

Document clustering and its application to web search

Claudio Carpineto
Fondazione Ugo Bordonni

Research work in collaboration with:

G. Romano, A. Bernardini, M. D'Amico, S. Mizzaro, S. Osinski, D. Weiss

Overview

Clustering methods:

K-Means

Quality-threshold clustering

Hierarchical agglomerative clustering

Concept lattices

Why so many clustering algorithms?

Web Clustering Engines:

Issues

Systems

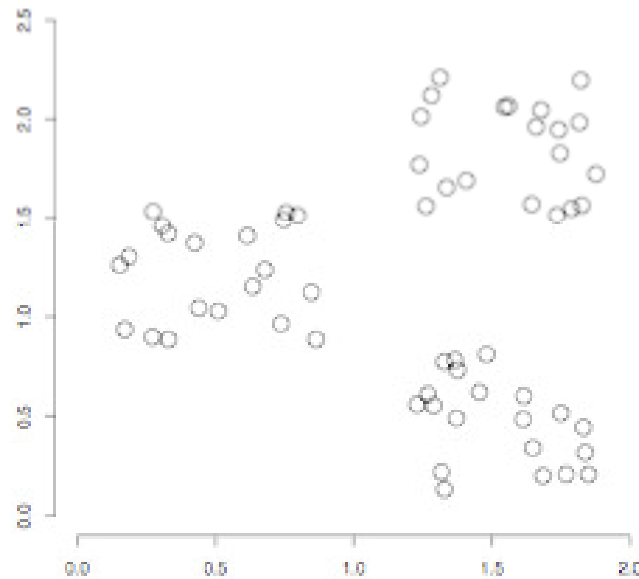
KeySRC

Problems

New Directions

Mobile search

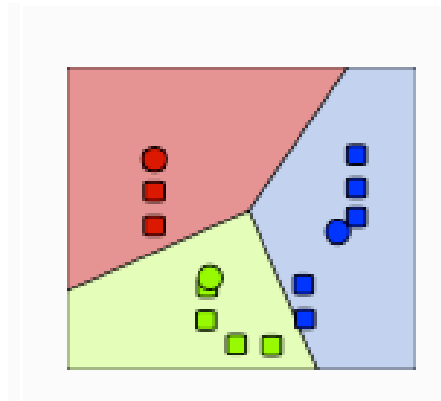
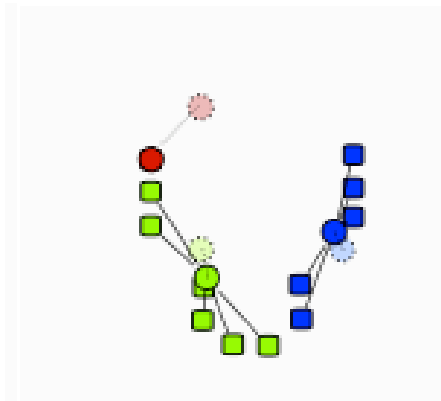
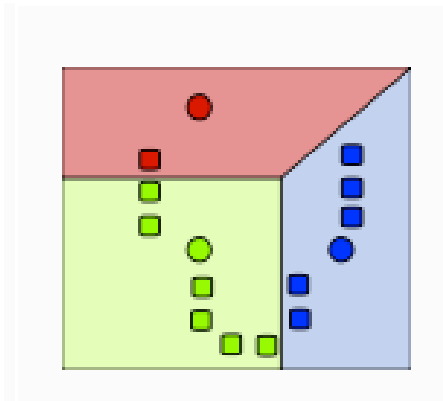
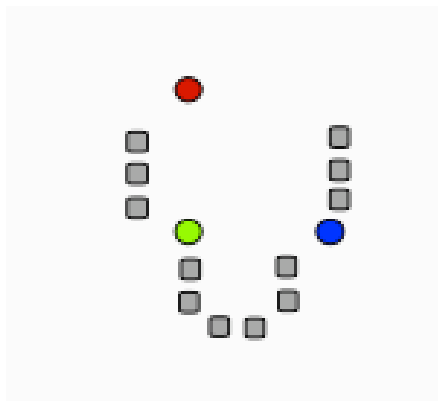
The **goal** of clustering is to maximize the *inter-class* similarity and minimize the *intra-class* similarity



Keys to clustering are the object representation and the distance measure

K-means algorithm

$$V = \sum_{i=1}^k \sum_{x_j \in \mathcal{S}_i} (x_j - \mu_i)^2$$



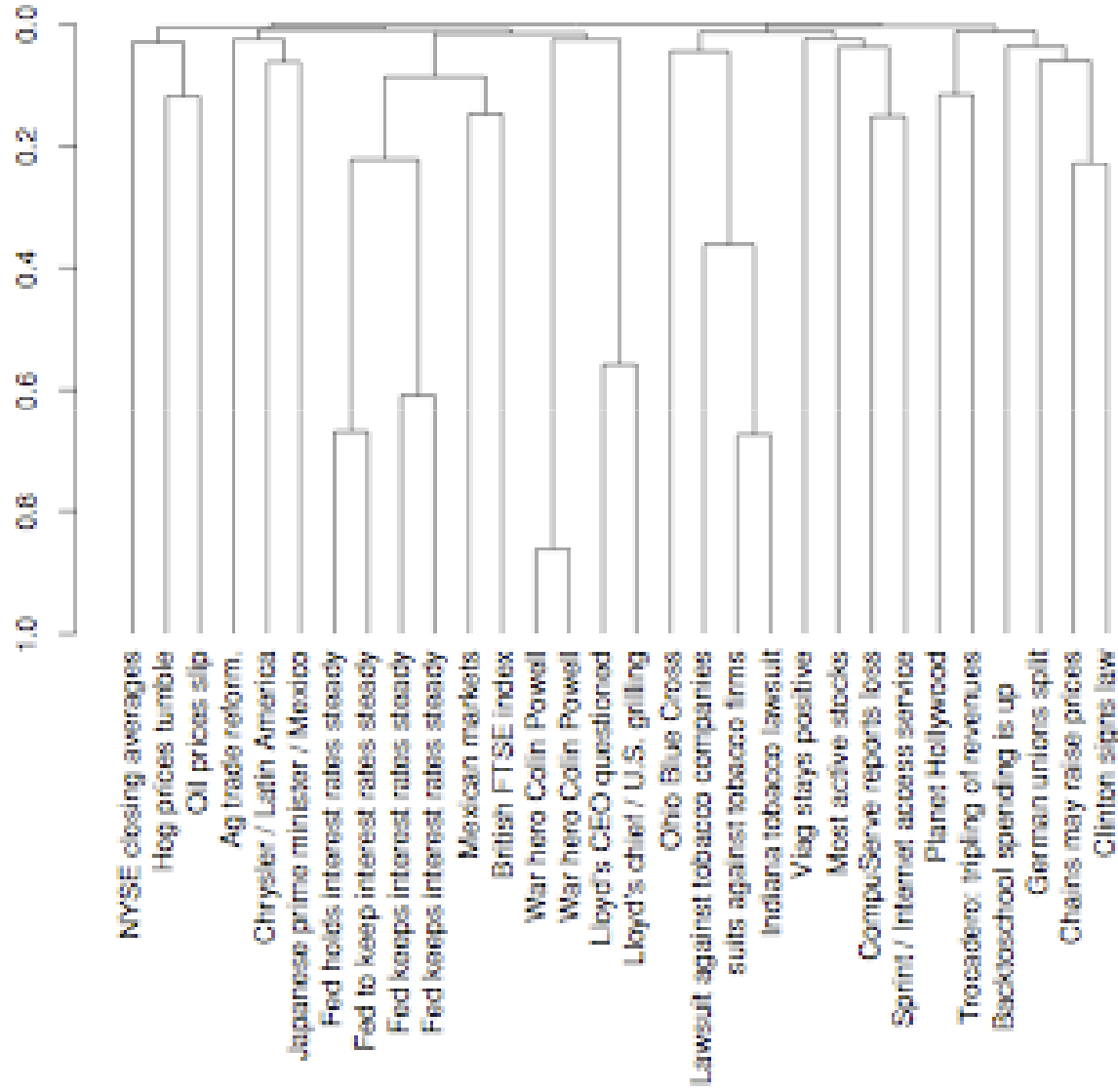
Quality threshold (QT) clustering

1. Choose a *maximum diameter* for clusters.
2. Build a candidate cluster for each point by including *the closest point*, the next closest, and so on, until the diameter of the cluster surpasses the threshold.
3. Save the candidate cluster with *the most points* as the first true cluster, and remove all points in the cluster from further consideration.
4. *Recurse* with the reduced set of points.

