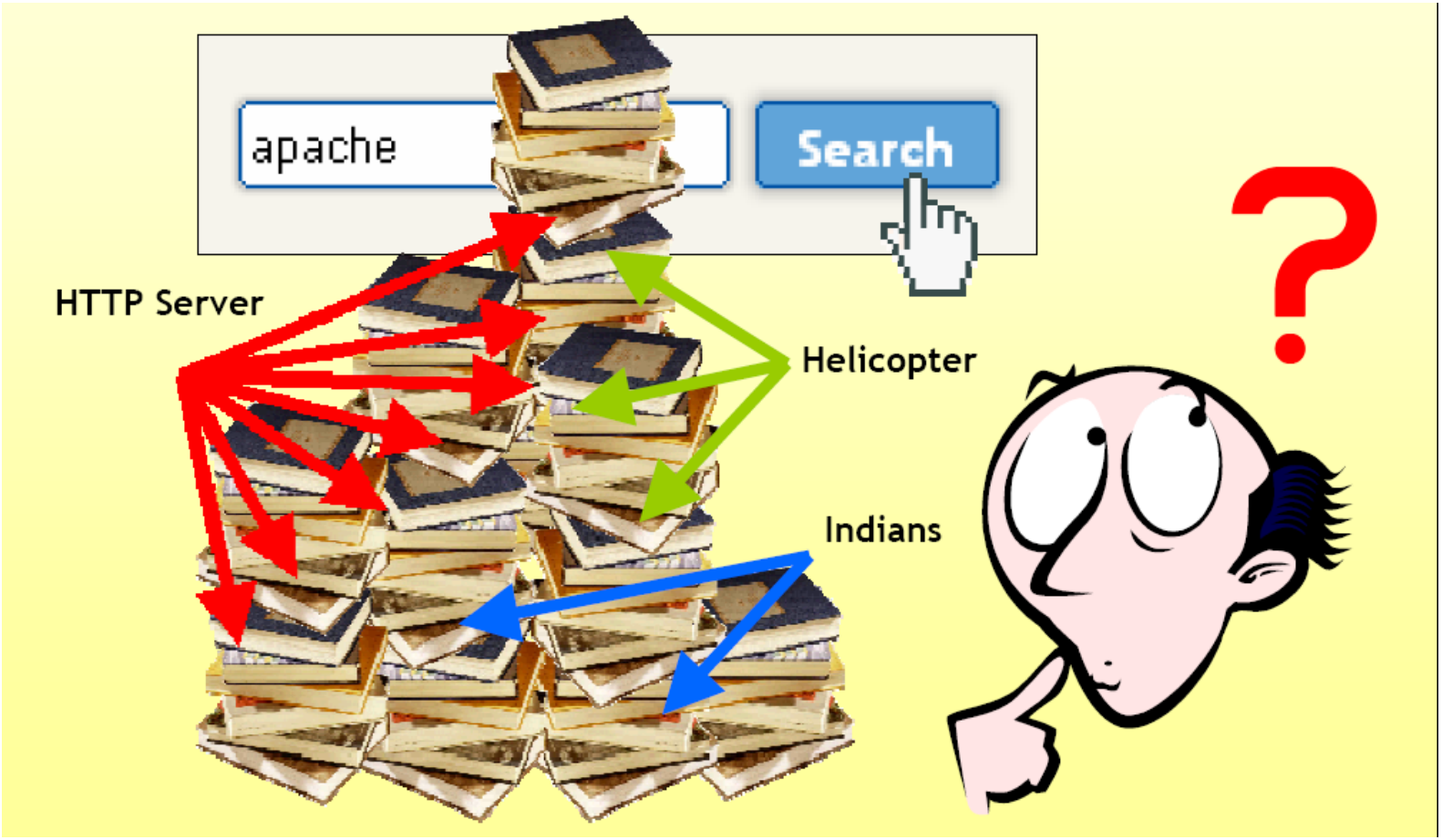
Pająki

Wyniki

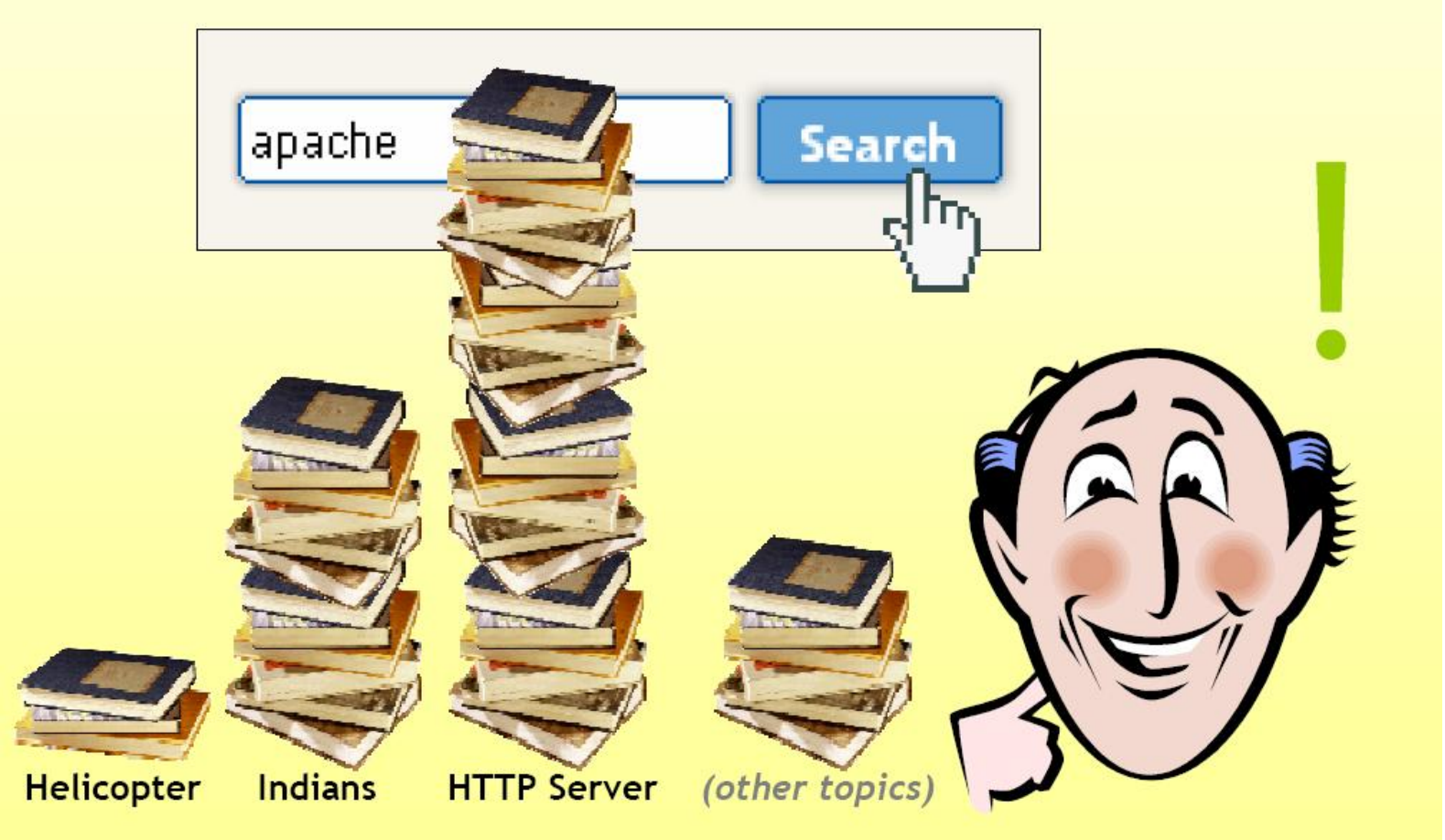
zapytania

Lista rankingowa nie jest doskonała!

