



O szukaniu sensu w stogu siana

Algorytmy grupowania wyników z wyszukiwarek internetowych
i propozycje ich ulepszenia przy wykorzystaniu wiedzy lingwistycznej.

Dawid Weiss

Instytut Informatyki
Politechnika Poznańska

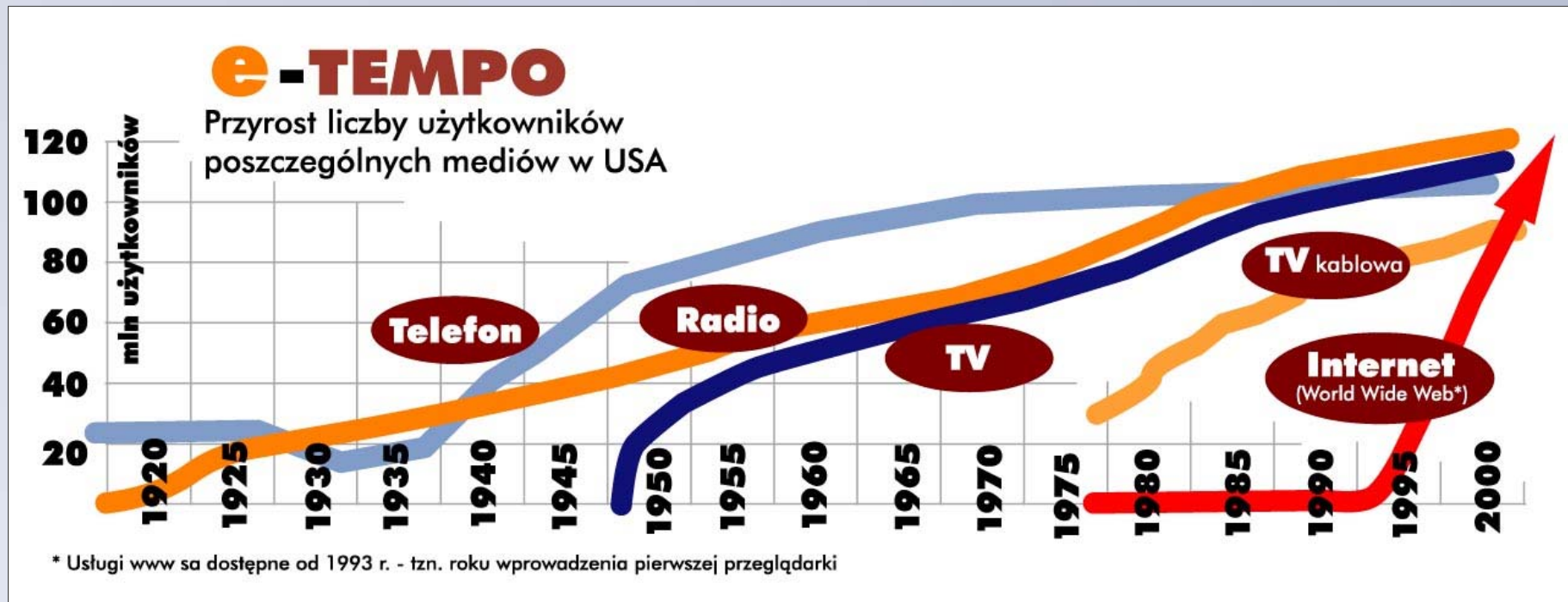
local:8
Carrot
Sort: [flat]
All group
Knowledge (23)
Data Mining
Visual Data
Data Mining
Data Mining
Data Mining
Questi
Data Mining
K-State
and Data
Data Mining
Data Mining
Data Mining
Text Data
Data War
Data Mining (6)
Research
Internatio
Data War
Pattern (4)
TSA Book
Analytical
Data Mining
Experiences
Questions
Data Mining
Petra per
Promotion
other) (1)

Plan prezentacji

- o istocie problemu słów kilka
- algorytmy grupowania wyników wyszukiwania
 - model wektorowy - Term Vector Model
 - model powtarzających się fraz - alg. STC
 - inne modele
 - lingo - rodzimy algorytm, „tej”
- propozycje wykorzystania wiedzy lingwistycznej
- system Carrot²

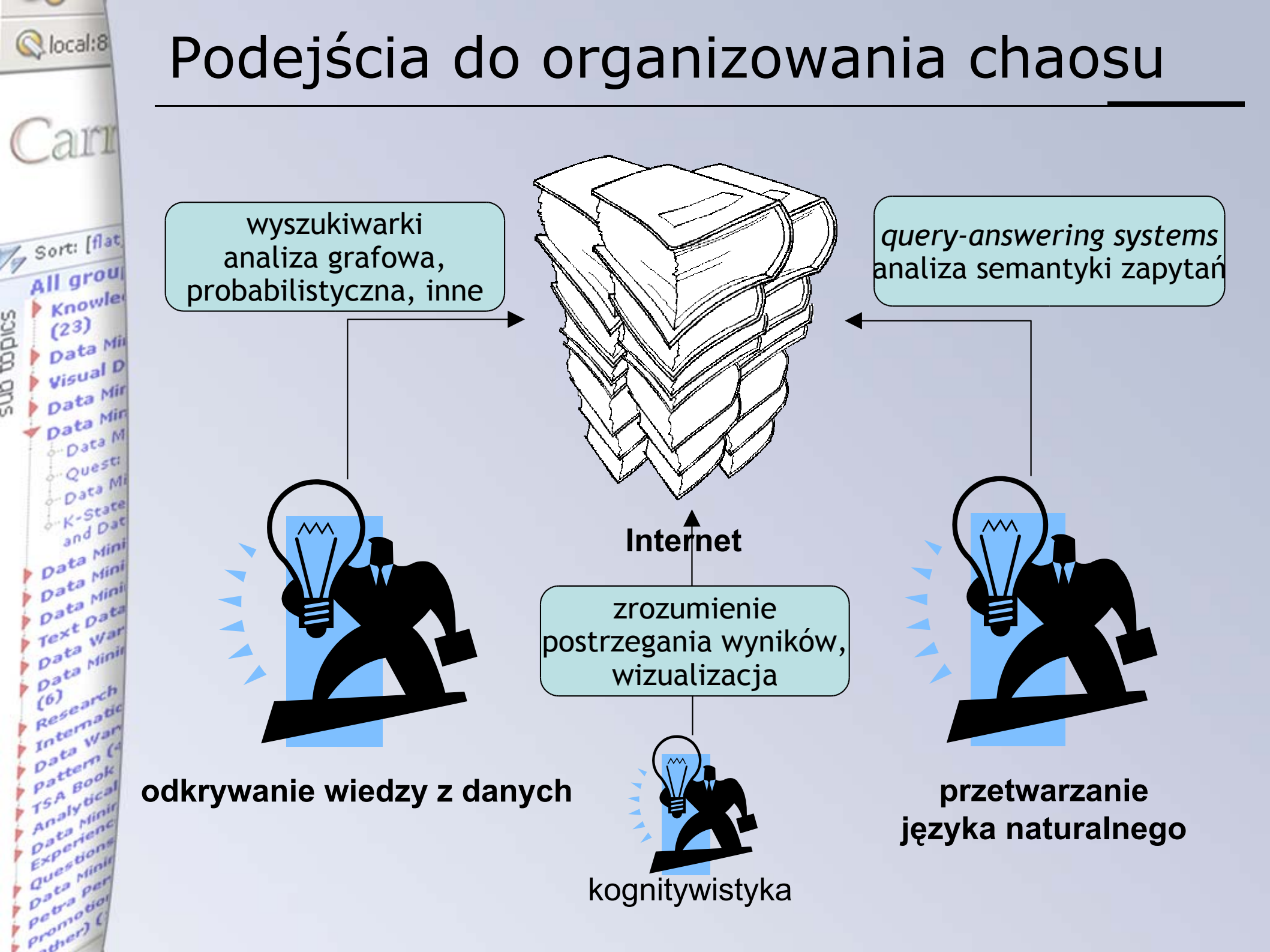
W czym tkwi problem?