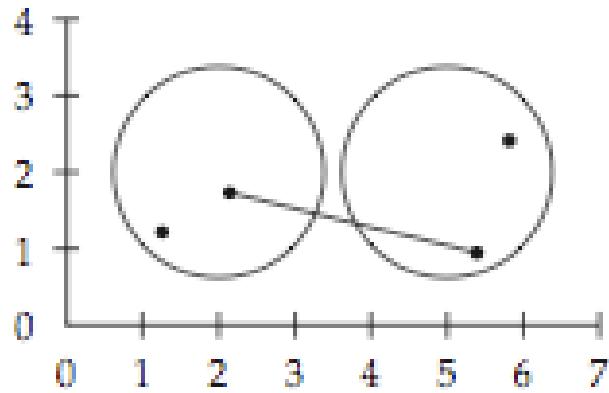
K-means versus QT clustering

	k-means	QT
Deterministic?	no	yes
Need to specify cluster number?	yes	no
Computationally intensive?	no	yes
Every object must be clustered?	yes	no

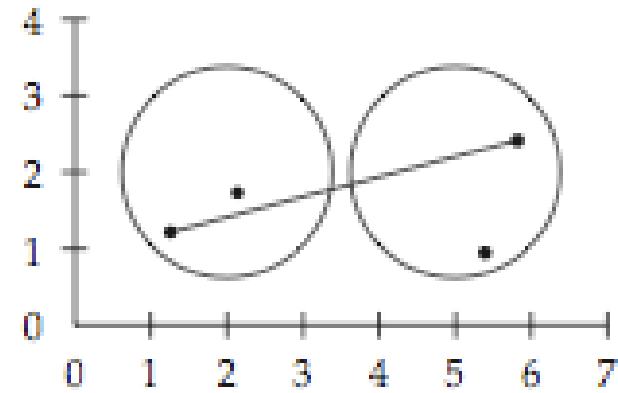
Hierarchical agglomerative clustering



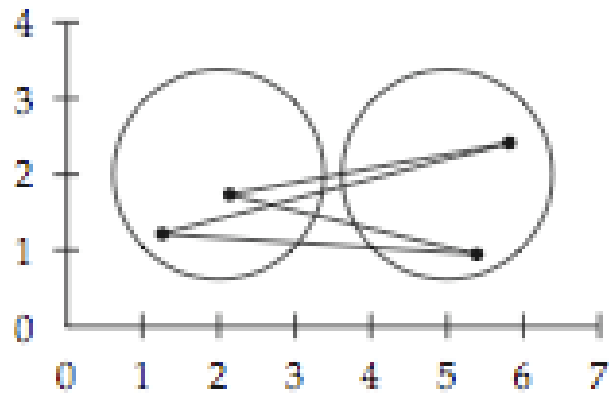
Cluster similarity measures



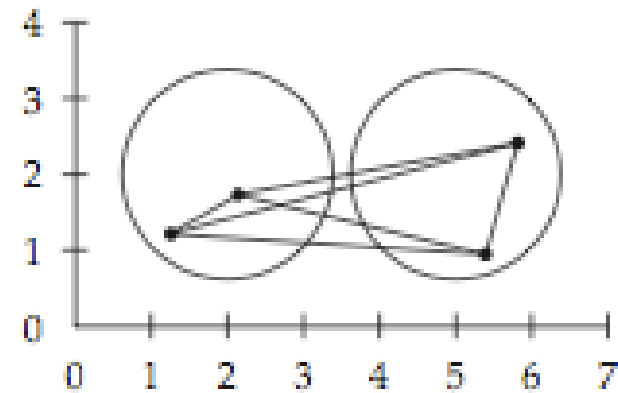
(a) single link: maximum similarity



(b) complete link: minimum similarity

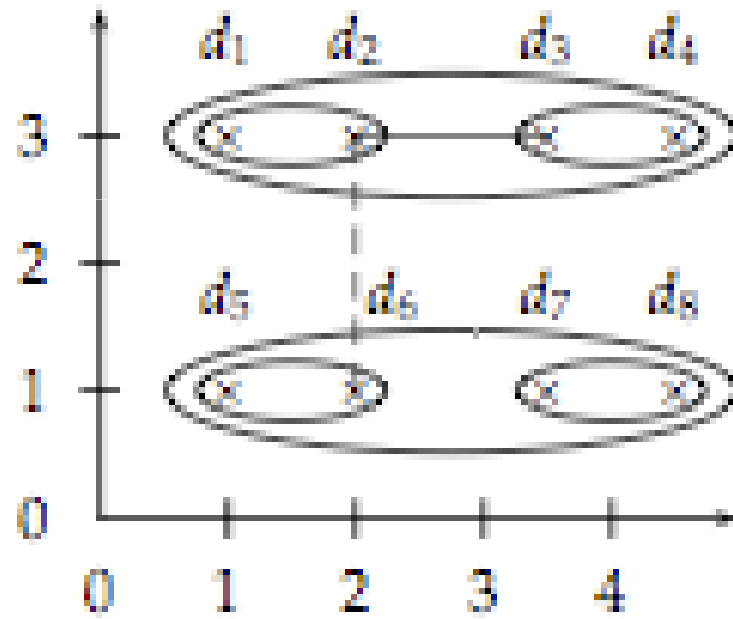


(c) centroid: average inter-similarity

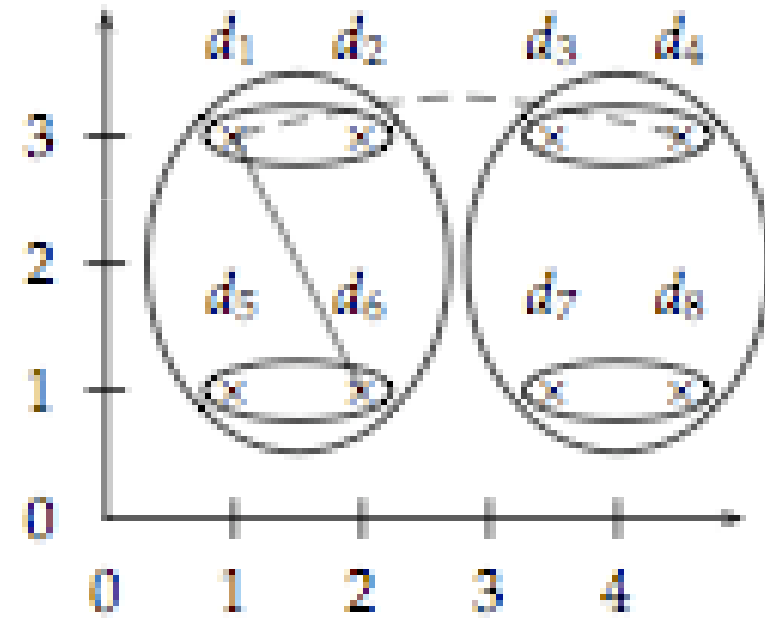


(d) group-average: average of all similarities

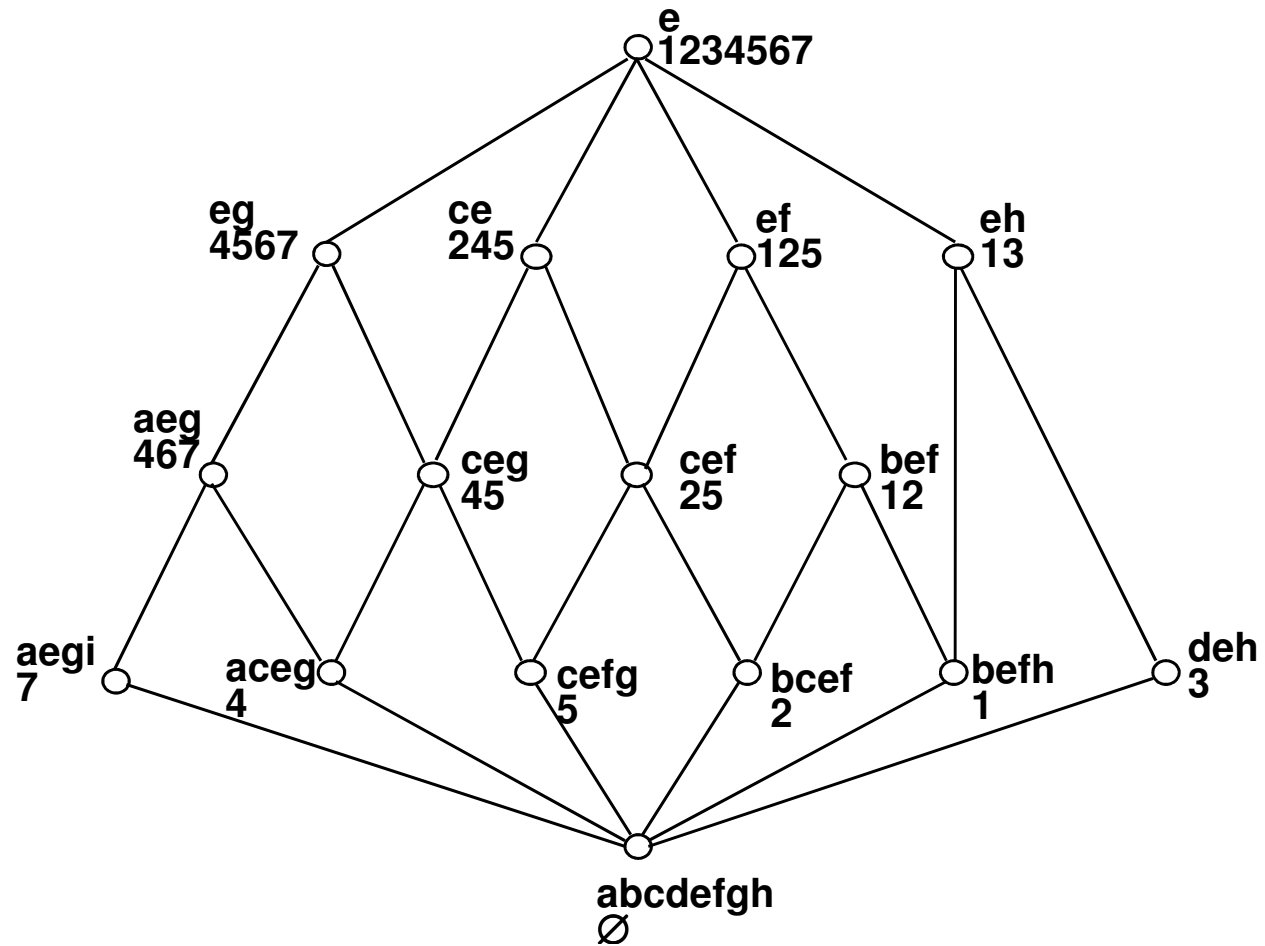
Single-link clustering



Complete-link clustering



		breathes in water (a)	can fly (b)	has beak (c)	has hands (d)	has skeleton (e)	has wings (f)	lives in water (g)	vivipar - ous (h)	produce light (i)
1	Bat		x			x	x		x	
2	Eagle		x	x		x	x			
3	Monkey				x	x			x	
4	Parrot fish	x		x		x		x		
5	Penguin			x		x	x	x		
6	Shark	x				x		x		
7	Lantern fish	x				x		x		x