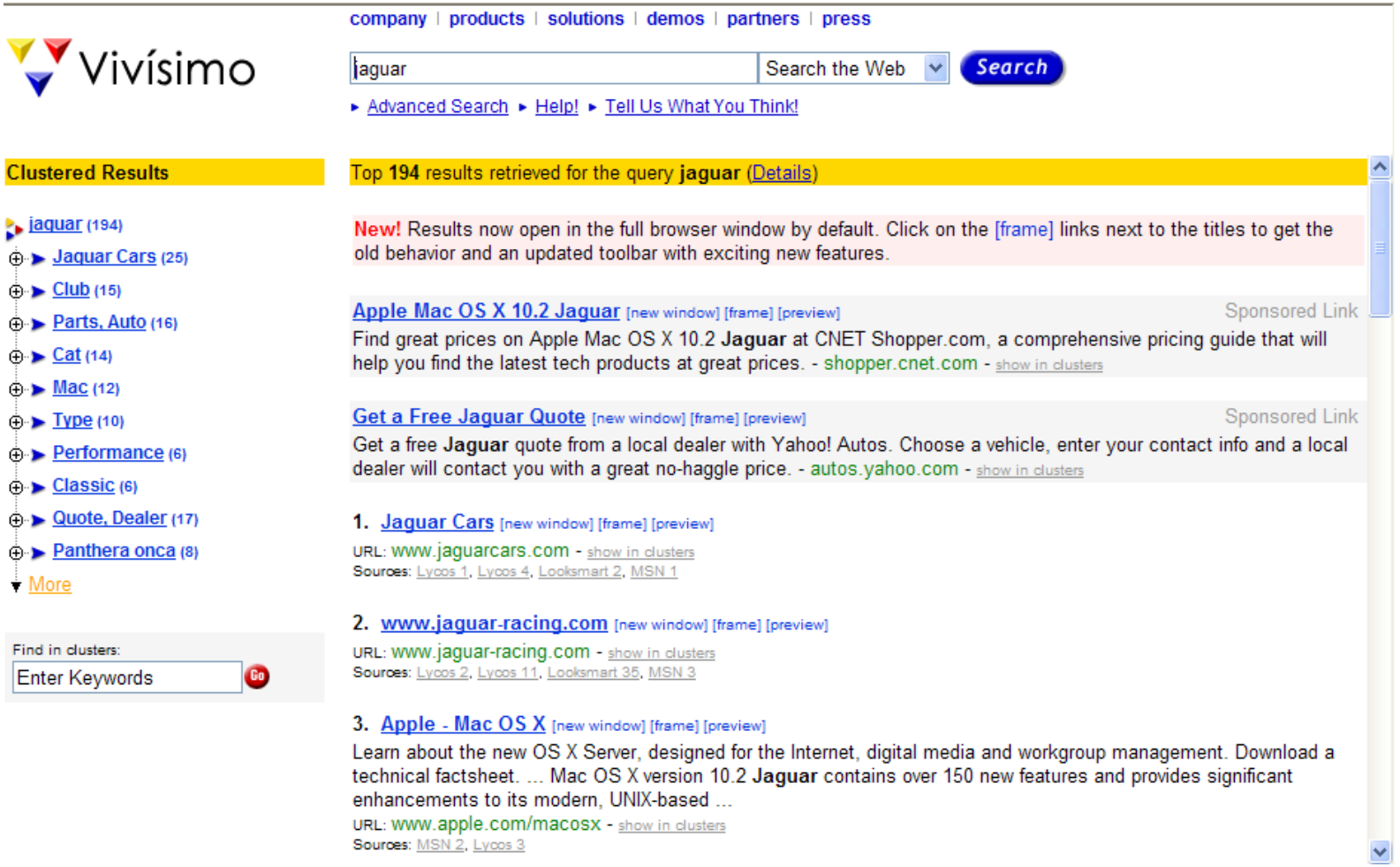
Lista rankingowa nie jest doskonała!



Grupowanie wyników wyszukiwania (ang. *SRC: Search Result Clustering*)



SRC korzysta z krótkich fragmentów tekstu (snippets)



SRC czy grupowanie dokumentów?

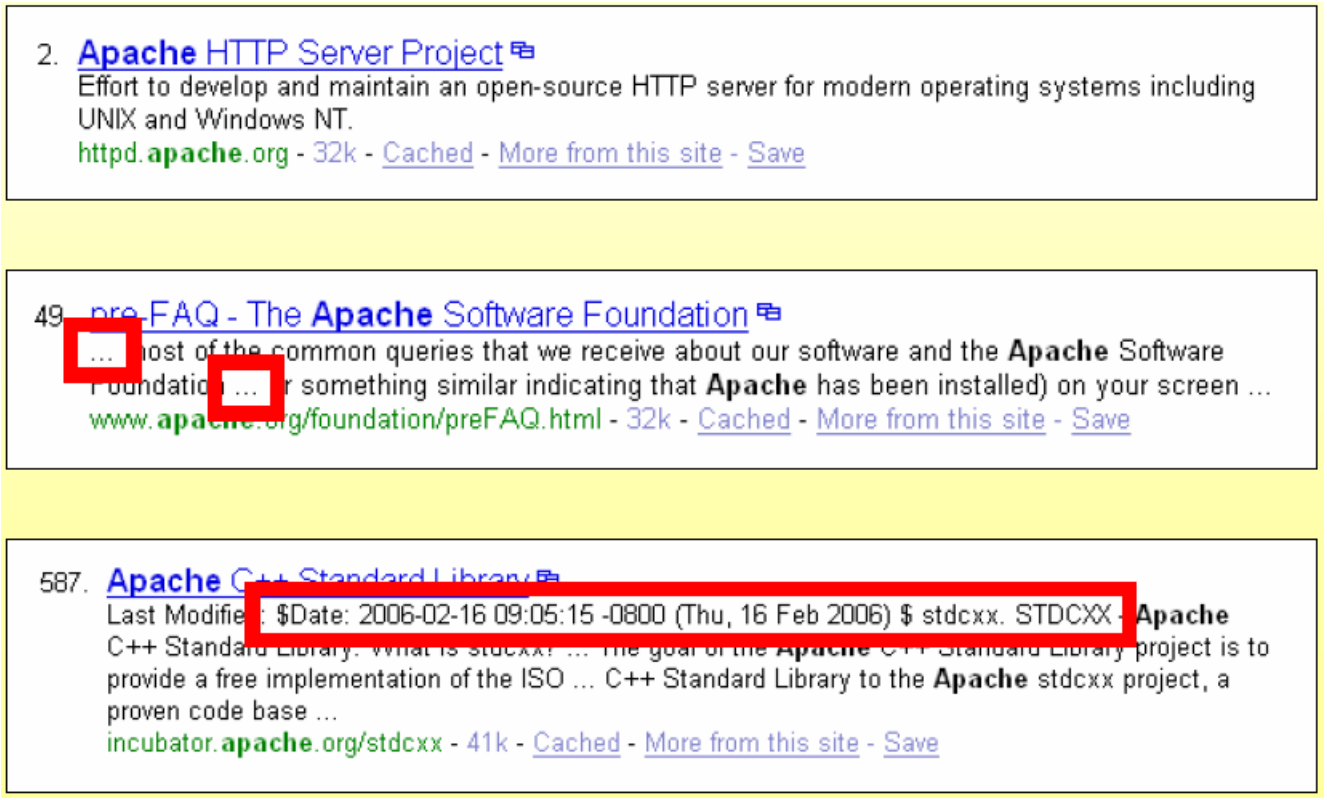
Grupowanie dokumentów:

- Miliardy stron;
- Ich treści ciągle się zmieniają;
- Skalowalność wzg. liczby dokumentów
- Są to niestrukturalne i różnorodne dane;
- dodatkowe informacje:
  - hiperłącze,
  - przejścia między stronami (click-through data), itp.

SRC

- Próba 100~400 wyników wyszukiwania
- Informacje są aktualne
- Działa na bieżąco
- Skalowalność wzg. potrzeby użytkownika
- Zbyt mała, zaszumiona informacja → gorsza jakość grup

Problemy w SRC



2. **Apache** HTTP Server Project
Effort to develop and maintain an open-source HTTP server for modern operating systems including UNIX and Windows NT.
httpd.**apache**.org - 32k - Cached - More from this site - Save

49. pre FAQ - The **Apache** Software Foundation
... most of the common queries that we receive about our software and the **Apache** Software Foundation ... r something similar indicating that **Apache** has been installed) on your screen ...
www.**apache**.org/foundation/preFAQ.html - 32k - Cached - More from this site - Save

587. **Apache** C++ Standard Library
Last Modifie : $Date: 2006-02-16 09:05:15 -0800 (Thu, 16 Feb 2006) $ stdcxx. STDCXX - **Apache** C++ Standard Library. What is stdcxx? ... The goal of the **Apache** C++ Standard Library project is to provide a free implementation of the ISO ... C++ Standard Library to the **Apache** stdcxx project, a proven code base ...
incubator.**apache**.org/stdcxx - 41k - Cached - More from this site - Save

Wymagania

- Kryteria oceniania jakości metod SRC:
  - *Semantyczność*: dokumenty w jednej grupie powinny dotyczyć tego samego tematu
  - *Znaczenie etykiet grup*: powinny one dobrze opisać zawartość całej grupy.
  - *Mała liczba grup*: należy pokryć jak najwięcej dokumentów używając przy tym jak najmniej grup.
- Te kryteria są raczej subiektywne aniżeli obiektywne.

Model wektorowy dokumentów

- $\boldsymbol{T}=\{t_1,\ldots,t_n\}$ – zbiór wybranych wyrazów (słów, fraz)
- Dokument $\boldsymbol{d}_i = [w_{i,1},\ldots,w_{i,n}]$ gdzie
  $w_{i,j}$ jest wagą wyrazu t_j w dokumencie $\boldsymbol{d}_i$
- Schemat ważenia wyrazów – TFxIDF

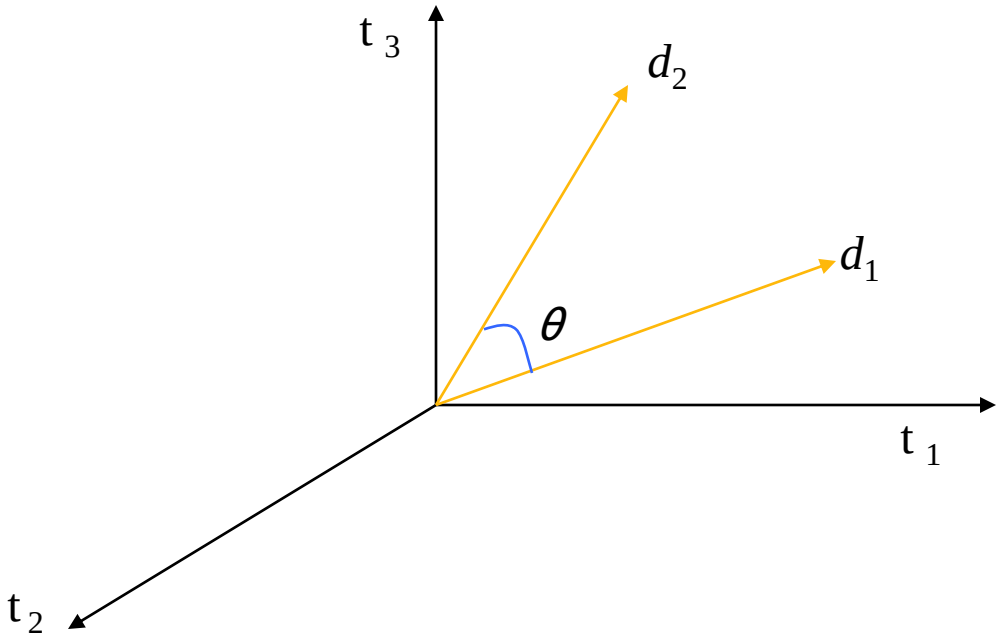
$$w_{i,j} = f_{i,j} \times \log \frac{N}{df(t_j)}$$

$w_{i,j}$: częstość występowania wyrazu t_j w dokumencie $\boldsymbol{d}_i$
N : liczba dokumentów
$df(t_j)$: liczba dokumentów zawierających t_j

Podobieństwo dokumentów

- Miara cosinusa:

$$sim(d_j, d_k) = \frac{\vec{d}_j \cdot \vec{d}_k}{\left|\vec{d}_j\right|\left|\vec{d}_k\right|} = \frac{\sum_{i=1}^{n} w_{i,j} w_{i,k}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2} \sqrt{\sum_{i=1}^{n} w_{i,k}^2}}$$