- WWW - krótka historia wielkiego wybuchu



Ilość $\neq$ jakość

- większa ilość informacji nie przekłada się na jej jakość
- internet kiedyś był o wiele bardziej *wiarygodny*
- wyszukiwarki nie ułatwiają dostępu do jakościowo lepszej wiedzy
 - wyszukiwarki szukają dokumentów pasujących do *pytań*, a nie odpowiedzi
 - wyszukiwarki zazwyczaj nie tłumaczą struktury zwracanych wyników



Podejścia do organizowania chaosu

wyszukiwarki
analiza grafowa,
probabilistyczna, inne

query-answering systems
analiza semantyki zapytań

Internet

zrozumienie
postrzegania wyników,
wizualizacja

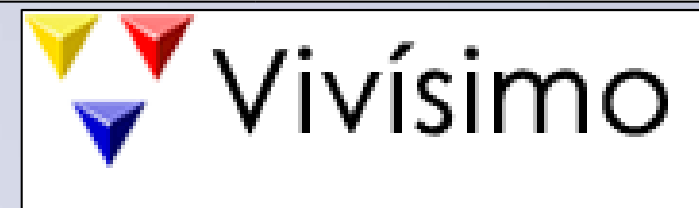
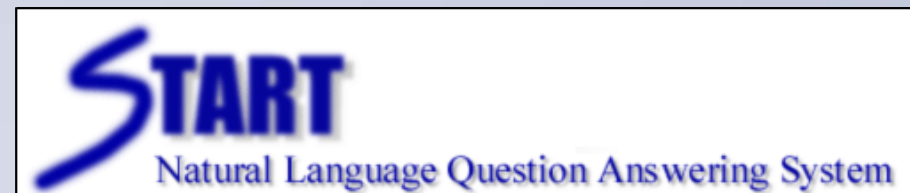
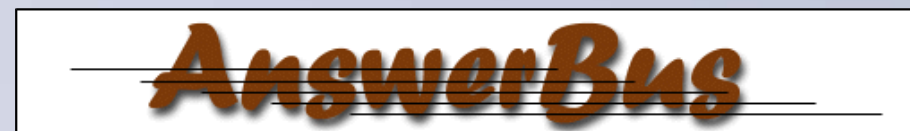
odkrywanie wiedzy z danych

przetwarzanie
języka naturalnego

kognitywistyka

Przykład – zapytanie „apache”

- systemy wyszukujące dokumenty zawierające postawione pytanie
 - AllTheWeb
 - Altavista
 - Google
- systemy odpowiadające na pytania
 - System START
 - System AnswerBus
- systemy organizujące wyniki
 - Vivisimo
 - Carrot²



„apache” w Google

[Zaawansowane szukanie](#)[Ustawienia](#)[Narzędzia językowe](#)[Porady](#)☒ Szukaj w Internecie ☐ Szukaj na stronach kategorii: Polski[WWW](#)[Grafika](#)[Grupy dyskusyjne](#)[Katalog](#)Szukano **apache**.Wyniki **1 - 10** z ok. **23,800,000**. Znalezione w **0.07** sek.Kategoria: [Computers > Software > Internet > Servers > WWW > Apache](#)

[Welcome! - The Apache Software Foundation](#)

The **Apache** Software Foundation. ... Welcome! The **Apache** Software Foundation provides support for the **Apache** community of open-source software projects. ...

Opis: Main site of this top rated internet web server.

Kategoria: [Computers > Software > ... > Projects > Internet > Web > Servers](#)

[www.apache.org/](#) - 11k - [Kopia](#) - [Podobne strony](#)

[Welcome! - The Apache HTTP Server Project](#)

The **Apache** HTTP Server Project. ... HTTP standards. **Apache** has been the most popular web server on the Internet since April of 1996. The ...

Opis: The most popular web server on the net.

Kategoria: [Computers > Software > Internet > Servers > WWW > Apache](#)

[httpd.apache.org/](#) - 8k - [Kopia](#) - [Podobne strony](#)

[Apache HTTP Server Version 1.3 Documentation](#)

FAQ, SiteMap, Directives, Modules, Search. **Apache** HTTP Server Version 1.3. ... Maintained by the **Apache** HTTP Server Documentation Project. **Apache** HTTP Server. Index

[httpd.apache.org/docs/](#) - 10k - [Kopia](#) - [Podobne strony](#)

[[Więcej wyników z httpd.apache.org](#)]

[JAPAN APACHE USERS GROUP](#)

... きたる 8/15~8/17 に開催されるコミックマーケット64にユーザー有志が参加します。詳細は 日本**Apache**ユーザ会(仮 ...

Opis: ウェブサーバー**Apache**の概要、初心者向け情報、マニュアル、FAQ等を掲載。

Kategoria: [World > Japanese > ... > インターネット > サーバー](#)

[www.apache.or.jp/](#) - 10k - 17 Paź 2003 - [Kopia](#) - [Podobne strony](#)

„apache” w AllTheWeb

[advanced search](#) :: [customize preferences](#) :: [submit site](#) :: [help](#)

• • • find it all • • •

Results in: ☐ Any Language ☒ Polish, English

Web | News | Pictures | Video | Audio | FTP files

1 - 10 of 7,406,010 Results for **apache** Offensive content filter: [On](#) - [Off](#)

[Get 128-bit SSL Web Site Encryption!](#) (sponsored)
Looking for SSL? Turn to the SSL experts! Protect your servers with 128-bit SSL encryption from VeriSign. Register for a free Internet security guide today. Click here.
<http://www.verisign.com>

[Apache Stock Report](#) (sponsored)
Get quotes, ratings, analyst opinions, returns, performance charts, valuations, company financials, owners, news and more. All from an independent resource you can trust - Morningstar.
<http://www.Morningstar.com>

[Compare Movie Prices and Shop Online](#) (sponsored)
Smart shoppers, this is the site for deals on movies. Compare prices on thousands of movies, search by title, actor, or director. Find anything from movies to computers.
<http://www.pricegrabber.com>

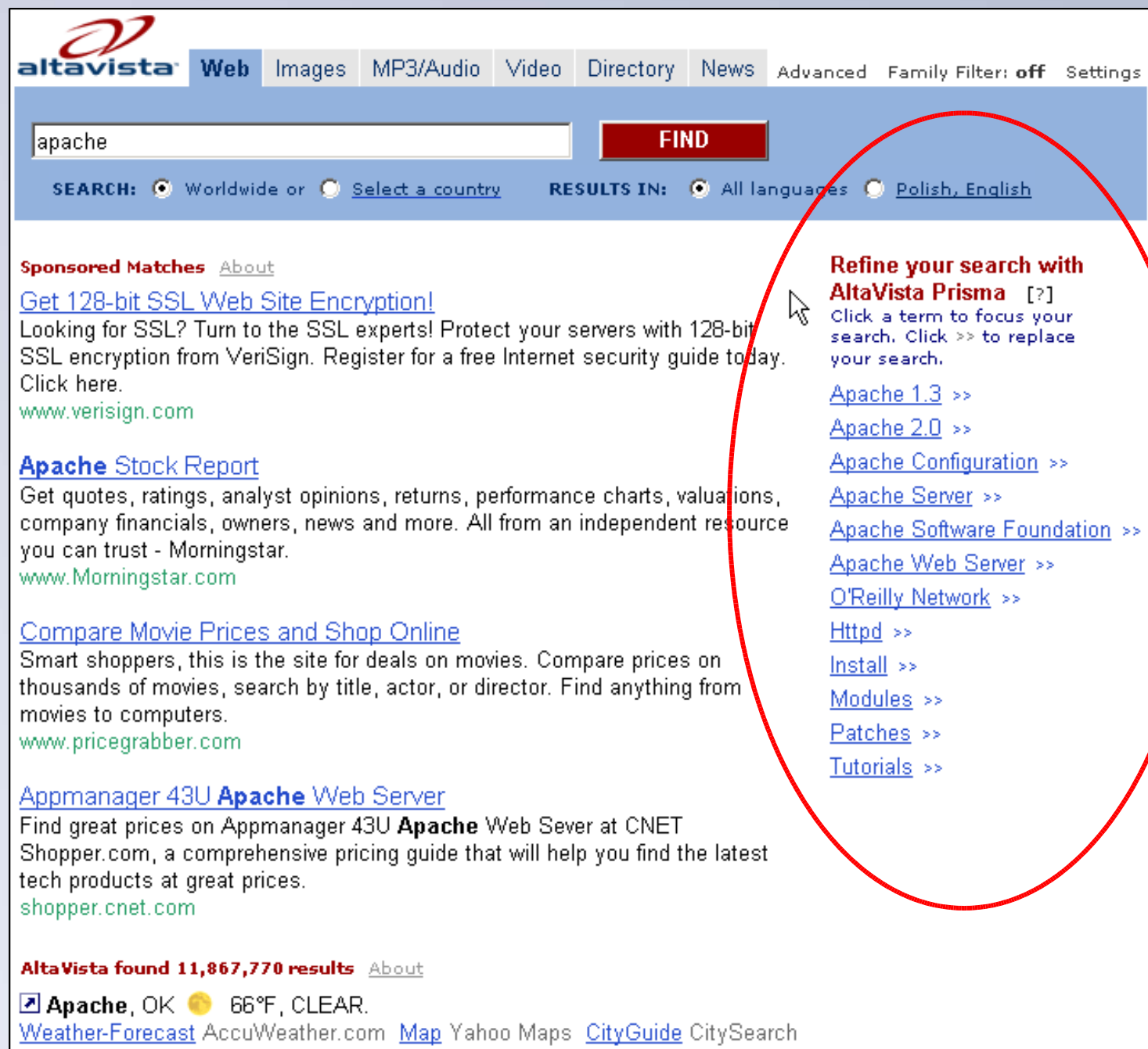
[more results like these](#) »