Concept lattice (definition)

Context = (G, M, I)

gIm or $(g, m) \in I$ means “object g has property m ”

For $A \subseteq G, B \subseteq M$ define:

$A' = \{ m \in M \mid gIm \text{ for all } g \in A \}$

$B' = \{ g \in G \mid gIm \text{ for all } m \in B \}$

A concept of (G, M, I) is a pair (A, B) where

$A \subseteq G, B \subseteq M, A' = B, B' = A$ (A and B are called extent and intent)

A subset $X \subseteq G$ is an extent if and only if $X'' = X$; the concept is (X'', X')

A subset $Y \subseteq M$ is an extent if and only if $Y'' = Y$; the concept is (Y', Y'')

$(A_1, B_1) \leq (A_2, B_2)$ if $A_1 \subseteq A_2$ (or $B_1 \supseteq B_2$)

$C(G, M, I, \leq)$ is a complete lattice and(The Basic Theorem of concept lattices)

NextNeighbours

Input: Context (G, M, I)

Output: The concept lattice $L = (C, E)$ of (G, M, I)

1. $C := \{(G, G')\}$
2. $E := \emptyset$;
3. $currentLevel := \{(G, G')\}$
4. **while** $currentLevel \neq \emptyset$
5. $nextLevel := \emptyset$
6. **for each** $(X, Y) \in currentLevel$
7. $lowerNeighbours := FindLowerNeighbours((X, Y))$
8. **for each** $(X_1, Y_1) \in lowerNeighbours$
9. **if** $(X_1, Y_1) \notin C$ **then**
10. $C := C \cup \{(X_1, Y_1)\}$
11. $nextLevel := nextLevel \cup \{(X_1, Y_1)\}$
12. Add edge $(X_1, Y_1) \rightarrow (X, Y)$ to E
13. $currentLevel := nextLevel$

function *FindLowerNeighbours* $((X, Y))$

/ Returns the lower neighbours of a concept */*

1. $candidates := \emptyset$
2. **for each** $m \in M \setminus Y$
3. $X_1 := (Y \cup \{m})'$
4. $Y_1 := X_1'$
5. **if** $(X_1, Y_1) \notin candidates$ **then**
6. $candidates := candidates \cup \{(X_1, Y_1)\}$
7. **return** maximally general $candidates$

Why so many clustering algorithms?

Clusters and outliers are in the eye of the beholder

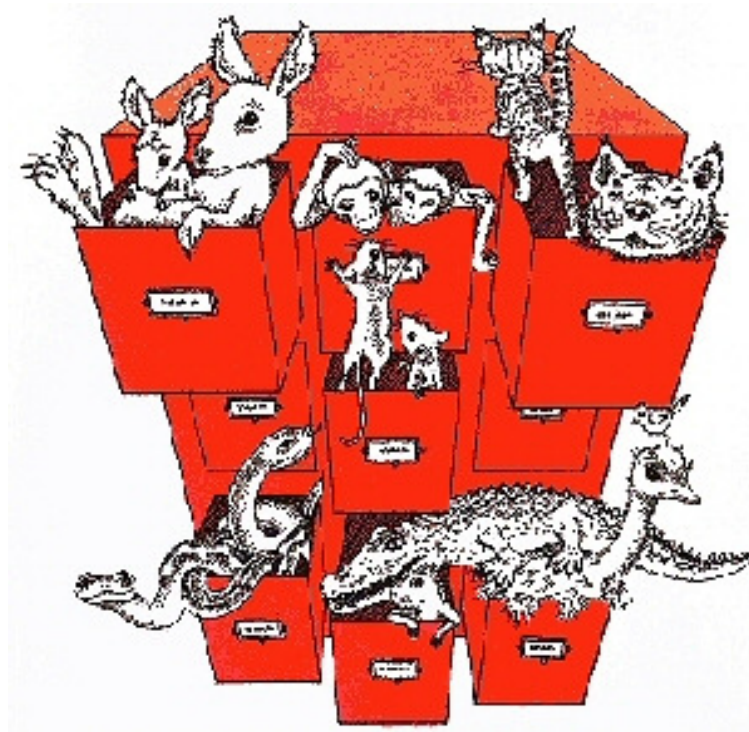
The strategy of cluster analysis is structure seeking although its operation is structure-imposing (e.g., by a clustering criterion)

The main distinction is in the underlying model (e.g., representatives, tree search, graph search)

Clustering often translates into optimization problem solved by approximate algorithms (e.g., k-means, HAC, fuzzy concept lattices)

The clustering algorithm must be compatible with the data structure (e.g., k-means cannot find non-convex clusters)

Web search results clustering



Search engines vs clustering engines (broad query)

The screenshot shows the Clusty search engine interface. At the top, there is a navigation bar with links for 'web', 'news', 'images', 'wikipedia', 'blogs', 'jobs', and 'more'. A search bar contains the word 'rome' and a 'Search' button. To the right of the search bar are links for 'advanced preferences'. Below the search bar, the Clusty logo is visible on the left. The main content area displays 'Top 235 results of at least 8,927,000 retrieved for the query rome (definition) (details)'. A sidebar on the left lists various clusters: 'All Results (240)', 'Italy (59)', 'Ancient (32)', 'Travel (28)', 'Hotels (27)', 'Pictures (26)', 'Georgia (23)', 'Empire (14)', 'Tours, Sightseeing (12)', 'University (16)', and 'School (11)'. Below the sidebar is a 'find in clusters' search box and a 'Font size' selector. The main results area shows a table for 'Current stock trading for (ROME)' with columns for 'SYMBOL', 'LAST', 'CHANGE', 'OPEN', and 'PREV CLOSE'. Below the table are sponsored results for hotels in Rome, including '500 Hotels in Rome', 'Hotels in Rome', and 'Rome Hotels'. The search results section lists five items: 1. 'Rome - Wikipedia, the free encyclopedia', 2. 'HBO: Rome', 3. 'Ancient Rome - Wikipedia, the free encyclopedia', 4. 'The ROME Operating System', and 5. 'Rome.info > Rome tourist information, Ancient Rome travel guide'. Each result includes a brief description and a link to the source.

web news images wikipedia blogs jobs more »

Clusty

rome Search advanced preferences

clusters sources sites remix

All Results (240)

- Italy (59)
- Ancient (32)
- Travel (28)
- Hotels (27)
- Pictures (26)
- Georgia (23)
- Empire (14)
- Tours, Sightseeing (12)
- University (16)
- School (11)

more | all clusters

find in clusters: Find

Font size: A A A A

Top 235 results of at least 8,927,000 retrieved for the query rome (definition) (details)

Current stock trading for (ROME)

SYMBOL	LAST	CHANGE	OPEN	PREV CLOSE
ROME				

Sponsored Results

- [500 Hotels in Rome](#) - Good availability and great rates. Save up to 75% on your booking! - [www.booking.com/Rome-Hotels](#)
- [Hotels in Rome](#) - 55 charming hotels selected. Up to 50% discount. No prepayment. - [www.Hotels-Direct-Rome.com](#)
- [Rome Hotels](#) - Stay Near Rome Paris France Compare Photos, Hotel Info & Rates. - [www.europe-hotelrooms.com](#)

Search Results

- [Rome - Wikipedia, the free encyclopedia](#)
Rome (pronounced /roʊm/: Italian: Roma, pronounced : Latin: Roma) is the capital city of Italy and Lazio. [2] and is Italy's largest and most populous city, with 2,705,317 residents, [3] an urban area of 3,457,690 [4] as well as a metropolitan area of about 4 million inhabitants spread over a 5,352 km² area. [5]History · Administration · Geography · Demographics · Cityscape · Economy [en.wikipedia.org/wiki/Rome](#) - [cache] - Live, Ask
- [HBO: Rome](#)
The homepage for HBO's official 'Rome' website. ... An intimate drama of love and betrayal, masters and slaves, and husbands and wives, 'Rome' chronicles the epic times that saw ... [www.hbo.com/rome](#) - [cache] - Live, Ask
- [Ancient Rome - Wikipedia, the free encyclopedia](#)
Ancient Rome was a civilization that grew out of a small agricultural community founded on the Italian Peninsula as early as the 10th century BC. Located along the Mediterranean Sea, it became one of the largest empires in the ancient world. [1] In its centuries of existence, Roman civilization shifted from a monarchy to an oligarchic republic to an increasingly autocratic empire.History · Society · Culture · Roman Technological ... [en.wikipedia.org/wiki/Ancient_Rome](#) - [cache] - Live, Ask
- [The ROME Operating System](#)
Lightweight, very modular component-based, multitasking, embedded OS; developed, used for many research projects. Goal: manage fast data streams in multimedia environments; for ... [rome.sourceforge.net](#) - [cache] - Live, Open Directory
- [Rome.info > Rome tourist information, Ancient Rome travel guide](#)
Italy - Rome travel guide, tourist information on Rome and Vatican, pictures of Rome, sights and attractions, Rome entertainment, restaurants, hotels etc... [www.rome.info](#) - [cache] - Live, Open Directory, Ask

Search engines vs clustering engines (broad query)


The screenshot displays a search engine interface with a navigation bar at the top containing links for Web, Wiki, Images, News (Boss), Yahoo, MSN, Jobs, PubMed, PUT, and Blogs. A search bar contains the query 'fascism' and a 'Search' button. Below the search bar, there are two tabs: 'Tree' and 'Visualization'. The 'Tree' tab is active, showing a hierarchical list of topics related to fascism, such as 'Italian Fascism (10)', 'History of Fascism (9)', and 'Wikipedia Fascism (6)'. The main content area displays the top 100 results for the query 'fascism', with a total of 1290000 results. The results are listed in a numbered format, each with a title, a brief description, and a URL. The first result is 'Fascism - Wikipedia, the free encyclopedia', followed by 'fascism - Definition from the Merriam-Webster Online Dictionary', 'George W Bush and the 14 points of fascism - Project for the OLD ...', 'Fascism: The Concise Encyclopedia of Economics | Library of ...', 'Modern History Sourcebook: Mussolini: What is Fascism, 1932', 'PublicEye.org - What is Fascism?', 'What is Fascism?', and 'Fascism Anyone?'. The footer of the page contains the query 'fascism -- Source: Web (100 results, 3437 ms) -- Clusterer: Lingo (111 ms)' and version information 'v3.1-dev | build 74 | 2009-03-03 11:06 © 2002-2009 Stanislaw Osinski, Dawid Weiss'.