Istniejące metody

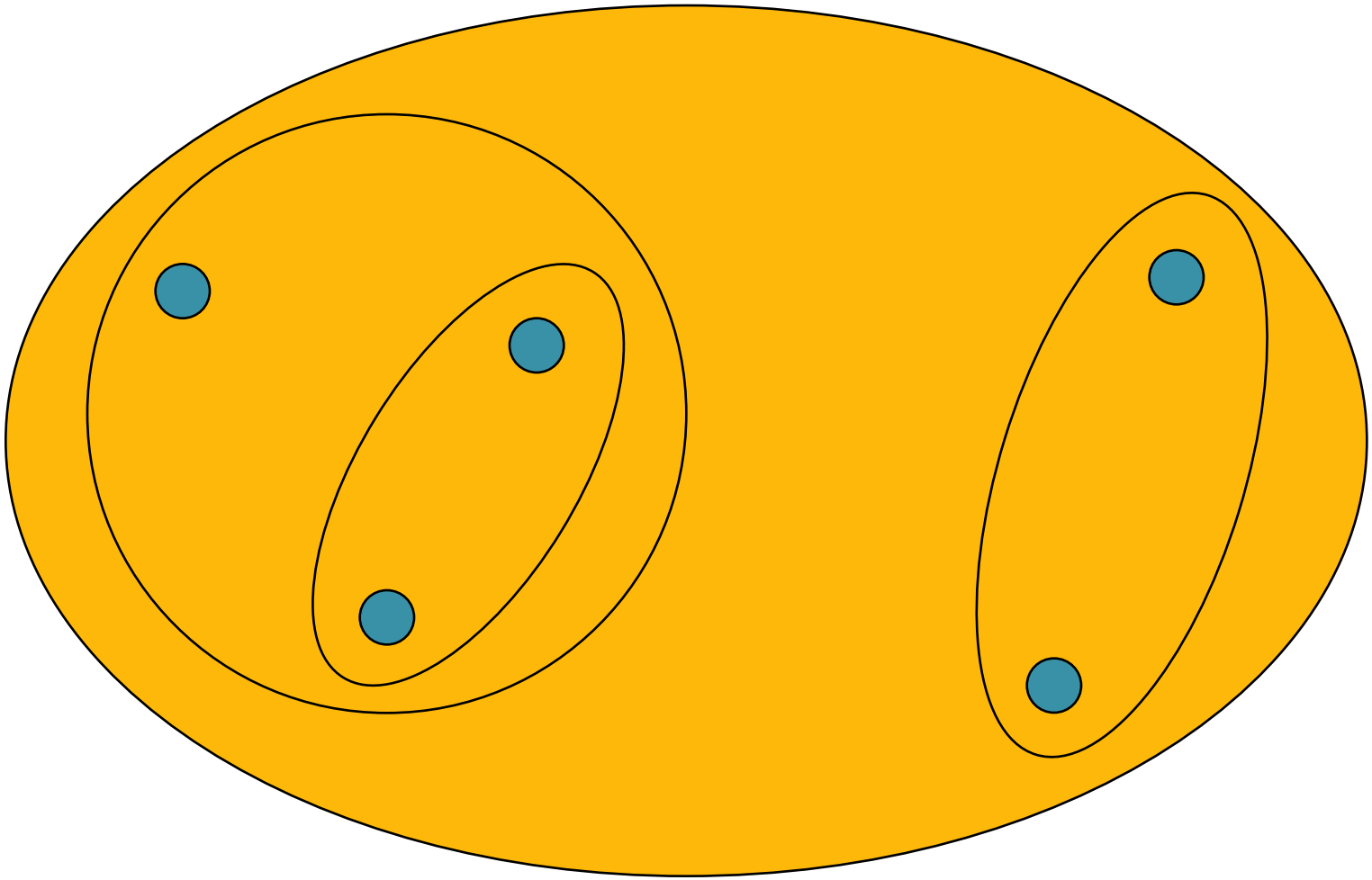
Name	Word Flat	Sentences Flat	Word Hier.	Sentences Hier.	Online/ Software
WebCat	+				+
Retriever	+				
Scatter/Gather	+				
Wang *et al.*	+				
Grouper		+			
Carrot		+			+
Lingo		+			+
Microsoft		+			
FICH			+		+
Credo			+		+
IBM			+		
SHOC				+	
CIIRarchies				+	+
LA				+	
Highlight				+	+
SnakeT				+	+
Mooter			+		+
Vivisimo				+	+

Klasyfikacja algorytmów grupowania

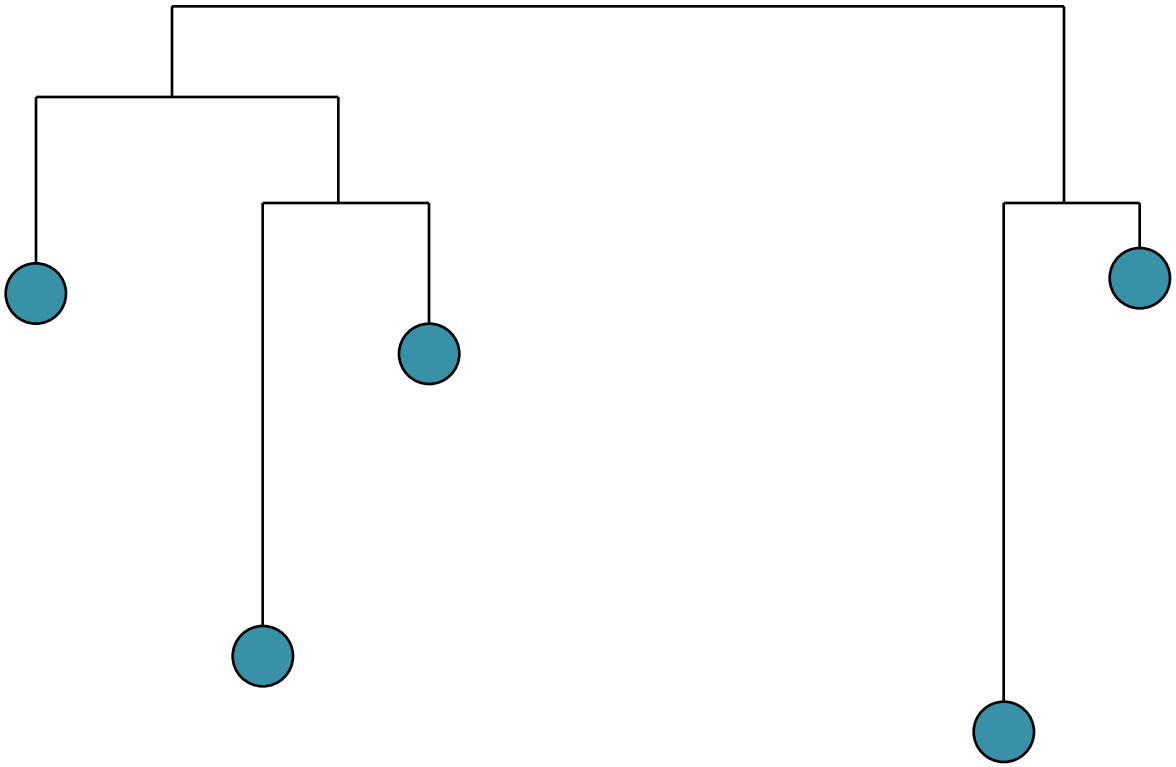
- Płaska struktura czy hierarchiczna?
- Czy grupy są rozłączne?
- Ostry czy miękki podział?
- Przyrostowa metoda?
- Czy liczba grup jest z góry zadana?
- Czy miary odległości lub podobieństwa muszą być zadane z góry?