News : [Test Page for the SSL/TLS-aware Apache ...](#) - [Infosync](#) - Found: 2003-10-16, 9:03 AM
[SOT Linux Advisories: mysql, apache](#) - [Linux Today](#) - Found: 2003-10-15, 11:03 AM
[more news](#) »

[Welcome! - The Apache HTTP Server Project](#)
... mirror Documentation **Apache** 1.3 **Apache** 2.0 Get Involved Mailing ... Number One HTTP Server On The Internet The **Apache** HTTP Server Project is an effort to develop and maintain ...
Description: The most popular web server on the net.
[more hits from:](#) <http://httpd.apache.org/> - 8 KB

[Apache Digital - AMP Resource Site](#)
Apache offers Linux compatible hardware, Linux servers, linux PC's, linux ... current offerings. This is meant for search engines and spiders. **Apache** Digital offers **Apache** training, **Apache** support, **Apache**

„apache“ w Altavista.com



The screenshot shows the Altavista search engine interface. The search bar contains the text 'apache' and a red 'FIND' button. Below the search bar, there are options for 'SEARCH:' (Worldwide or Select a country) and 'RESULTS IN:' (All languages or Polish, English). The search results are displayed in a list format. A red circle highlights the right side of the results, specifically the 'Refine your search with AltaVista Prisma' section and the list of links.

altavista Web Images MP3/Audio Video Directory News Advanced Family Filter: **off** Settings

apache **FIND**

SEARCH: ☒ Worldwide or ☐ Select a country **RESULTS IN:** ☒ All languages ☐ Polish, English

Sponsored Matches [About](#)

[Get 128-bit SSL Web Site Encryption!](#)
Looking for SSL? Turn to the SSL experts! Protect your servers with 128-bit SSL encryption from VeriSign. Register for a free Internet security guide today. Click here.
www.verisign.com

[Apache Stock Report](#)
Get quotes, ratings, analyst opinions, returns, performance charts, valuations, company financials, owners, news and more. All from an independent resource you can trust - Morningstar.
www.Morningstar.com

[Compare Movie Prices and Shop Online](#)
Smart shoppers, this is the site for deals on movies. Compare prices on thousands of movies, search by title, actor, or director. Find anything from movies to computers.
www.pricegrabber.com

[Appmanager 43U Apache Web Server](#)
Find great prices on Appmanager 43U **Apache** Web Sever at CNET Shopper.com, a comprehensive pricing guide that will help you find the latest tech products at great prices.
shopper.cnet.com

AltaVista found 11,867,770 results [About](#)

 **Apache**, OK  66°F, CLEAR.
[Weather-Forecast](#) AccuWeather.com [Map](#) Yahoo Maps [CityGuide](#) CitySearch

Refine your search with AltaVista Prisma [?]
Click a term to focus your search. Click >> to replace your search.

- [Apache 1.3](#) >>
- [Apache 2.0](#) >>
- [Apache Configuration](#) >>
- [Apache Server](#) >>
- [Apache Software Foundation](#) >>
- [Apache Web Server](#) >>
- [O'Reilly Network](#) >>
- [Httpd](#) >>
- [Install](#) >>
- [Modules](#) >>
- [Patches](#) >>
- [Tutorials](#) >>

„apache“ w systemie START

START's reply

==> What is apache?

Apache is a freely available Web [server](#) that is distributed under an "open source" license. Version 2.0 runs on most [UNIX](#)-based operating systems (such as [Linux](#), [Solaris](#), Digital UNIX, and [AIX](#)), on other UNIX/[POSIX](#)-derived systems (such as Rhapsody, [BeOS](#), and BS2000/OSD), on AmigaOS, and on [Windows 2000](#). According to the Netcraft (www.netcraft.com) Web server survey in February, 2001, 60% of all Web sites on the Internet are using Apache (62% including Apache derivatives), making Apache more widely used than all other Web servers combined.

Apache complies with the newest level of the Hypertext Transport Protocol, [HTTP 1.1](#). Free support is provided through a bug reporting system and several [Usenet](#) newsgroups. Several companies offer priced support.

Read more about it:

- >> [SearchWebManagement.com](#) provides more information about "Apache Management."
- >> [You can download Apache, read a FAQ, and find other information at the Apache Web site.](#)

Source: whatis.com

Which movie in [The Internet Movie Database](#) do you mean:

- [Apache, The \(1925\)](#)
- [Apache, The \(1928\)](#)
- [Apache \(1954\)](#)

[Apache](#) is a city located in [the USA](#), in the state of [Oklahoma](#).

Source: START KB

„apache“ w AnswerBus

AnswerBus

Type in your question in English, French, Spanish, German, Italian or Portuguese.

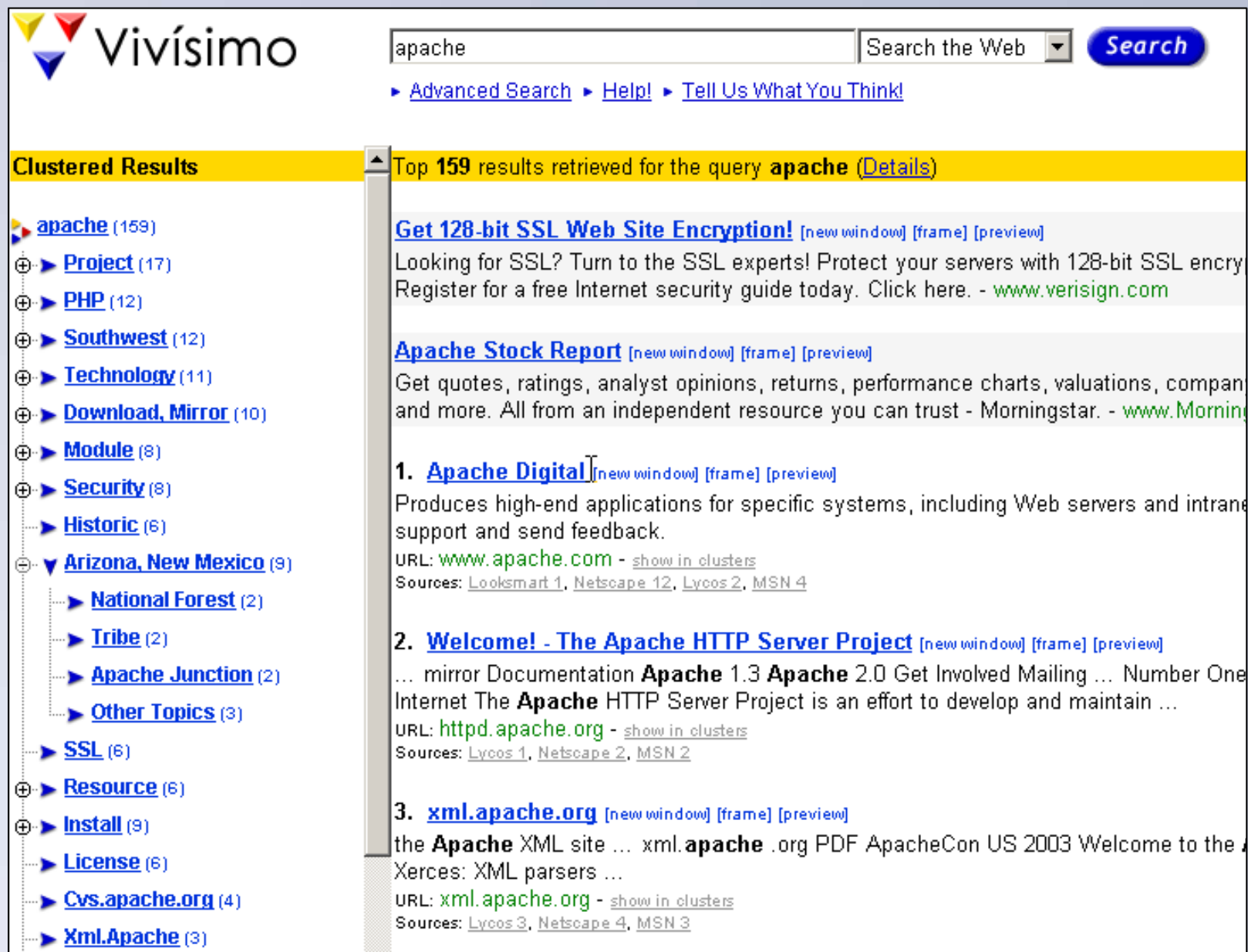
Question:

what is apache?