Search results for **fascism** (1290000 results):

- 1 [Fascism - Wikipedia, the free encyclopedia](#)
Fascism is a radical, authoritarian nationalist ideology that aims to create a single-party state with a government led by a dictator who seeks national ...
<http://en.wikipedia.org/wiki/Fascism> [Ask, Google, Live, Yahoo]
- 2 [fascism - Definition from the Merriam-Webster Online Dictionary](#)
Definition of **fascism** from the Merriam-Webster Online Dictionary with audio pronunciations, thesaurus, Word of the Day, and word games.
<http://www.merriam-webster.com/dictionary/fascism> [Ask, Google, Wikia]
- 3 [George W Bush and the 14 points of fascism - Project for the OLD ...](#)
This page is a collection of news articles dating from the start of the Bush presidency divided into topics relating to each of the 14 points of **fascism**.
...
<http://www.oldamericancentury.org/14pts.htm> [Ask, Google, Yahoo]
- 4 [Fascism: The Concise Encyclopedia of Economics | Library of ...](#)
As an economic system, **fascism** is socialism with a capitalist veneer. The word derives from fasces, the Roman symbol of collectivism and power: a tied ...
<http://www.econlib.org/library/Enc/Fascism.html> [Ask, Cui, Google, Live, Yahoo]
- 5 [Modern History Sourcebook: Mussolini: What is Fascism, 1932](#)
Fascism, the more it considers and observes the future and the development ... Fascism denies that the majority, by the simple fact that it is a majority, ...
<http://www.fordham.edu/halsall/mod/mussolini-fascism.html> [Ask, Cui, Google, Live, Wikia]
- 6 [PublicEye.org - What is Fascism?](#)
Some general ideological features outlined by Matthew N. Lyons.
<http://www.publiceye.org/eyes/whatfasc.html> [Ask, Google, Yahoo]
- 7 [What is Fascism?](#)
The seeds of **fascism**, however, were planted in Italy. "**Fascism** is reaction," said Mussolini, but reaction to what? The reactionary movement following World ...
<http://remember.org/hist.root.what.html> [Ask, Google]
- 8 [Fascism Anyone?](#)
Jul 25, 2004 ... It is **fascism**. And **fascism's** principles are wafting in the air today, ... German and Italian **fascism** form the historical models that define ...
http://www.secularhumanism.org/library/fi/britt_23_2.htm [Ask, Google, Wikia]

Query: **fascism** -- Source: Web (100 results, 3437 ms) -- Clusterer: Lingo (111 ms) v3.1-dev | build 74 | 2009-03-03 11:06 © 2002-2009 Stanislaw Osinski, Dawid Weiss

Search engines vs clustering engines (ambiguous query)



Enter a query:

English Italiano [help](#) [terms of use](#) [about](#)

- [metamorphosis \(100\)](#)
 - [music \(15\)](#)
 - [hilary duff \(6\)](#)
 - [punk myspace \(2\)](#)
 - [rolling stones \(2\)](#)
 - [philip glass \(2\)](#)
 - [cd \(2\)](#)
 - [other \(1\)](#)
 - [insects \(11\)](#)
 - [kafka \(11\)](#)
 - [hilary duff \(9\)](#)
 - [insect \(8\)](#)
 - [butterfly \(7\)](#)
 - [life \(5\)](#)
 - [definition \(4\)](#)
 - [dictionary \(4\)](#)
 - [cd \(4\)](#)
 - [philip glass \(3\)](#)
 - [rolling stones \(3\)](#)
 - [star trek \(3\)](#)
 - [wikipedia \(3\)](#)
 - [spectrasonics \(3\)](#)
 - [other \(31\)](#)

[Amazon.com: Metamorphosis: Music: Hilary Duff](#)
Amazon.com: Metamorphosis: Music: Hilary Duff by Hilary Duff ... Although dubbing her first album Metamorphosis, the disc is anything but. ...
<http://www.amazon.com/Metamorphosis-Hilary-Duff/dp/B0000AGWES>

[Amazon.com: Metamorphosis: Music: The Rolling Stones,Rolling Stones](#)
Amazon.com: Metamorphosis: Music: The Rolling Stones,Rolling Stones by The Rolling Stones,Rolling Stones ... Heart Of Stone (Metamorphosis has the only release ...
<http://www.amazon.com/Metamorphosis-Rolling-Stones/dp/B00006AW2F>

[MySpace.com - METAMORPHOSIS - Vienna/Prag/St.Petersburg, AT - Pop Punk / Ambient / Acoustic - www.myspace.com/...](#)
MySpace music profile for METAMORPHOSIS with tour dates, songs, videos, pictures, blogs, band information, downloads and more
<http://www.myspace.com/contaminatedchamberpop>

[Metamorphosis - Hilary Duff](#)
Metamorphosis album by Hilary Duff including album title, track listings, release dates, guest artists, record label info and user reviews on AOL Music.
<http://music.aol.com/album/metamorphosis/690078>

[MySpace.com - metamorphosis - Lima - Indie / Hardcore / Punk - www.myspace.com/metamorphosisperu](#)
MySpace music profile for metamorphosis with tour dates, songs, videos, pictures, blogs, band information, downloads and more ... necesito un concert de METAMORPHOSIS ...
<http://www.myspace.com/metamorphosisperu>

[Metamorphosis \[CD\] | Target Official Site](#)
Shop for Metamorphosis at Target. Choose from a wide range of Music. Expect More, Pay Less at Target.com
<http://www.target.com/Metamorphosis-Duff-Hilary/dp/B0000AGWES>

[Philip Glass: Music: Solo Piano](#)
Metamorphosis I - V. Mad Rush. Wichita Vortex Sutra ... Metamorphosis was written in 1988 and takes its name from a play based on Kafka's short story. ...
<http://www.philipglass.com/music/solo-piano.php>

Web Clustering Engines:

- *Issues*
- Systems
- Problems
- New Directions

Features of clustering engines

Advantages:

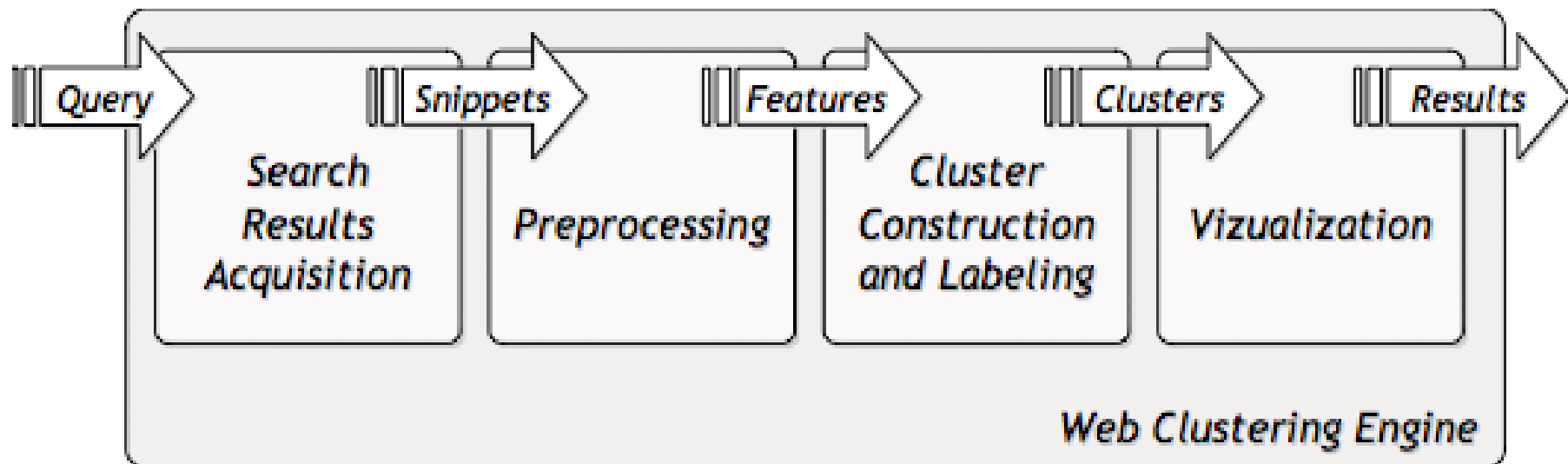
- Fast subtopic retrieval
- Exploring unknown or dynamic domains
- Filtering out irrelevant results

Mostly good when plain search engines fail

Search results clustering (post-retrieval)
versus
traditional document clustering (pre-retrieval)

Clustering type	Cluster labels	Cluster computation	Input data	Cluster number	Cluster intersection	GUI
Search results clustering	Natural language	On-line	Snippets	Variable	Overlapping	Yes
Document clustering	Centroid	Off-line	Documents	Fixed	Disjoint	No

Architecture of Web clustering engines



Search results acquisition

Search engine	Protocol	Queries per day	Results per search	Terms of service
Alexa	SOAP or REST	n/a	20	Paid service (per-query).
Gigablast	REST/XML	100	10	Non-commercial use only.
Google	SOAP	1 000	10	Unsupported as of December 5, 2006. Non-commercial use only.
Google CSE	REST/XML	n/a	20	Custom search over selected sites/ domains. Paid service if XML feed is required.
MSN Search	SOAP	10 000	50	Per application-ID query limit. Non-commercial use only.
Yahoo!	REST/XML	5 000	100	Per-IP query limit. No commercial restrictions (except Local Search).