- Z użyciem odległości
  - Hierarchiczna struktura
    - Agglomerative Hierarchical Clustering (AHC)
  - Płaska struktura
    - K-centroidów (możliwe rozmycie)
    - Inkrementalna (Single-pass)
- Inne
  - Suffix Tree Clustering (Grouper)
  - SOM (Kohonen)
  - Latent Semantic Indexing (LSI) (zmniejsza wymiar)

Grupowanie hierarchiczne (AHC)

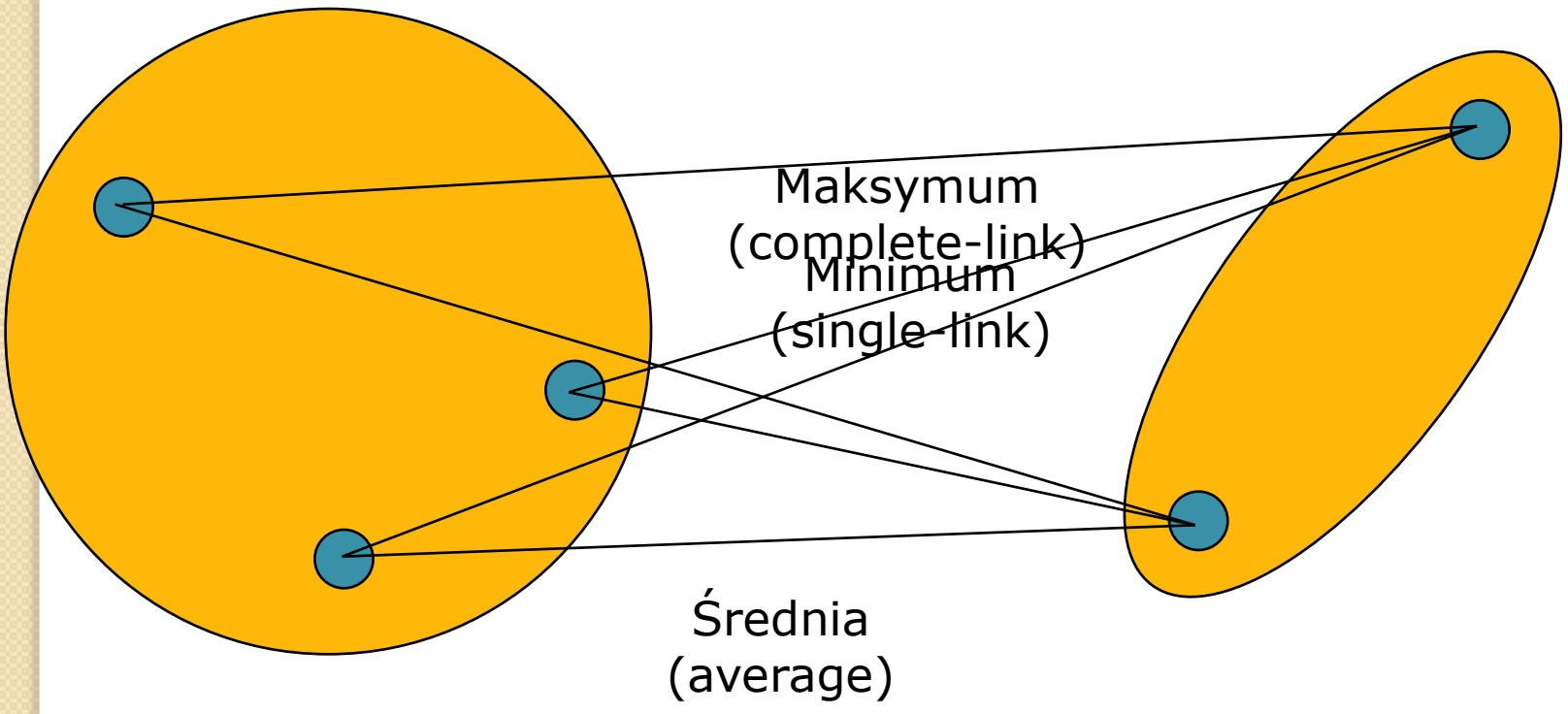


Wynik grupowania: hierarchia pojęć

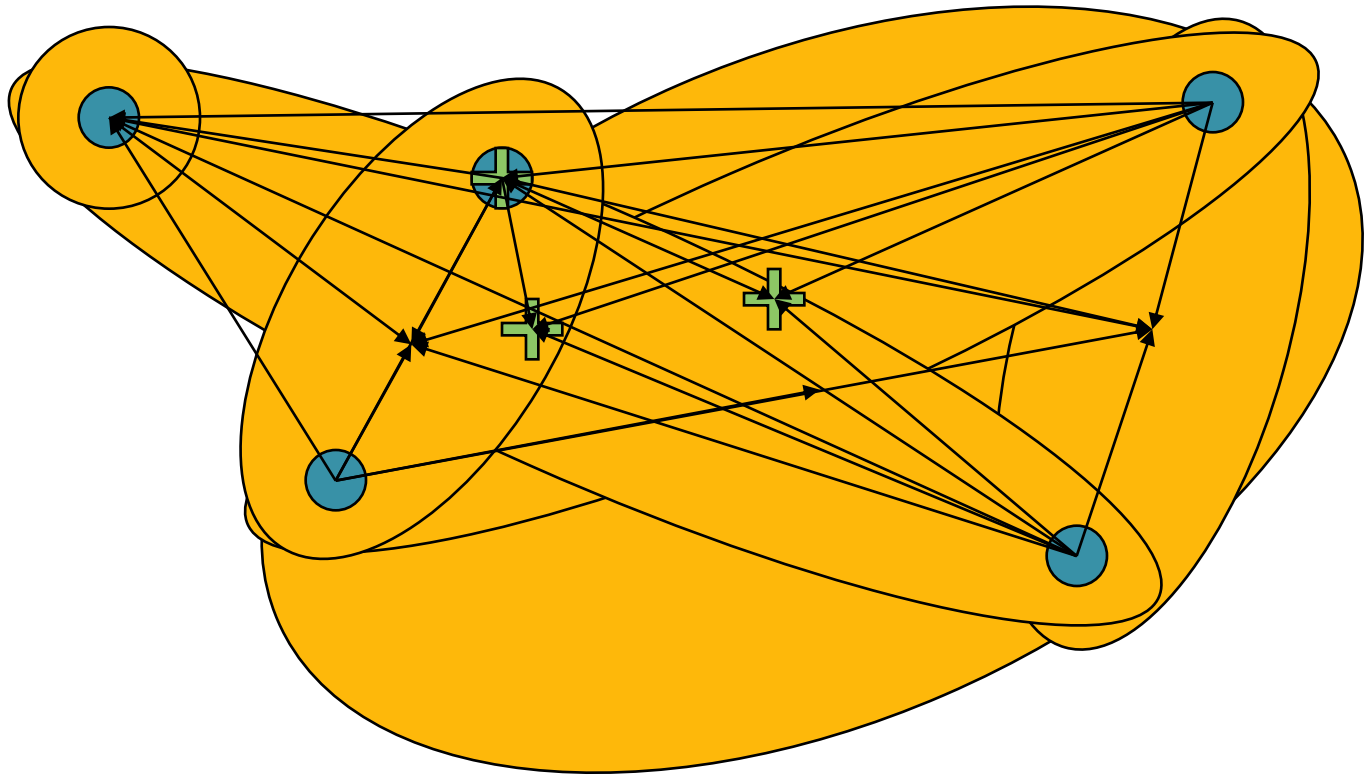


Różne wersje AHC