Possible answers: [XML](#) [TXT](#)

- [AxKit brings the power of XML web publishing and dynamic XML based applications to the Apache web server using the mod_perl framework and the Apache C API to combine the power of XML with the performance of a native Apache solution.](#)
- [The Apache Cultural Museum, located on the 288 acre site at Fort Apache, provides insight into the culture of the Apache people.](#)
- [The Apache are composed of six regional groups: the Western Apache, Chiricahua, Mescalero, Jicarilla, Lipan, and Kiowa Apache.](#)
- [Throttles Apache; mod_security Drives Apache 2.2 with AAA; Apache Deflates; Triple Book Reviews](#)
- [The Apache Group is pleased to announce the release of the 1.3.27 version of the Apache HTTP Server.](#)
- [Credits Apache-SSL was written by Ben Laurie, who is also an Apache core team and board member, and an OpenSSL core team member.](#)
- [The Apache HTTP Server is a project of the Apache Software Foundation.](#)
- [The Apache Office of Tourism, is located in General Crook's cabin \(built in 1870\), at Fort Apache.](#)
- [Commercial support Commercial support is available for both Apache and Apache-SSL.](#)
- [The Apache HTTP Server Project is proud to announce the eighth public release of Apache 2.0.](#)

„apache“ w Vivísimo.com



The screenshot shows the Vivísimo search engine interface. At the top, the Vivísimo logo is on the left, and a search bar with the text 'apache' is on the right. To the right of the search bar is a dropdown menu labeled 'Search the Web' and a blue 'Search' button. Below the search bar are links for 'Advanced Search', 'Help!', and 'Tell Us What You Think!'. The main content area is divided into two columns. The left column, titled 'Clustered Results', lists various categories and sub-topics related to 'apache', such as 'Project', 'PHP', 'Southwest', 'Technology', 'Download, Mirror', 'Module', 'Security', 'Historic', 'Arizona, New Mexico', 'National Forest', 'Tribe', 'Apache Junction', 'Other Topics', 'SSL', 'Resource', 'Install', 'License', 'Cvs.apache.org', and 'XmlApache'. The right column, titled 'Top 159 results retrieved for the query apache (Details)', displays three search results. The first result is 'Get 128-bit SSL Web Site Encryption!' from verisign.com. The second result is 'Apache Stock Report' from Morningstar. The third result is '1. Apache Digital' from www.apache.com. Each result includes a brief description, a URL, and a list of sources.

Vivísimo

apache Search the Web Search

Advanced Search Help! Tell Us What You Think!

Clustered Results

Top 159 results retrieved for the query **apache** (Details)

apache (159)

- Project (17)
- PHP (12)
- Southwest (12)
- Technology (11)
- Download, Mirror (10)
- Module (8)
- Security (8)
- Historic (6)
- Arizona, New Mexico (9)
 - National Forest (2)
 - Tribe (2)
 - Apache Junction (2)
 - Other Topics (3)
- SSL (6)
- Resource (6)
- Install (9)
- License (6)
- Cvs.apache.org (4)
- XmlApache (3)

Get 128-bit SSL Web Site Encryption! [new window] [frame] [preview]
Looking for SSL? Turn to the SSL experts! Protect your servers with 128-bit SSL encryption. Register for a free Internet security guide today. Click here. - www.verisign.com

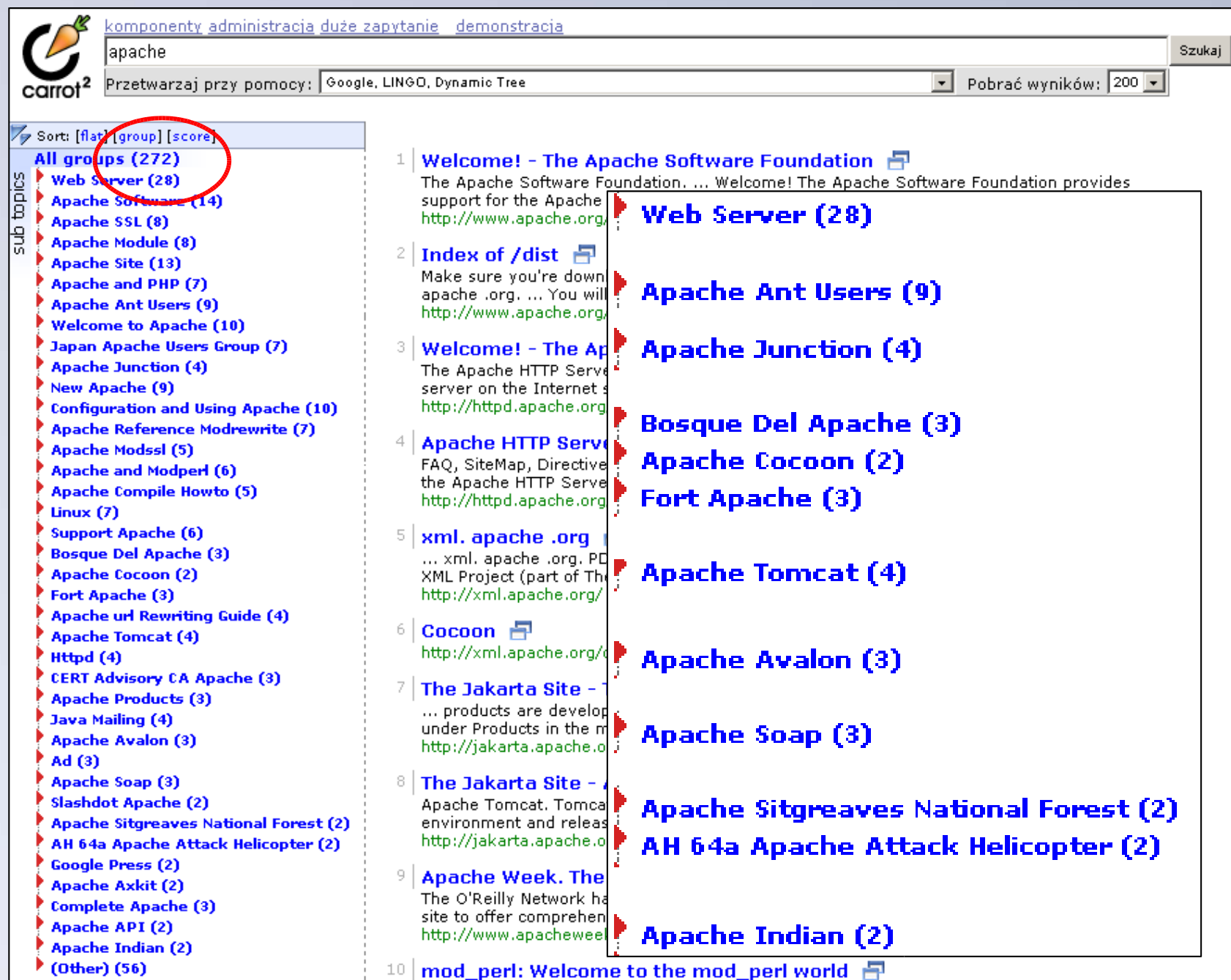
Apache Stock Report [new window] [frame] [preview]
Get quotes, ratings, analyst opinions, returns, performance charts, valuations, company information and more. All from an independent resource you can trust - Morningstar. - www.Morningstar.com

1. Apache Digital [new window] [frame] [preview]
Produces high-end applications for specific systems, including Web servers and intranet support and send feedback.
URL: www.apache.com - show in clusters
Sources: Looksmart 1, Netscape 12, Lycos 2, MSN 4

2. Welcome! - The Apache HTTP Server Project [new window] [frame] [preview]
... mirror Documentation **Apache** 1.3 **Apache** 2.0 Get Involved Mailing ... Number One Internet The **Apache** HTTP Server Project is an effort to develop and maintain ...
URL: httpd.apache.org - show in clusters
Sources: Lycos 1, Netscape 2, MSN 2

3. xml.apache.org [new window] [frame] [preview]
the **Apache** XML site ... **xml.apache.org** PDF ApacheCon US 2003 Welcome to the Apache Xerces: XML parsers ...
URL: xml.apache.org - show in clusters
Sources: Lycos 3, Netscape 4, MSN 3

„apache” w systemie Carrot²



The screenshot displays the Carrot2 search engine interface. At the top, the Carrot2 logo is visible on the left, and navigation links (komponenty, administracja, duże zapytanie, demonstracja) are on the right. The search bar contains the term 'apache'. Below the search bar, a dropdown menu shows 'Przetwarzaj przy pomocy: Google, LINGO, Dynamic Tree'. To the right of the search bar is a 'Szukaj' button and a 'Pobrać wyników: 200' dropdown.

On the left side, there is a 'sub topics' sidebar with a 'Sort: [flat] [group] [score]' dropdown. The 'All groups (272)' section is expanded, showing a list of sub-topics. The 'Web Server (28)' sub-topic is circled in red.

The main search results area on the right lists 10 items. The first 9 items are grouped under the 'Web Server (28)' sub-topic, which is also circled in red. The 10th item, 'mod_perl: Welcome to the mod_perl world', is not part of this group.