Preprocessing of search results

1. Language recognition
2. Tokenization
3. Shallow language preprocessing
4. Feature selection

Cluster construction and labeling

“Description comes first”: from **data-centric** to **description-centric** clustering algorithms

Keyphrase-based Search Results Clustering (1)

http://keysrc.fub.it

The screenshot displays the KeySRC search engine interface. At the top, there is a navigation bar with links for Home, Preferences, Links, Documents, and Contact. The KeySRC logo is prominently displayed on the left. A search bar contains the query 'data mining' and a 'Search' button. Below the search bar, the results are organized into two main sections: a left sidebar for 'All results (99)' and a main content area for 'TOP 99 RESULTS OF RETRIEVED FOR THE QUERY DATA MINING'.

All results (99)

- » data mining knowledge discovery (21)
- » data mining and predictive modeling (10)
- » data mining tools (6)
- » business intelligence (8)
- » data mining applications (7)
- » data mining techniques (5)
- » machine learning (5)
- » data mining helps (3)
- » mining software (8)
- » data mining solutions (5)

More clusters

TOP 99 RESULTS OF RETRIEVED FOR THE QUERY DATA MINING

DATA MINING - WIKIPEDIA
Article about knowledge-discovery in databases (KDD), the practice of automatically searching large stores of **data** for patterns.
http://en.wikipedia.org/wiki/Data_mining

DATA MINING: DEFINITION FROM ANSWERS.COM
data mining n. The automatic extraction of useful, often previously unknown information from large databases or **data** ... **Data Mining** For Investing ...
<http://www.answers.com/topic/data-mining>

DATA MINING - WIKIPEDIA, THE FREE ENCYCLOPEDIA
Data mining will not uncover patterns that are present in the domain, but not in ... The term **data mining** is often used to apply to the two separate processes of ...
http://en.wikipedia.org/wiki/Subject-based_data_mining

AN INTRODUCTION TO DATA MINING
Data mining, the extraction of hidden predictive information from large ... **Data mining** tools predict future trends and behaviors, allowing businesses to ...
<http://www.theartling.com/text/dmwhite/dmwhite.htm>

ELECTRONIC STATISTICS TEXTBOOK: DATA MINING TECHNIQUES
Outlines the crucial concepts in **data mining**, defines the **data** warehousing process, and offers examples of computational and graphical exploratory **data** analysis ...
<http://www.statsoft.com/textbook/stdatmin.html>

DATA MINING: TEXT MINING, VISUALIZATION AND SOCIAL MEDIA
Commentary on text **mining**, **data mining**, social media and **data** visualization. ... relevant publications in top **data mining**, information retrieval and natural ...
<http://datamining.typepad.com/>

DATA MINING: WHAT IS DATA MINING?
Outlines what knowledge discovery, the process of analyzing **data** from different perspectives and summarizing it into useful information, can do and how it works.
<http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm>

Keyphrase-based Search Results Clustering (2)

http://keysrc.fub.it

The screenshot shows the KeySRC search results page for the query 'data mining'. The page features a blue header with the KeySRC logo, navigation links (Home, Preferences, Links, Documents, Contact), and a search bar containing 'data mining'. The main content area is divided into a left sidebar with a cluster list and a right main area with document details.

KeySRC Home Preferences Links Documents Contact

data mining Search

All results (99)

- » data mining knowledge discovery (21)
- » data mining and predictive modeling (10)
- » data mining tools (6)
- » business intelligence (8)
- » data mining applications (7)
- » data mining techniques (5)
- » machine learning (5)
- » data mining helps (3)
- » mining software (8)
- » data mining solutions (5)

More clusters

YOU ARE IN "DATA MINING TOOLS" CLUSTER WITH 6 DOCUMENTS

AN INTRODUCTION TO DATA MINING
Data mining, the extraction of hidden predictive information from large ... Data mining tools predict future trends and behaviors, allowing businesses to ...
<http://www.theartling.com/text/dmwhite/dmwhite.htm>

TWO CROWS WHITE PAPER: "SCALABLE DATA MINING"
White paper on the reasons scalable data mining solutions are needed and the main methods of achieving scalability. ... Making Data Mining Tools Scalable ...
<http://www.twocrows.com/whitep.htm>

OPEN DIRECTORY - COMPUTERS: SOFTWARE: DATABASES: DATA MINING: TOOL VENDORS
Advanced Software Applications - Data mining, analysis, and decision support ... Provides neural network software for data mining and forecasting as well as ...
http://www.dmoz.org/Computers/Software/Databases/Data_Mining/Tool_Vendors/

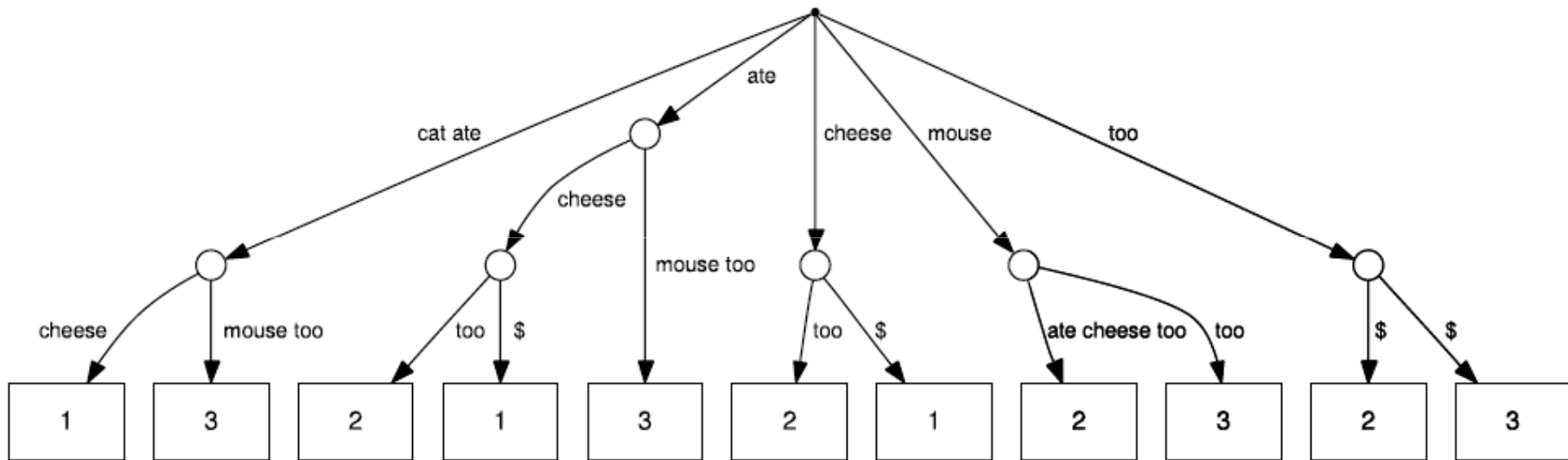
DATA MINING - DATA MINING SOFTWARE (SOFTWARE) - THE DATA MINE WIKI
All currently available data mining software, tools and applications whose ... You can add Data Mining tools / software to this list. Don't know where to start? ...
<http://www.the-data-mine.com/bin/view/Software/DataMiningSoftware>

UNDERSTANDING DATA MINING
... theory, whereas data mining emerged from computational ... Data mining techniques do not require the user to know the ... Today's data mining tools ...
http://www.qbase.us/pdf/Technology/White%20Paper_Data%20Mining.pdf

A COMPARISON OF LEADING DATA MINING TOOLS
on Knowledge Discovery & Data Mining. Friday, August 28, 1998. New York, New York ... Compare and Summarize Data Mining Tools which: ...
http://www.datamininglab.com/pubs/kdd98_elder_abbott_nopics_bw.pdf

Generalized suffix tree

(from Zamir and Etzioni, 1998)