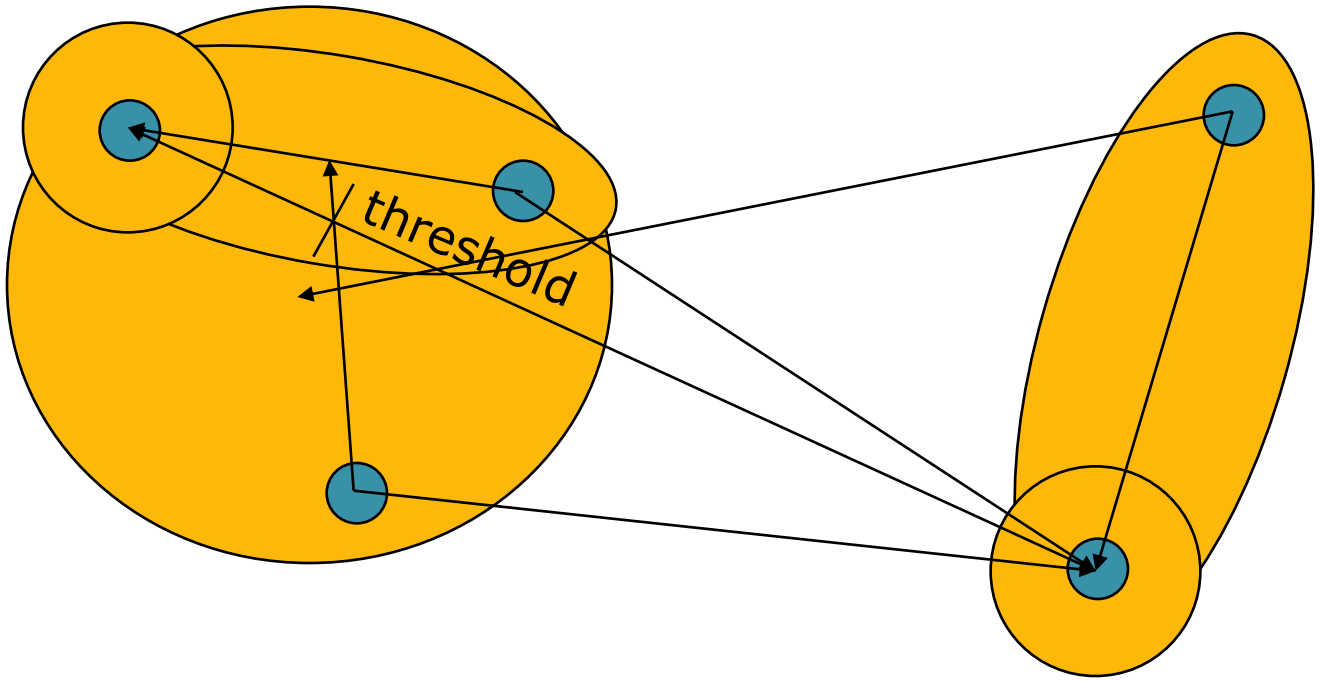
- Istnieją różne metody mierzenia podobieństwa grup



K-centroidów (k=3)



Metoda inkrementalna (single-pass)



„Grouper”
(Zamir and Etzioni 1997, 1999)

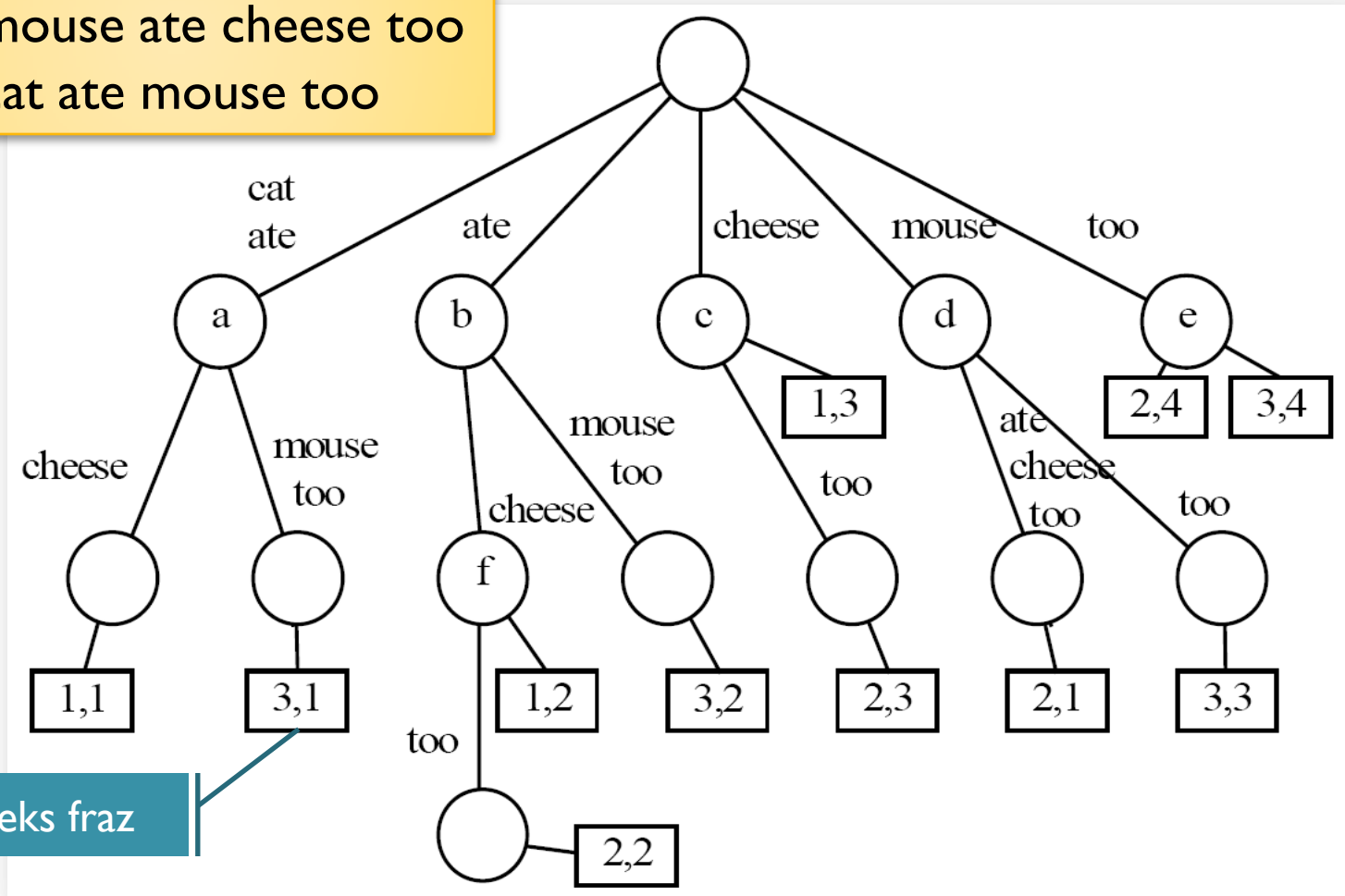
- Działa na bieżąco (online)
- Grupuje wyniki wyszukiwania (snippets)
- Grupuje dokumenty, które mają wiele wspólnych fraz
- Grupowanie drzewem sufiksowym (STC - Suffix Tree Clustering)
  - Czas liniowy
  - Metoda inkrementalna
  - Grupy nie są rozłączne
  - Może być hierarchiczna.