Rank	Result	Sub-topic
1	Welcome! - The Apache Software Foundation The Apache Software Foundation. ... Welcome! The Apache Software Foundation provides support for the Apache ... http://www.apache.org/	Web Server (28)
2	Index of /dist Make sure you're down apache .org. ... You will ... http://www.apache.org/	
3	Welcome! - The Apache HTTP Server The Apache HTTP Server on the Internet ... http://httpd.apache.org/	
4	Apache HTTP Server FAQ, SiteMap, Directive ... http://httpd.apache.org/	
5	xml.apache.org ... xml.apache.org. PD ... http://xml.apache.org/	
6	Cocoon http://xml.apache.org/	
7	The Jakarta Site - ... products are develop ... http://jakarta.apache.org/	
8	The Jakarta Site - Apache Tomcat. Tomcat ... http://jakarta.apache.org/	
9	Apache Week. The The O'Reilly Network ha ... http://www.apacheweek.org/	
10	mod_perl: Welcome to the mod_perl world	

Grupowanie w AllTheWeb?

Apache Corporation - Oil and Gas Exploration and Development

... **Apache** Corporation is ... **Apache** makes second Exmouth ... **Apac**

Description: **Apache** Corporation - Oil Exploration & Development, Natu

<http://www.apachecorp.com/> - 26 KB

Result Page: [1](#) [2](#) [3](#) [4](#) [5](#) [6](#) [7](#) [8](#) [9](#) [10](#) [Next](#) »

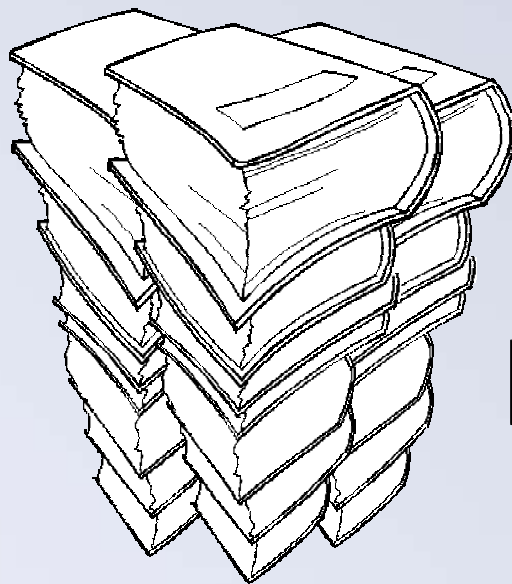
Clusters: [apache, www, servers \(17\)](#) 
[modules, apache, www \(9\)](#)
[apache, native americans \(6\)](#)
[development tools, mod perl, modules \(5\)](#)
[apache junction, arizona, us \(4\)](#)
[ssi, internet, programming \(3\)](#)
[apache, www, servers \(3\)](#) 
[applications, mod perl, modules \(3\)](#)
[application servers, java, languages \(3\)](#)
[guides, apache, www \(3\)](#)
[business, apache junction, arizona \(3\)](#)
[« fewer clusters](#)

See our [last 10 queries](#)

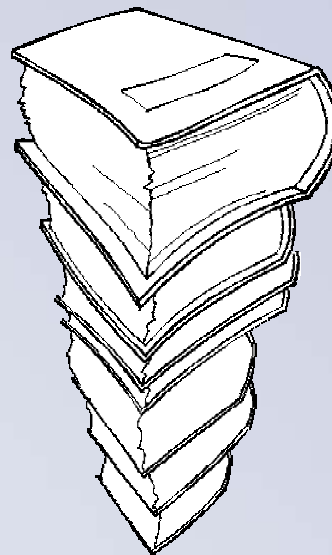
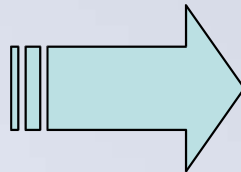
[Try your query on AltaVista](#)

Definicja problemu

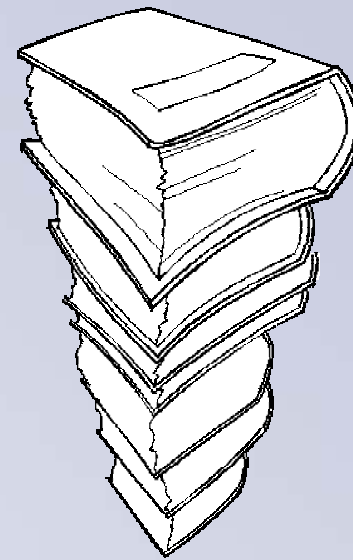
Search Results Clustering polega na efektywnym utworzeniu sensownych grup tematycznie powiązanych dokumentów, oraz ich zwięzłym opisaniu w sposób zrozumiały dla człowieka.



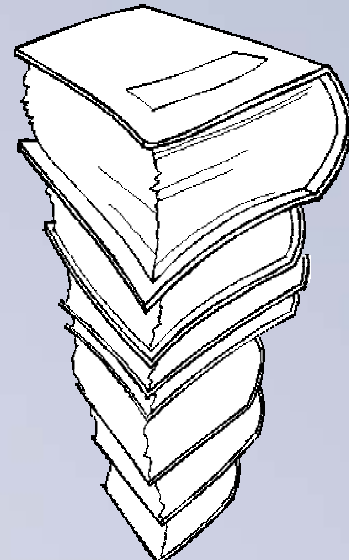
Apache



serwer WWW



helikopter



indianie



Problem nie jest trywialny...

- nie jest znana liczba oczekiwanych grup
- miara podobieństwa dokumentów jest trudna do zdefiniowania
- grupy mogą się nakładać
- znalezienie opisu dla grup nie jest proste
- wymagana szybkość wykonywania (on-line)
- dokumenty mogą być wielojęzyczne
- opisy są zazwyczaj krótkie (*snippets*) i niepełne

The logo for Carrot2 is a vertical strip on the left side of the slide. It features a yellow background with a red border. At the top, it says 'local:8'. Below that, the word 'Carrot' is written in a large, stylized font. Further down, there is a list of sub-topics including 'All group', 'Knowledge (23)', 'Data Mining', 'Visual Data Mining', 'Data Mining', 'Data Mining', 'Quest', 'Data Mining', 'K-State and Data', 'Data Mining', 'Data Mining', 'Data Mining', 'Text Data', 'Data War', 'Data Mining (6)', 'Research', 'Internation', 'Data War', 'Pattern (4)', 'TSA Book', 'Analytical', 'Data Mining', 'Experiences', 'Questions', 'Data Mining', 'Petra per', and 'Promotion (other)'.

Plan prezentacji

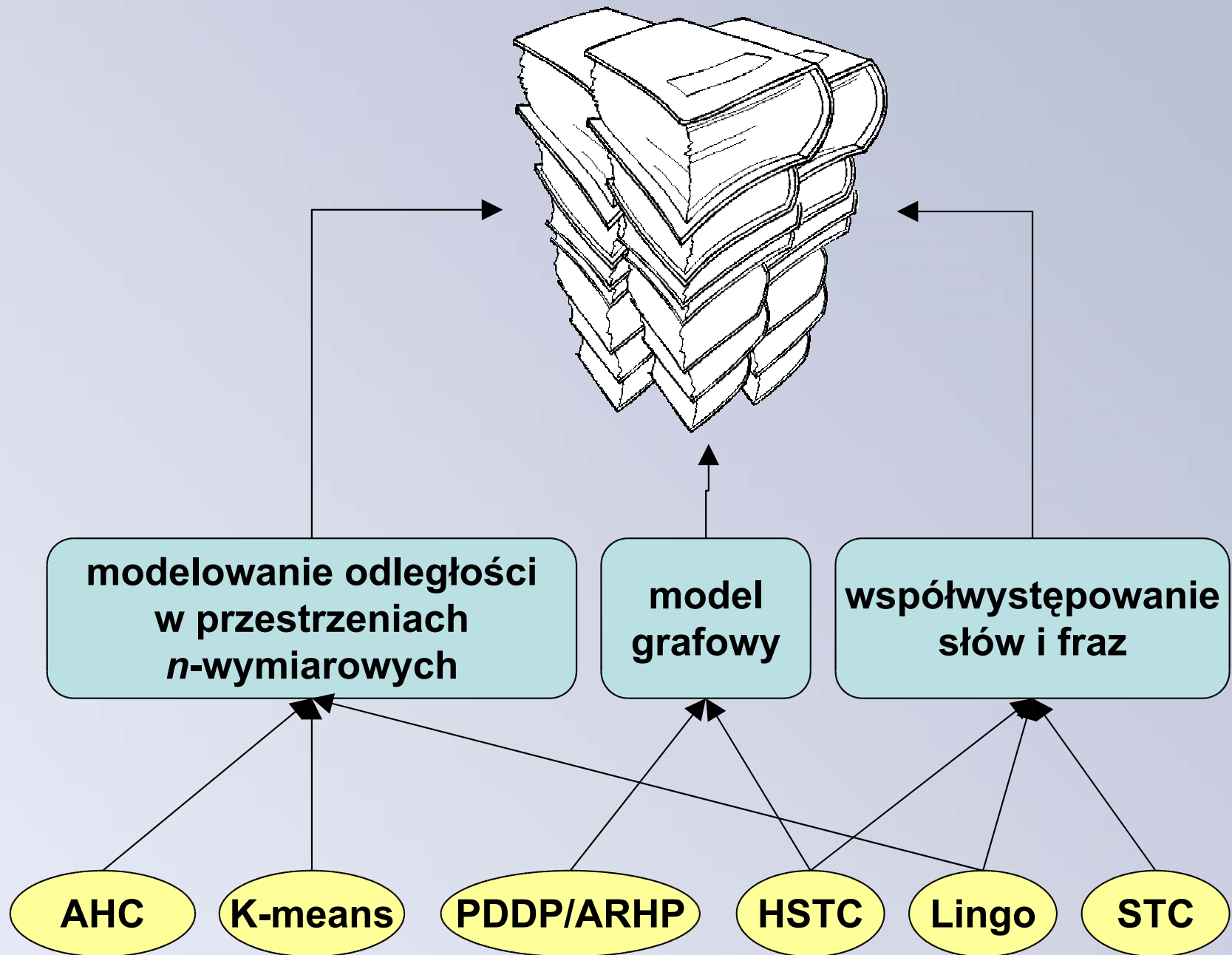
- o istocie problemu słów kilka
- **algorytmy grupowania wyników wyszukiwania**
 - model wektorowy - Term Vector Model
 - model powtarzających się fraz - alg. STC
 - inne modele
 - lingo - rodzimy algorytm, „tej”
- propozycje wykorzystania wiedzy lingwistycznej
- demonstracja systemu Carrot²