- 1) Cat ate cheese
- 2) Mouse ate cheese too
- 3) Cat ate mouse too

KeySRC algorithm

1. Search results preprocessing
2. Construction of Generalized Suffix Tree (GST)
3. Extraction of keyphrases from GST
(internal nodes of GST + ≤ 4 words + POS tagging)
4. Keyphrase clustering

$$Thr = const \frac{ics(c_1) |c_1| + ics(c_2) |c_2|}{|c_1| + |c_2|}$$

5. Label assignment

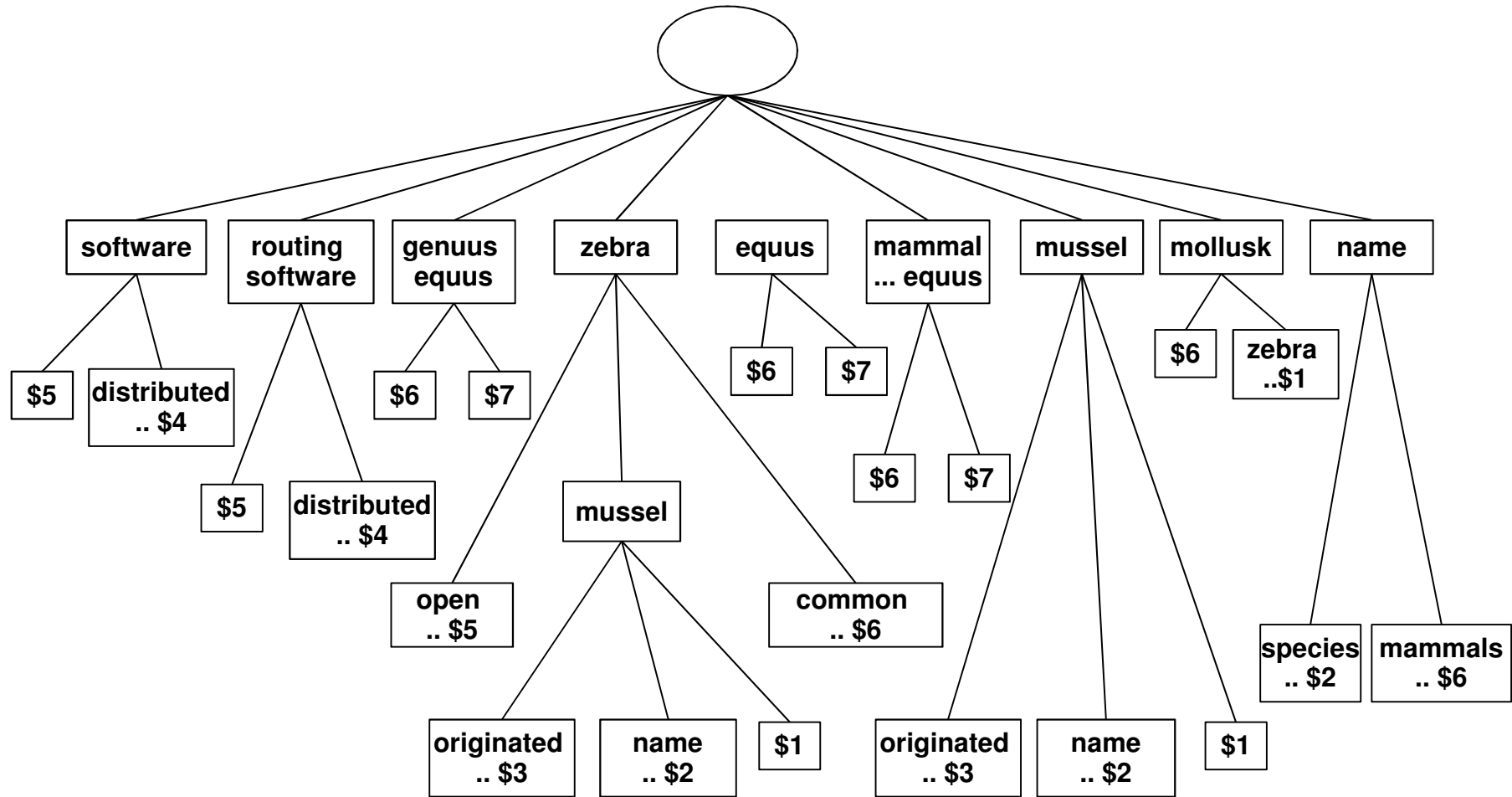
$$Score = KPfreq \sum_{t \in W_{KP}} wfreq$$

6. Cluster ranking

Search results for query “zebra”

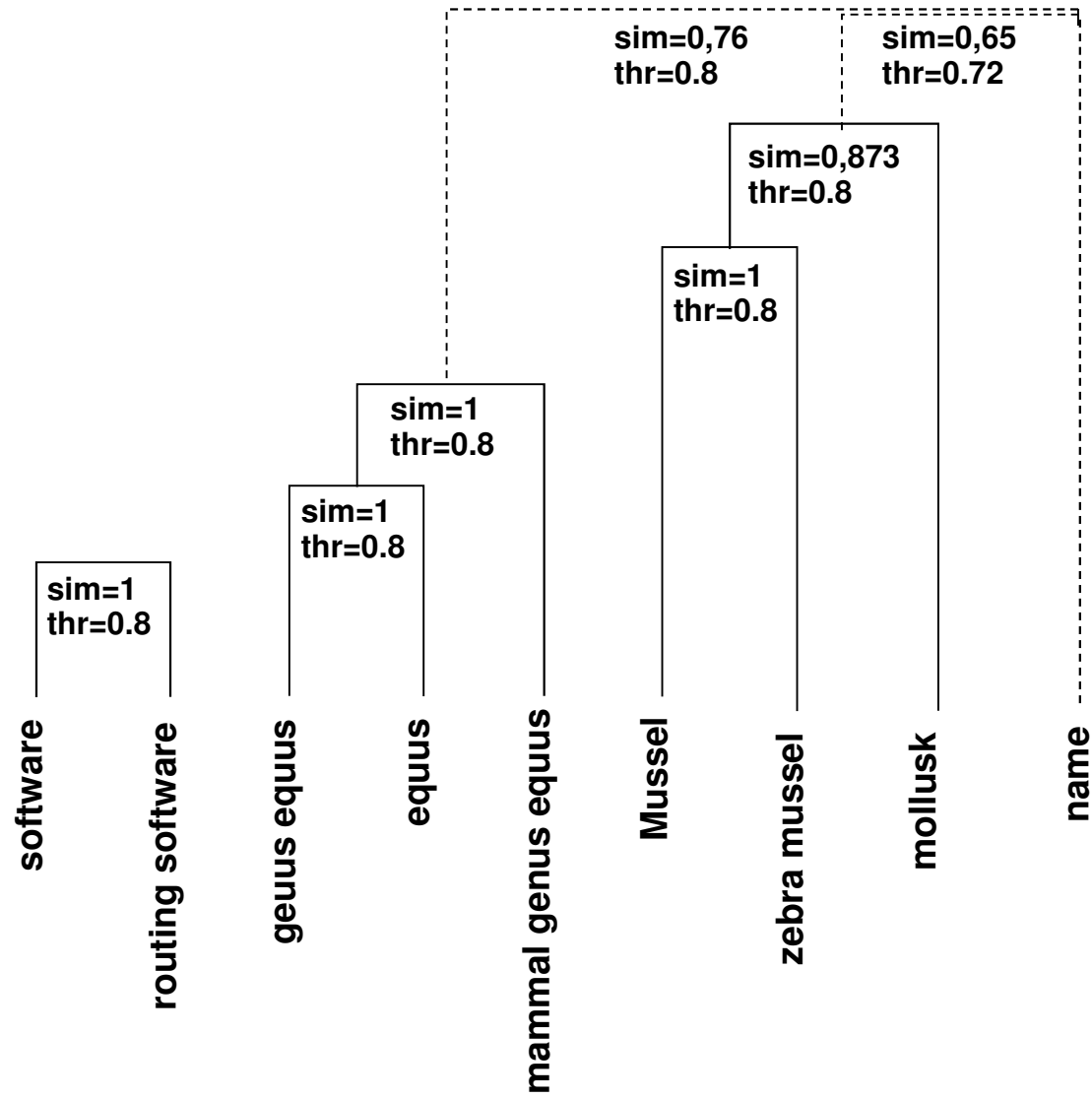
- *D1*: Harmful aquatic hitchhikers: mollusks, zebra mussel.
- *D2*: Zebra mussel: name of a species of mollusks.
- *D3*: Zebra mussel originated in the Balkans, Poland.
- *D4*: Free routing software distributed under GNU license.
- *D5*: Zebra is open source TCP/IP routing software.
- *D6*: Zebra is the common name for some mammals of the genus equus.
- *D7*: Horselike African mammals of the genus equus.

GST of “zebra” results

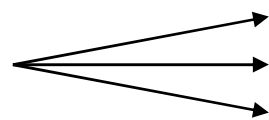


	A	B	C	D	E	F	G	H	I
A	1	1	0	0	0	0	0	0	0
B	1	1	0	0	0	0	0	0	0
C	0	0	1	1	0	0	0	1	0.5
D	0	0	1	1	0	0	0	1	0.5
E	0	0	0	0	1	1	0,81	0	0.41
F	0	0	0	0	1	1	0,81	0	0.41
G	0	0	0	0	0,81	0,81	1	0	0.5
H	0	0	1	1	0	0	0	1	0.5
I	0	0	0	0.5	0.5	0.41	0.41	0.5	1

Table 1. Similarity matrix for the keyphrases extracted from the seven documents. A = software, B = routing software, C = genus equus, D = equus, E = mussel, F = zebra mussel, G = mollusk, H = mammal genus equus, I = name.