Algorytm STC (Suffix Tree Clustering)

- Krok 1: Czyszczenie danych:
  - Normalizacja (stemming, stop-words elimination)
  - Identyfikacja fraz i zdań.
  - Eliminacja znaków interpunkcyjnych.
- Krok 2: Budowa drzewa sufiksowego:
  - Stworzenie grup bazowych
  - Ocena grup bazowych za pomocą ich rozmiaru i ocen fraz
- Krok 3: Łączenie grup bazowych:
  - Grupy mające „dużą część wspólną” są połączone.

Drzewo sufiksowe = minimalne drzewo zawierające sufiksy wszystkich napisów

1. cat ate cheese
2. mouse ate cheese too
3. cat ate mouse too



Odwrotny indeks fraz

Krok 2 – Identyfikacja grup bazowych

a	cat ate	1,3
b	ate	1,2,3
c	cheese	1,2
d	mouse	2,3
e	too	2,3
f	ate cheese	1,2

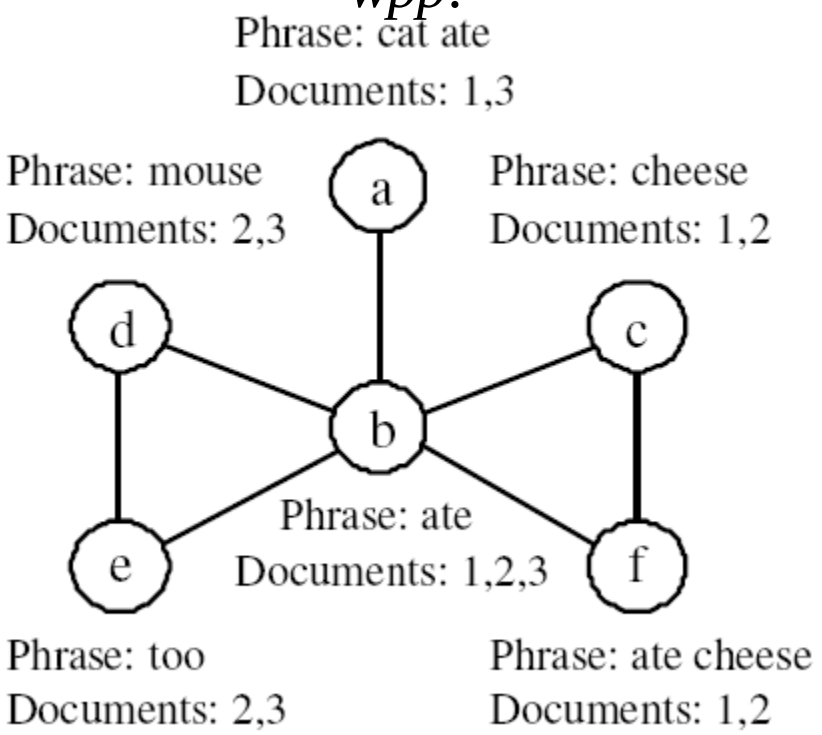
- Wierzchołki reprezentują grupy dokumentów mających wspólną frazę
- Każda grupa *B* definiowana przez frazę *P* jest oceniona przez $S(B) = |B|f(|P|)$

Krok 3 – Łączenie grup bazowych

- Podobieństwo między grupami bazowymi:

$$sim = \begin{cases} 1 & \frac{|B_n \cap B_m|}{|B_n|} > 0.5 \; oraz \; \frac{|B_n \cap B_m|}{|B_m|} > 0.5 \\ 0 & wpp. \end{cases}$$

- Łączymy grupy algorytmem przyrostowym



„Lingo”
(S.Osiński, D.Weiss)

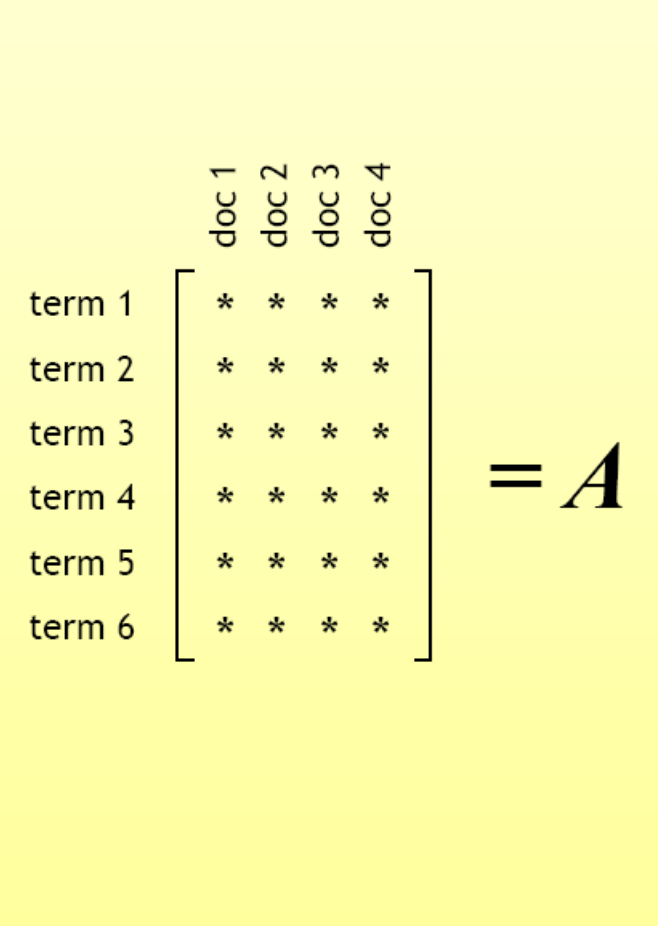
- Korzysta z rozkładu macierzy wzg. wartości osobliwych (SVD)
- Reprezentacja zbioru dokumentów (snippets) w przestrzeni rzutowej o małym wymiarze
- Wektory osobliwe wyznaczają etykiety grup
- Dokumenty są dopisane do grup według miary cosinusa.
- Implementacja:

  Carrot2: Search Results Clustering Framework

  http://www.carrot2.org

  http://carrot.cs.put.poznan.pl

Rozkład wzg. wartości osobliwych
(ang. SVD - Singular Value Decomposition)



- $\boldsymbol{A}$ – macierz m x n

$$\boldsymbol{A} = \boldsymbol{U}\,\boldsymbol{\Sigma}\,\boldsymbol{V}^T$$

  - Kolumny $\boldsymbol{U}$ – *wektory własne* $\boldsymbol{A}\boldsymbol{A}^T$
  - Kolumny $\boldsymbol{V}$ – *wektory własne* $\boldsymbol{A}^T\boldsymbol{A}$
  - $\boldsymbol{\Sigma} = diag(\sigma_1, \ldots, \sigma_n)$:
    $\sigma_1, \ldots, \sigma_n$ wartości osobliwe $\boldsymbol{A}$
    $\sigma_1 > \ldots > \sigma_k \ldots > \sigma_n$