All the real knowledge which we possess, depends on methods by which we distinguish the similar from the dissimilar.

- Genera plantarum, Linnaeus

Modelowanie podobieństwa



Vector Space Model

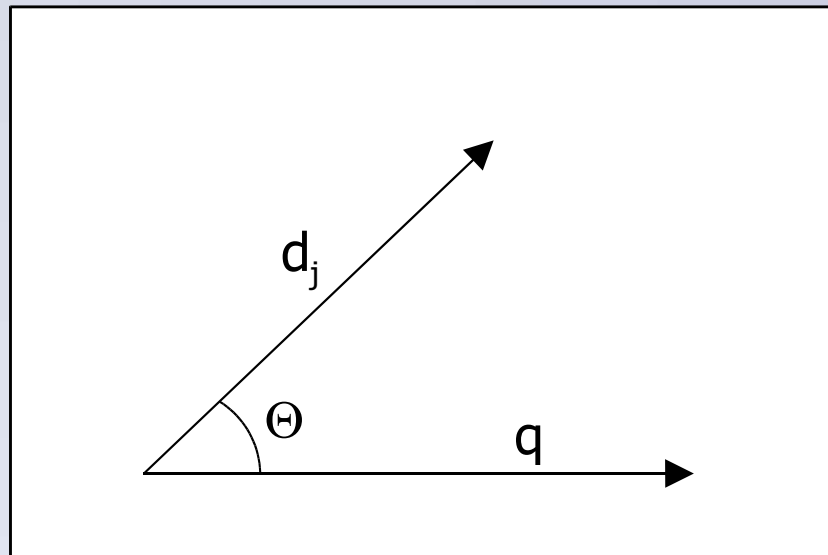
- model oparty na algebrze liniowej
 - zbiór unikalnych słów $T=t_1, t_2, \dots, t_n$
 - dokumenty ($D=d_1, d_2, \dots, d_m$) reprezentowane jako n -wymiarowe wektory
 - $d_i=[w_{i1}, w_{i2}, \dots, w_{in}]$, gdzie w_{ij} jest wagą j -tego słowa w dokumencie i

Ważenie słów

- wagi słów - jak dane słowo jest charakterystyczne dla dokumentu?
 - wiele różnych form:
 - binarna - $w_{ij}=1$ lub $w_{ij}=0$
 - częstość wystąpień - $w_{ij}=tf_{ij}(t_j)$
 - *tfidf* (Salton) - $w_{ij}=tf_{ij}(t_j)*\log(N/df_{ij}(t_j))$
 - rola lematyzacji w procesie obliczania wag
 - rola *stop words* w procesie obliczania wag
- macierz *term-document* (A)
 - kolumny - dokumenty
 - wiersze - słowa

Pojęcie bliskości w macierzy A

- jesteśmy zainteresowani *kątem* jaki tworzą między sobą wektory dokumentów
 - Identyeczny kąt $\rightarrow$ dokumenty są złożone z identycznych słów $\rightarrow$ dokumenty są podobne



Przykład – macierz A

The $t=5$ terms:

- T1: Information
- T2: Singular
- T3: Value
- T4: Computations
- T5: Retrieval

The $d=7$ documents:

- D1: Large Scale Singular Value Computations
- D2: Software for the Sparse Singular Value Decomposition
- D3: Introduction to Modern Information Retrieval
- D4: Linear Algebra for Intelligent Information Retrieval
- D5: Matrix Computations
- D6: Singular Value Analysis of Cryptograms
- D7: Automatic Information Organization

$$A = \begin{pmatrix} 0.00 & 0.00 & 0.71 & 0.71 & 0.00 & 0.00 & 1.00 \\ 0.58 & 0.71 & 0.00 & 0.00 & 0.00 & 0.71 & 0.00 \\ 0.58 & 0.71 & 0.00 & 0.00 & 0.00 & 0.71 & 0.00 \\ 0.58 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.71 & 0.71 & 0.00 & 0.00 & 0.00 \end{pmatrix}$$

Przykład, c.d.

The query: *Singular Value*

$$\hat{q} = (0.00 \quad 0.71 \quad 0.71 \quad 0.00 \quad 0.00)^T$$

$$r = \begin{pmatrix} 0.00 \\ 0.71 \\ 0.71 \\ 0.00 \\ 0.00 \end{pmatrix}^T \begin{pmatrix} 0.00 & 0.00 & 0.71 & 0.71 & 0.00 & 0.00 & 1.00 \\ 0.58 & 0.71 & 0.00 & 0.00 & 0.00 & 0.71 & 0.00 \\ 0.58 & 0.71 & 0.00 & 0.00 & 0.00 & 0.71 & 0.00 \\ 0.58 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.71 & 0.71 & 0.00 & 0.00 & 0.00 \end{pmatrix} = \begin{pmatrix} 0.81 \\ 1.00 \\ 0.00 \\ 0.00 \\ 0.00 \\ 1.00 \\ 0.00 \end{pmatrix}^T$$

The **matched documents**:

D1: Large Scale Singular Value Computations

D2: Software Library for the Sparse Singular Value Decomposition

D6: Singular Value Analysis of Cryptograms

Algorytmy grupowania a macierz A

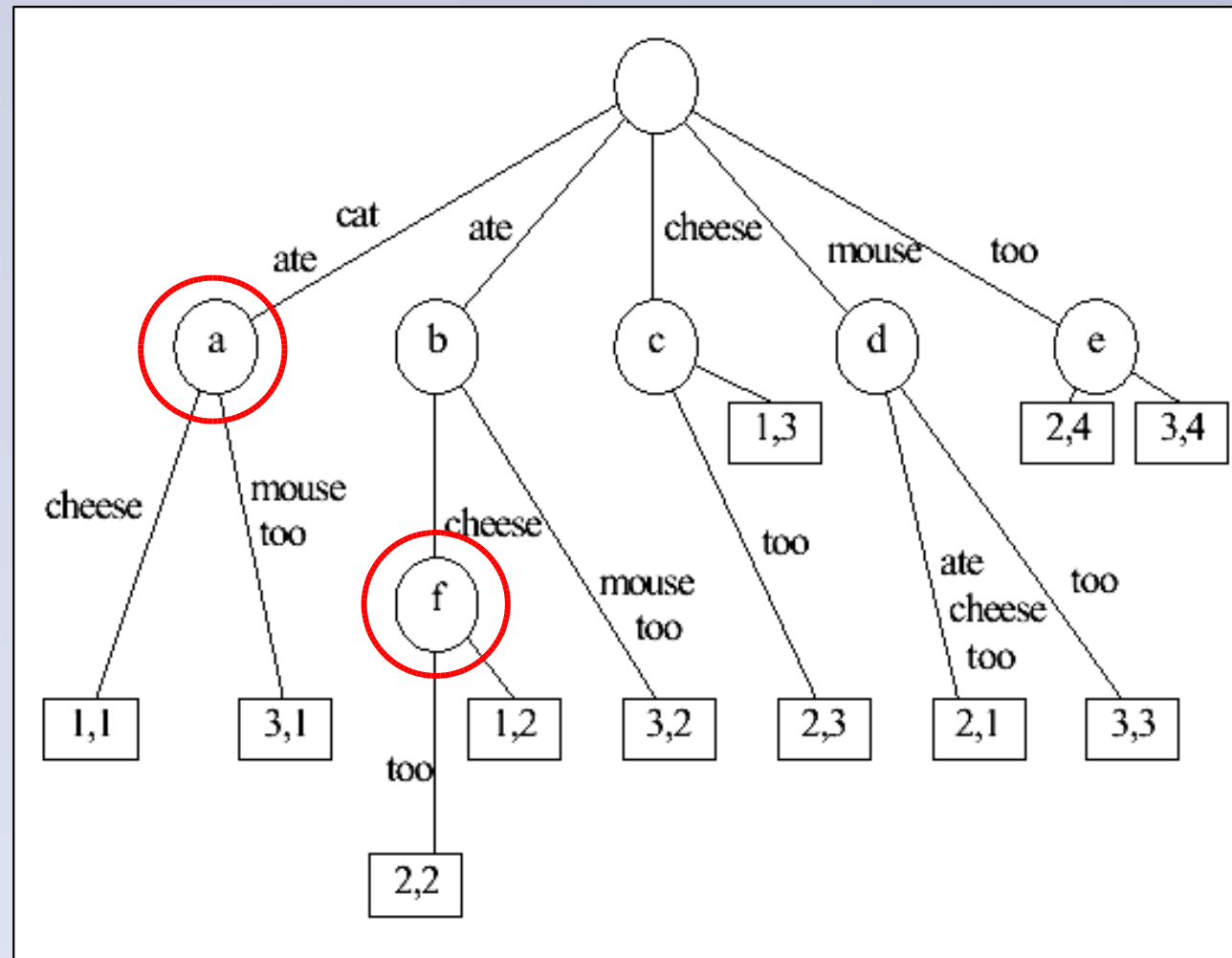
- wykorzystanie informacji o bliskości dokumentów w A
 - zastosowanie mają wszelkie metody analizy skupień w danych numerycznych
- problemy
 - grupy zazwyczaj sferyczne
 - każde słowo jest traktowane oddzielnie
 - problemy ze znalezieniem opisu grup
 - problem z naturalnym kryterium stopu dla większości algorytmów numerycznych

Algorytm STC – wykorzystanie fraz

- Suffix Tree Clustering, Oren Zamir, O. Etzioni
 - fraza = sekwencja występujących po sobie słów
 - algorytm rozważa wspólne podfrazy występujące w dokumentach

Przykład - STC

Drzewo
sufiksów



frazy
częste

Wejściowe „dokumenty” :

- (1) “cat ate cheese”
- (2) “mouse ate cheese too”
- (3) “cat ate mouse too”

Grupy:

- [a] (1,3) cat ate
(cat’s culinary habits)
- [f] (1,2) ate cheese
(cheese dining preference)



STC w praktyce

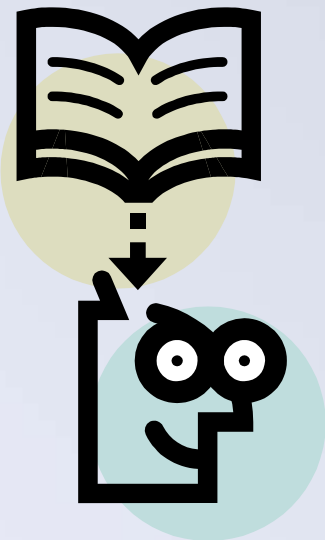
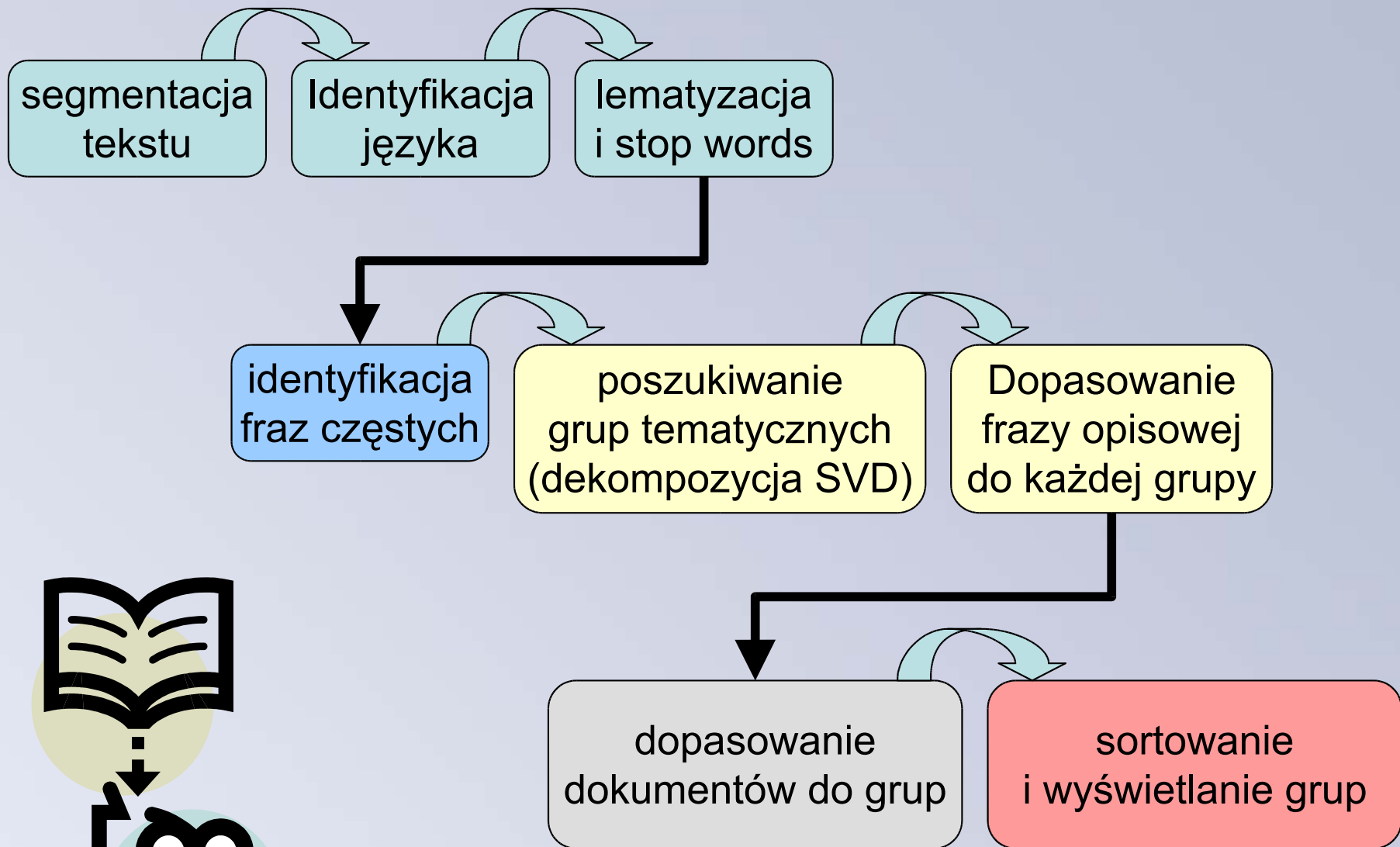
- zalety
 - brak numerycznej miary odległości
 - frazy stanowią zazwyczaj dobre opisy grup
 - liniowa złożoność - $O(N)$
- wady
 - słabo radzi sobie z szumem (*home page*)
 - problemy z separacją małych grup
 - wrażliwość na progi i język dokumentów



Algorytm Lingo

- Label INduction Grouping Algorithm
- autor: Stanisław Osiński, Politechnika Poznańska, uczestnik projektu Carrot²
- odwrócony proces grupowania - najpierw opisy, później dokumenty (jak w Vivisimo)
- opiera się na metodzie rozkładu macierzy A przy pomocy SVD
 - redukcja szumów
 - ujawnia ukryte związki między dokumentami
- wykorzystuje frazy częste aby opisać odnalezione grupy

Lingo - kroki algorytmu



Dekompozycja SVD

- pomijając szczegóły matematyczne...
 - kolumny macierzy U tworzą ortogonalną bazę w przestrzeni kolumn macierzy A
 - wektory te wykazują podobieństwo do „tematów” obecnych w A

$$\underbrace{\begin{pmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{pmatrix}}_A = \underbrace{\begin{pmatrix} * & * & * \\ * & * & * \\ * & * & * \end{pmatrix}}_U \underbrace{\begin{pmatrix} \bullet & & \\ & \bullet & \\ & & \bullet \end{pmatrix}}_{\Sigma} \underbrace{\begin{pmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{pmatrix}}_{V^T}$$