Final clusters



- zebra mussel (D1, D2, D3)
- mammals of the genus Equus (D6, D7)
- routing software (D4, D5)

Keyword-based visualization

RANK	MEMBERS
1	<u>handgun</u> , <u>revolver</u> , <u>shotgun</u> , <u>pistol</u> , <u>rifle</u> , <u>machine gun</u> , <u>sawed-off shotgun</u> , <u>submachine gun</u> , <u>gun</u> , <u>automatic pistol</u> , <u>automatic rifle</u> , <u>firearm</u> , <u>carbine</u> , <u>ammunition</u> , <u>magnum</u> , <u>cartridge</u> , <u>automatic</u> , <u>stopwatch</u>
236	<u>whitefly</u> , <u>pest</u> , <u>aphid</u> , <u>fruit fly</u> , <u>termite</u> , <u>mosquito</u> , <u>cockroach</u> , <u> flea</u> , <u>beetle</u> , <u>killer</u> , <u>bee</u> , <u>maggot</u> , <u>predator</u> , <u>mite</u> , <u>houseplant</u> , <u>cricket</u>
471	<u>supervision</u> , <u>discipline</u> , <u>oversight</u> , <u>control</u> , <u>governance</u> , <u>decision making</u> , <u>jurisdiction</u>
706	<u>blend</u> , <u>mix</u> , <u>mixture</u> , <u>combination</u> , <u>juxtaposition</u> , <u>combine</u> , <u>amalgam</u> , <u>sprinkle</u> , <u>synthesis</u> , <u>hybrid</u> , <u>melange</u>

<input type="checkbox"/> Cluster 1 Size: 8 control drive accident program office design front-wheel invent
<input type="radio"/> AP: Auto Maker Recalls 285,000 Front-wheel Drive Vehicles AP900525-0242
<input type="radio"/> SJMN: USED CARS ARE OUTSELLING NEW AT DEALERSHIPS SJMN91-062570
<input type="radio"/> ZF: AutoTrack (brief article) (computer-aided design software from Savoy Computing) (
<input type="radio"/> AP: Army Commander Breaks Arm in Car Accident AP880905-0143
<input type="radio"/> ZF32-294-735 ZF32-294-735
<input type="checkbox"/> Cluster 2 Size: 25 battery california technology mile state recharge impact officia
<input type="radio"/> WSJ: Nissan Unveils Electric Car Claims 'Fastest' Recharge WSJ910826-0053
<input type="radio"/> WSJ: Autos: GM Says It Plans an Electric Car, but Details Are Spotty ---- By Joseph B.
<input type="radio"/> WSJ: Autos: Auto Makers Strive to Get Up to Speed On Clean Cars for the California Mar
<input type="radio"/> WSJ: Technology: Nissan Plans Electric Car With Very Fast Recharging WSJ910625-00
<input type="radio"/> SJMN: NISSAN JOINS ELECTRIC CAR RACE WITH BEST BATTERY SJMN91-06
<input type="checkbox"/> Cluster 3 Size: 48 import j. rate honda toyota trk light veh drop mazda percentag
<input type="radio"/> WSJ: U.S. Car Sales Fell 12.9% in Late May As Signs of Recovery Detour Detroit ---- I
<input type="radio"/> WSJ: Economy: Auto Sales Fell 4.5% in Late February; Dealers Report No Postwar Rebou
<input type="radio"/> WSJ: Car, Truck Sales Fell 21.3% in Late April, In Lowest Annual pace Since December
<input type="radio"/> WSJ: U.S. Car Sales Edged Higher At End of July ---- Auto Makers Keep Making Slow P
<input type="radio"/> WSJ: Economy: Car Sales Rose Slightly in Latest 10 Days; Greenspan Says Rate Cuts to A
<input type="checkbox"/> Cluster 4 Size: 16 export international unit japan trade manufacturer citation gery

Folder-based visualization

The screenshot shows a web browser window with the Clusty Search interface. The search term is "metamorphosis". The left sidebar displays a hierarchical folder structure of search results. The main content area shows a list of search results, with the first result selected, displaying a snippet of text from a document.

Clusty Search » metamorphosis

web news images wikipedia blogs jobs more »

metamorphosis Search advanced preferences

clusters sources sites

All Results (203) remix

- Insects (18)
 - Biology (4)
 - Stages, Adult (3)
 - Gregor Samsa (3)**
 - Cycle, Butterflies (2)
 - Incomplete Metamorphosis (3)
 - Other Topics (3)
- Butterfly (21)
- Images (16)
- Franz Kafka (13)
 - Found himself transformed (3)
 - Project Gutenberg (2)
 - Study Guide (2)
 - Other Topics (6)
- Definition, Dictionary (9)
- Biology (6)
 - Design (7)
- Artist (8)
 - Women (5)
- Painting, Face (6)

find in clusters: Find

Cluster **Insects** » **Gregor Samsa** contains 3 documents.

Search Results

- metamorphosis**
As **Gregor Samsa** awoke one morning from uneasy dreams he found himself transformed in his bed into a gigantic **insect**. He was lying on his hard, ... What has happened to me? he thought. ...
victorian.fortunecity.com/vermeer/287/metamorphosis.htm - [cache] - Ask
- The Metamorphosis**
Franz Kafka The **Metamorphosis** by kafka ... The **Metamorphosis** -The complete story ... As **Gregor Samsa** awoke one morning from uneasy dreams he found himself transformed in his bed into a gigantic **insect**. ...
www.kafka-franz.com/metamorphosis.htm - [cache] - Ask
- The Metamorphosis - Wikipedia, the free encyclopedia**
The **Metamorphosis** (German: Die Verwandlung) is a novella by Franz Kafka, first published in 1915. The story begins with a traveling salesman, **Gregor Samsa**, waking to find himself transformed into a "monstrous vermin" (see Lost in translation, below). Characters **Gregor Samsa** Greta...
en.wikipedia.org/wiki/The_Metamorphosis - [cache] - Ask

Search for more results like these

Nesting & zooming visualization



Selected Sources [2 of 3] [Add/Remove](#)
Yahoo!, Wikipedia

tiger

GROK [Search Options](#)

Working List

0 items in your list
[View your list](#)

[Email Map...](#)

[Export Map...](#)

Search within the map:

by keyword

Exclude

by date

all most recent

by source

<All sources>

by domain

<All domains>

[Hide Tools](#)

Layout Color

[Outline View](#)

Map View

250 total results

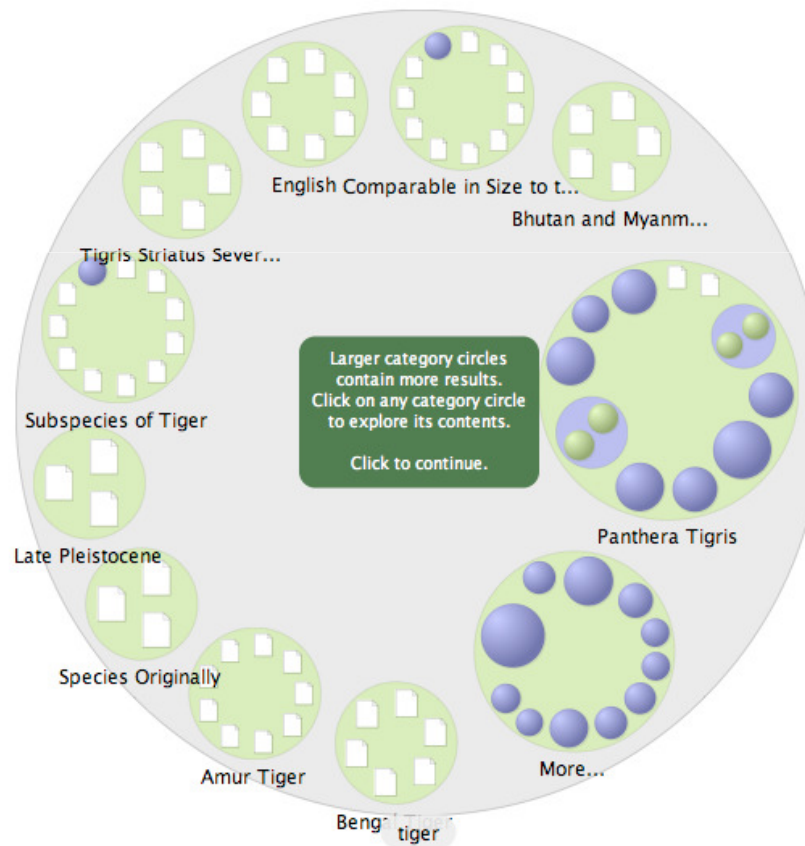
[ZOOM BACK](#)

[TOP](#)

[Expand View](#)

Detail: [Less](#) [Medium](#) [More](#)

[Expand Detail](#)



Biological reproduction

[Add to Working List](#) | [Post to del.icio.us](#) | [Email](#)

...Lions have also been known to breed with tigers (most often the Amur and Bengal subspecies) to create hybrids called ligers and...
http://en.wikipedia.org/wiki/Biological_reproduction - December 5, 2007
Source: Wikipedia

TigerDirect.com Best Deals - Computer Parts, PC Components, Computers & Electronics

[Add to Working List](#) | [Post to del.icio.us](#) | [Email](#)

Offers a wide selection of new and refurbished computer products. ... Tiger should be recognized as a true leader not only in the supply of tech items ...
<http://www.tigerdirect.com/> - 101k - December 3, 2007
Source: Yahoo!

India

[Add to Working List](#) | [Post to del.icio.us](#) | [Email](#)

...Bengal tiger or the Royal Bengal tiger (*Panthera tigris tigris*) is found in parts of India, Bangladesh, Nepal, Bhutan and Myanmar. It lives in varied habitats: grasslands, subtropical and tropical...
<http://en.wikipedia.org/wiki/India> - December 5, 2007
Source: Wikipedia

U.S. Census Bureau - TIGER/Line®

[Add to Working List](#) | [Post to del.icio.us](#) | [Email](#)

... based on the TIGER/Line files, with links to ordering information. ... MAF/TIGER Accuracy Improvement Project ... 2006 First Edition TIGER/Line® Files - No ...
<http://www.census.gov/geo/www/tiger/> - 16k - November 20, 2007
Source: Yahoo!

Quartz

[Add to Working List](#) | [Post to del.icio.us](#) | [Email](#)

...recorded in 1811. "Tiger's-eye" is a name for a golden-brown striped, chatoyant, fibrous variety of quartz used as a semi-precious gemstone. It was one of the many species originally described, as...
<http://en.wikipedia.org/wiki/Quartz> - December 5, 2007
Source: Wikipedia

Official Website for Tiger Woods

[Add to Working List](#) | [Post to del.icio.us](#) | [Email](#)

Official site for pro golfer Tiger Woods, complete with video interviews, photos, stats, and features.
<http://www.tigerwoods.com/> - 18k - December 3, 2007
Source: Yahoo!