- **Aproksymacja:**

$$\boldsymbol{A} \approx \boldsymbol{U}_k\,\boldsymbol{\Sigma}_k\,\boldsymbol{V}_k^T = \boldsymbol{U}_k\,\boldsymbol{C}_k$$

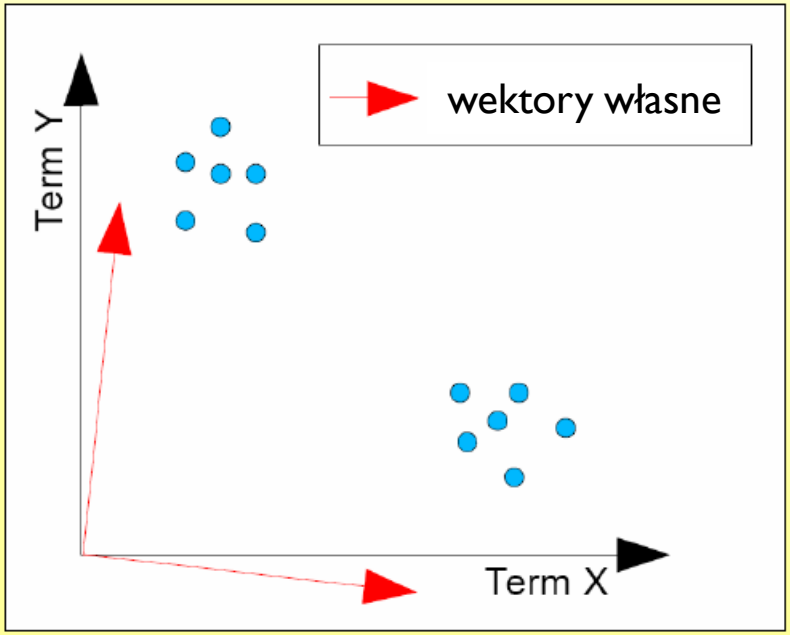
SVD

$$A = \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix} \approx \begin{bmatrix} * & * \\ * & * \\ * & * \\ * & * \\ * & * \\ * & * \end{bmatrix} \times \begin{bmatrix} * & * & * & * \\ * & * & * & * \end{bmatrix}$$

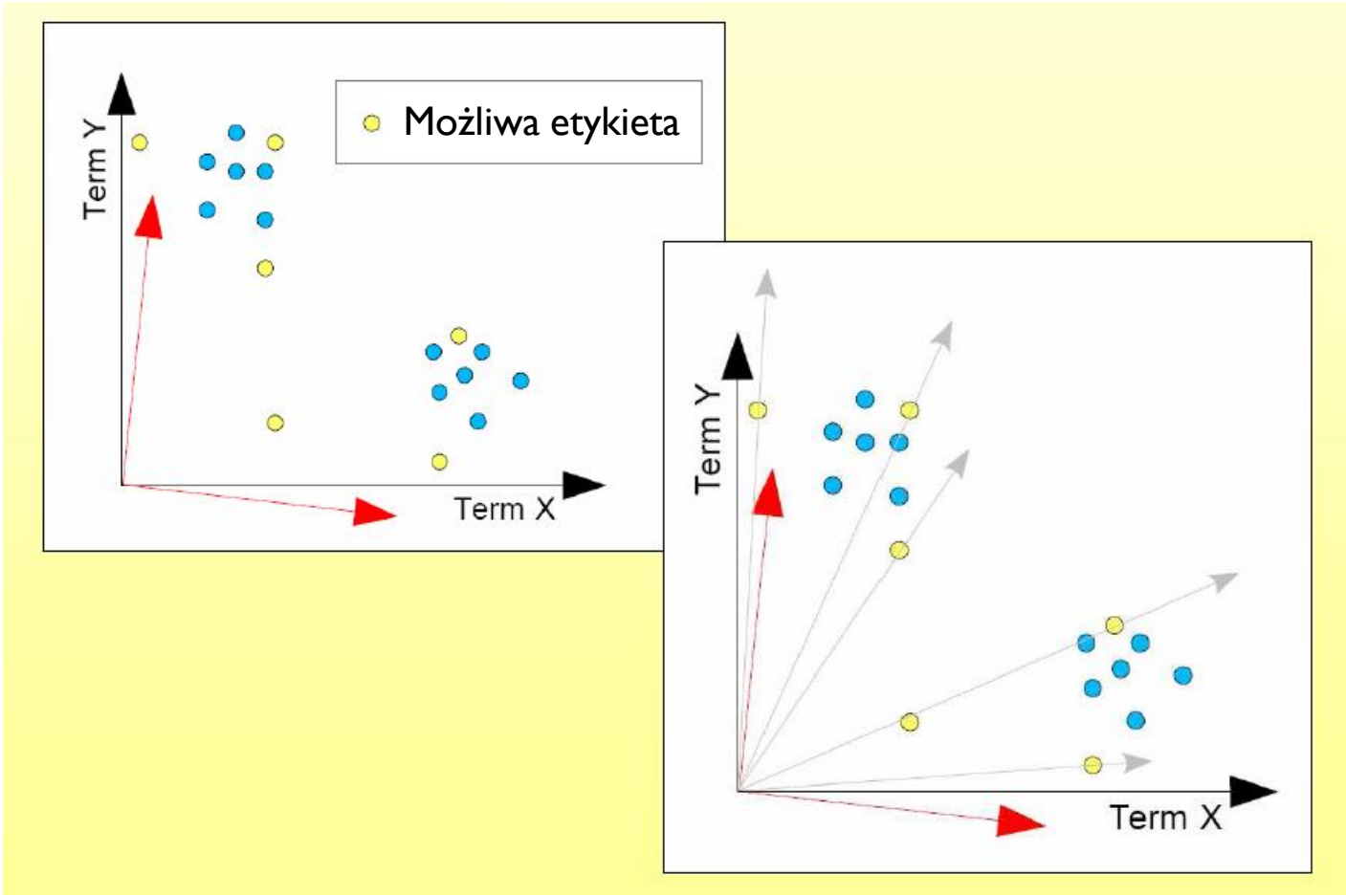
Współrzędne dokumentów w przestrzeni rzutowej

Wektory własne

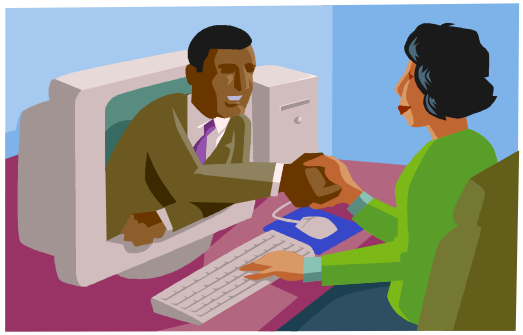
$$A \approx U_k C_k$$



SVD wyznacza etykiety grup



Konkluzje



- SRC – próba przyśpieszania procesu wyszukiwania informacji w internecie i w bibliotekach elektronicznych.
- Temat atrakcyjny również dla dużych graczy
- Problemy:
  - Brak obiektywnego kryterium oceny
  - Brak personalizacji
- Źródła informacji:
  - Historie procesów wyszukiwania w przeszłości
  - Publiczne katalogi internetowe
  - Leksykon semantycznych powiązań, np. Wordnet
  - Profil użytkownika