Dopasowanie fraz do tematów

- otrzymujemy poprzez wyliczenie kątów między frazami (kandydatami) a podmacierzą U
- jest to równoważne szukaniu *zapytań*, do których pasują pewne zbiory wyników

Przykład - Lingo

The $t=5$ terms:

T1: Information
T2: Singular
T3: Value
T4: Computations
T5: Retrieval

The $d=7$ documents:

D1: Large Scale Singular Value Computations
D2: Software for the Sparse Singular Value Decomposition
D3: Introduction to Modern Information Retrieval
D4: Linear Algebra for Intelligent Information Retrieval
D5: Matrix Computations
D6: Singular Value Analysis of Cryptograms
D7: Automatic Information Organization

The $p=2$ phrases:

P1: Singular Value
P2: Information Retrieval

$$U = \begin{pmatrix} 0.00 & 0.75 & 0.00 & -0.66 & 0.00 \\ 0.65 & 0.00 & -0.28 & 0.00 & -0.71 \\ 0.65 & 0.00 & -0.28 & 0.00 & 0.71 \\ 0.39 & 0.00 & 0.92 & 0.00 & 0.00 \\ 0.00 & 0.66 & 0.00 & 0.75 & 0.00 \end{pmatrix} \quad P = \begin{pmatrix} 0.00 & 0.56 & 1.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.71 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 0.00 \\ 0.71 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.83 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{pmatrix}$$

Zakładamy, że $r_A = 2$ (tzn, pod uwagę brane są dwa pierwsze wektory kolumnowe macierzy U).

Przykład – Lingo, c.d.

The $t=5$ terms:

T1: Information
T2: Singular
T3: Value
T4: Computations
T5: Retrieval

The $d=7$ documents:

D1: Large Scale Singular Value Computations
D2: Software for the Sparse Singular Value Decomposition
D3: Introduction to Modern Information Retrieval
D4: Linear Algebra for Intelligent Information Retrieval
D5: Matrix Computations
D6: Singular Value Analysis of Cryptograms
D7: Automatic Information Organization

The $p=2$ phrases:

P1: Singular Value
P2: Information Retrieval

$$M = U_k^T P = \begin{pmatrix} \mathbf{0.92} & 0.00 & 0.00 & 0.65 & 0.65 & 0.39 & 0.00 \\ 0.00 & \mathbf{0.97} & 0.75 & 0.00 & 0.00 & 0.00 & 0.66 \end{pmatrix}$$

The $e=2$ candidate cluster labels:

E1: Singular Value (score: 0.92)
E2: Information Retrieval (score: 0.97)

Dopasowanie dokumentów do grup

- Ponownie „odpytujemy” zbiór dokumentów, stosując zbiór opisów grup jako zapytania
- Używany jest ponownie Vector Space Model
- W fazie końcowej grupy są sortowane wg wzoru: $score = label_score * cluster_size$

Przykład – Lingo, c.d.

$$Q = \begin{pmatrix} 0.00 & 0.56 \\ 0.71 & 0.00 \\ 0.71 & 0.00 \\ 0.00 & 0.00 \\ 0.00 & 0.83 \end{pmatrix}$$

$$C = Q^T A = \begin{pmatrix} 0.69 & 1.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 & 1.00 & 0.00 & 0.00 & 0.56 \end{pmatrix}$$

Information Retrieval [1.0]

D3: Introduction to Modern Information Retrieval [1.00]

D4: Linear Algebra for Intelligent Information Retrieval [1.00]

D7: Automatic Information Organization [0.56]

Singular Value [0.95]

D2: Software for the Sparse Singular Value Decomposition [1.00]

D6: Singular Value Analysis of Cryptograms [1.00]

D1: Large Scale Singular Value Computations [0.69]

Other Topics

D5: Matrix Computations



LINGO – podsumowanie

- zalety
 - czytelne opisy grup
 - różnicowanie tematyczne grup (dzięki użyciu SVD)
 - dokument może należeć do więcej niż jednej grupy
- niedociągnięcia
 - płaska struktura grup
 - spore wymagania zasobowe (SVD)
 - brak możliwości obliczeń przyrostowych



Plan prezentacji

- o istocie problemu słów kilka
- algorytmy grupowania wyników wyszukiwania
 - model wektorowy - Term Vector Model
 - model powtarzających się fraz - alg. STC
 - inne modele
 - lingo - rodzimy algorytm, „tej”
- propozycje wykorzystania wiedzy lingwistycznej
- demonstracja systemu Carrot²



Motywacja

- metody frekwencyjne zawsze będą omyłne
- wydaje się, że nawet pobieżna analiza lingwistyczna może przynieść wymierną poprawę wyników



Możliwości

- dokładna lematyzacja
 - obecnie
 - j. polski - „lametyzator” własnej produkcji
 - j. angielski - alg. Portera
- przykłady obecnie błędnie interpretowane
 - j. polski
 - beton - betoniarka
 - komin - kominiarz - kominiarka
 - j. angielski
 - ‘new’ - ‘news’

Możliwości

- usuwanie (ang. *pruning*) nieczytelnych opisów
 - ograniczenie się tylko do niektórych części mowy
 - co z nazwami własnymi?
- dopuszczanie jedynie pewnych wzorców (pobieżna analiza składniowa)
- Przykład:
 - „[zawody w] programowaniu zespołowym”
 - „jak”
 - zarówno rzeczownik, jak i zaimek



Możliwości

- wykrywanie nazw własnych
 - zazwyczaj tworzą dobre opisy grup
 - realne raczej jedynie proste heurystyki
- problem zmiennego szyku zdania w j. polskim
 - potrzeba detekcji fraz nie na podstawie ich sekwencji

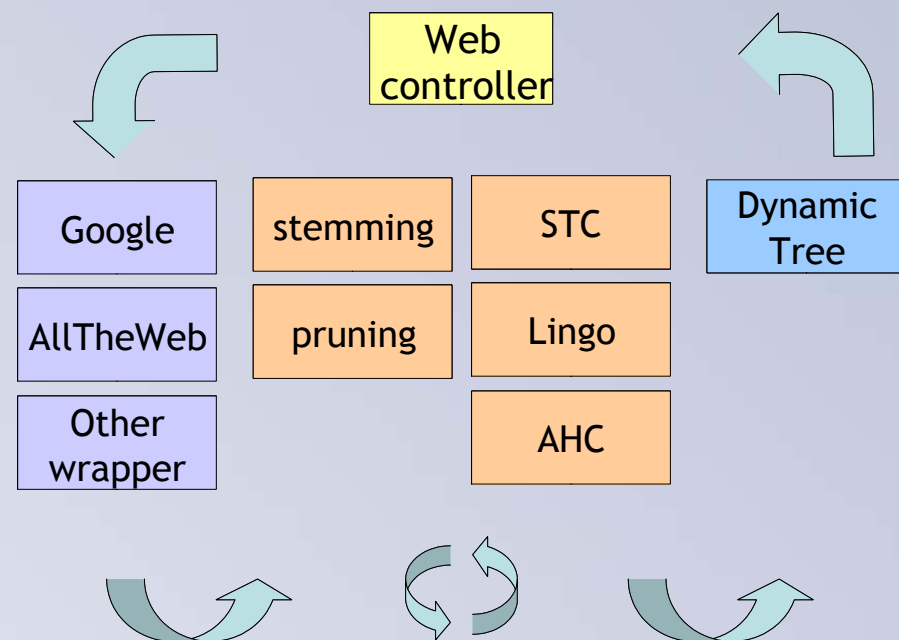
local:8
Carrot
Sort: [flat]
All group
Knowledge (23)
Data Mining
Visual Data
Data Mining
Data Mining
Data Mining
Questi
Data Mining
K-State
and Data
Data Mining
Data Mining
Data Mining
Text Data
Data War
Data Mining
(6)
Research
Internatio
Data War
Pattern (4
TSA Book
Analytical
Data Mining
Experiences
Questions
Data Mining
Petra per
Promotion
other) (3

Plan prezentacji

- o istocie problemu słów kilka
- algorytmy grupowania wyników wyszukiwania
 - model wektorowy - Term Vector Model
 - model powtarzających się fraz - alg. STC
 - inne modele
 - lingo - rodzimy algorytm, „tej”
- propozycje wykorzystania wiedzy lingwistycznej
- **system Carrot²**

O projekcie...

- komponentowy system przetwarzania danych
- wiele istniejących komponentów i narzędzi wspomagających
- projekt prowadzony na SourceForge
 - Licencja BSD-podobna
 - ~5 osób zaangażowanych w prace





Adresy, adresy...

- oficjalna strona projektu:
www.cs.put.poznan.pl/dweiss/carrot
- publiczna wersja demonstracyjna:
<http://carrot.cs.put.poznan.pl>



Dziękuję za uwagę

kontakt, uwagi:
dawid.weiss@cs.put.poznan.pl
<http://www.cs.put.poznan.pl/dweiss>