Graph-based visualization

The image shows a screenshot of the KartOO search engine interface. The search term is "tiger", and it has found 9,470,000 results. The results are visualized as a network graph where nodes represent websites and edges represent relationships between them. The nodes include:

- www.bbc.co.uk
- www.tiger.gov.uk
- www.iwantoneofthose.com
- www.tigerdirect.com
- www.greyscale.com
- www.census.gov
- www.tigerbayramblers.org.uk
- www.tigertiger.co.uk
- en.wikipedia.org
- www.animaladoptions.co.uk
- www.tigerwoods.com
- www.ebay.co.uk

The relationships (edges) between these nodes are labeled with terms such as "days", "parts", "work", "photos", "walk", "offers", "project", and "continue".

On the left side, there is a "Topics" list with the following items:

- bengal tiger
- lions and tiger
- information on the tiger
- group tiger
- tiger woods pga tour
- walk
- work
- photos
- continue
- project
- days
- parts
- offers

The interface also features a search bar, navigation buttons (back, forward, home, etc.), and a "display next map" button at the bottom right.

Web Clustering Engines:

- Issues
- *Systems*
- Problems
- New Directions

Research prototypes Web clustering engines

System name (algorithm alias)	Year	Text features	Cluster labels	Clustering method	On-line demo	Clusters structure	Source code
Grouper (STC)	1998	single words, phrases	phrases	STC	yes (dead)	flat, concept cloud	no
Lassi	2000	lexical affinities	lexical affinities	AHC	no (desktop)	hierarchy	no
CIIRarchies	2001	single words	word sets	language model/ graph analysis	yes (dead)	hierarchy	no
WICE (SHOC)	2002	single words, phrases	phrases	SHOC	yes (dead)	hierarchy	no
Carrot ² (Lingo)	2003	frequent phrases	phrases	Lingo	yes	flat	yes
Carrot ² (TRSC)	2004	words, tolerance rough sets	n-grams (of words)	TRSC	yes	flat (optional hierarchy)	yes
WebCat	2003	single words	words	k-Means	yes (dead)	flat	no
AIsearch	2004	single words	word sets	AHC + weighted centroid covering	yes (dead)	hierarchy	no
CREDO	2004	single words	word sets	concept lattice	yes	graph	no
DisCover	2004	single words, noun phrases	phrases	incremental cover- age optimization	no	hierarchy	no
SnakeT	2004	approximate sentences	phrases	approx. sent. coverage	yes	hierarchy	no
SRC	2004	n-grams (of words)	n-grams (of words)	SRC	yes	flat (paper) hierarchy (demo)	no
EigenCluster	2005	single words	three salient terms	divide-merge (hybrid)	yes	flat (optional hierarchy)	no
WhatsOnWeb	2006	single words	phrases	edge connectivity	yes	graph	no

Commercial Web clustering engines

Name	Company	Cluster labels	URL	Clusters structure
Accumo	Accumo	Phrases	www.accumo.com	Tree
Clusterizer	CyberTavern	Phrases	www.iboogie.com	Tree
Cowskid	Compara	Terms	www.cowskid.com	Flat
Fluster	Funnelback	Phrases	www.funnelback.com	Flat
Grokker	Grokker	Phrases	www.grokker.com	Graphical/Tree
KartOO	KartOO	Phrases	www.kartoo.com	Graphical/Tree
Lingo3G	Carrot Search	Phrases	www.carrot-search.com	Tree
Mooter	Mooter Media	Phrases	www.mooter.com	Graphical/Flat
WebClust	WebClust	Phrases	www.webclust.com	Tree
Vivísimo	Vivísimo	Phrases	www.vivisimo.com	Tree

Web Clustering Engines:

- Issues
- Systems
- **Problems**
- New Directions

Response times

Source	Avg. delay [s]	Std. dev. [s]
Yahoo! API	2.12	0.65
Google API	5.85	2.35
MSN API	0.56	0.43
Lucene (snippets)	1.78	0.50

Algorithm	50 snippets [s]	100 snippets [s]	200 snippets [s]	400 snippets [s]
CREDO	0.031	0.088	0.272	0.906
Lingo	0.025	0.138	0.243	0.243
Lingo3G	0.009	0.020	0.045	0.070
STC	0.007	0.014	0.030	0.070
TRSC	0.072	0.552	1.368	4.754

To improve efficiency:

Client-side processing, incremental processing, pretokenized docs...

Theoretical limitations of clustering engines

Quality and usability of clusters is still unsatisfactory:

- *incompleteness of clusters,*
- *lack of intra- and inter-cluster consistency,*
- *label expressiveness,*
- *different cluster granularity*

Evaluation of retrieval performance still an open issues


Evaluation method	Cluster validity	Classification accuracy	Label quality	Reach time	Subtopic reach time	User studies
Fully automatic	✓	✓		✓	✓	
No need for test collection	✓		✓			✓
Mathematical measures	✓	✓	✓	✓	✓	
Task oriented				✓	✓	✓
Labels handled			✓		✓	✓

Subtopic reach time

$$SRT_{list} = \frac{\sum_{i=1}^n \min_j(p_{i,j})}{n}$$

$$SRT_{cluster} = \frac{\sum_{i=1}^n \min_j(c_{i,j} + r_{i,j})}{n}$$

Label-driven Subtopic Reach Time



Enter a query:

English Italiano [help](#) [terms of use](#) [about](#)

- [metamorphosis \(100\)](#)
 - **music (15)**
 - [hilary duff \(6\)](#)
 - [punk myspace \(2\)](#)
 - [rolling stones \(2\)](#)
 - [philip glass \(2\)](#)
 - [cd \(2\)](#)
 - [other \(1\)](#)
 - [insects \(11\)](#)
 - [kafka \(11\)](#)
 - [hilary duff \(9\)](#)
 - [insect \(8\)](#)
 - [butterfly \(7\)](#)
 - [life \(5\)](#)
 - [definition \(4\)](#)
 - [dictionary \(4\)](#)
 - [cd \(4\)](#)
 - [philip glass \(3\)](#)
 - [rolling stones \(3\)](#)
 - [star trek \(3\)](#)
 - [wikipedia \(3\)](#)
 - [spectrasonics \(3\)](#)
 - [other \(31\)](#)

[Amazon.com: Metamorphosis: Music: Hilary Duff](#)
Amazon.com: Metamorphosis: Music: Hilary Duff by Hilary Duff ... Although dubbing her first album Metamorphosis, the disc is anything but. ...
<http://www.amazon.com/Metamorphosis-Hilary-Duff/dp/B0000AGWES>

[Amazon.com: Metamorphosis: Music: The Rolling Stones,Rolling Stones](#)
Amazon.com: Metamorphosis: Music: The Rolling Stones,Rolling Stones by The Rolling Stones,Rolling Stones ... Heart Of Stone (Metamorphosis has the only release ...
<http://www.amazon.com/Metamorphosis-Rolling-Stones/dp/B00006AW2F>

[MySpace.com - METAMORPHOSIS - Vienna/Prag/St.Petersburg, AT - Pop Punk / Ambient / Acoustic - www.myspace.com/...](#)
MySpace music profile for METAMORPHOSIS with tour dates, songs, videos, pictures, blogs, band information, downloads and more
<http://www.myspace.com/contaminatedchamberpop>

[Metamorphosis - Hilary Duff](#)
Metamorphosis album by Hilary Duff including album title, track listings, release dates, guest artists, record label info and user reviews on AOL Music.
<http://music.aol.com/album/metamorphosis/690078>

[MySpace.com - metamorphosis - Lima - Indie / Hardcore / Punk - www.myspace.com/metamorphosisperu](#)
MySpace music profile for metamorphosis with tour dates, songs, videos, pictures, blogs, band information, downloads and more ... necesito un concert de METAMORPHOSIS ...
<http://www.myspace.com/metamorphosisperu>

[Metamorphosis \[CD\] | Target Official Site](#)
Shop for Metamorphosis at Target. Choose from a wide range of Music. Expect More, Pay Less at Target.com
<http://www.target.com/Metamorphosis-Duff-Hilary/dp/B0000AGWES>

[Philip Glass: Music: Solo Piano](#)
Metamorphosis I - V. Mad Rush. Wichita Vortex Sutra ... Metamorphosis was written in 1988 and takes its name from a play based on Kafka's short story. ...
<http://www.philipglass.com/music/solo/piano/pba>

The AMBIENT (AMBIguous ENTRIES) test collection

Topic number	Topic description	# of Wikipedia subtopics	# of retrieved subtopics	# of relevant results
1	Aida	31	11	61
2	B-52	6	3	75
3	Beagle	11	4	86
4	Bronx	10	4	81
5	Cain	22	8	38
6	Camel	13	6	70
7	Coral Sea	6	4	45
8	Cube	26	10	48
9	Eos	21	8	65
10	Excalibur	26	8	32
11	Fahrenheit	13	8	72
12	Globe	18	11	53
13	Hornet	23	9	45
14	Indigo	28	15	40
15	Iwo Jima	10	7	90
16	Jaguar	22	6	80
17	La Plata	12	7	68
18	Labyrinth	16	6	26
19	Landau	37	15	40
20	Life on Mars	7	4	84
21	Locust	15	7	50
22	Magic Mountain	10	7	41
23	Matador	22	8	38
24	Metamorphosis	17	7	57
25	Minotaur	7	7	51
26	Mira	22	13	38
27	Mirage	31	9	34
28	Monte Carlo	10	5	72
29	Oppenheim	13	9	41
30	Out of Control	14	8	18
31	Pelican	24	7	63
32	Purple Haze	11	8	27
33	Raam	8	5	58
34	Rhea	19	6	52
35	Scorpion	32	12	44
36	The Little Mermaid	18	7	49
37	Tortuga	10	6	29
38	Urania	14	7	45
39	Wink	17	10	46
40	Xanadu	21	14	50
41	Zebra	29	10	71
42	Zenith	21	6	30
43	Zodiac	22	7	20
44	Zombie	25	10	34

Web Clustering Engines:

- Issues
- Systems
- Problems
- *New Directions*

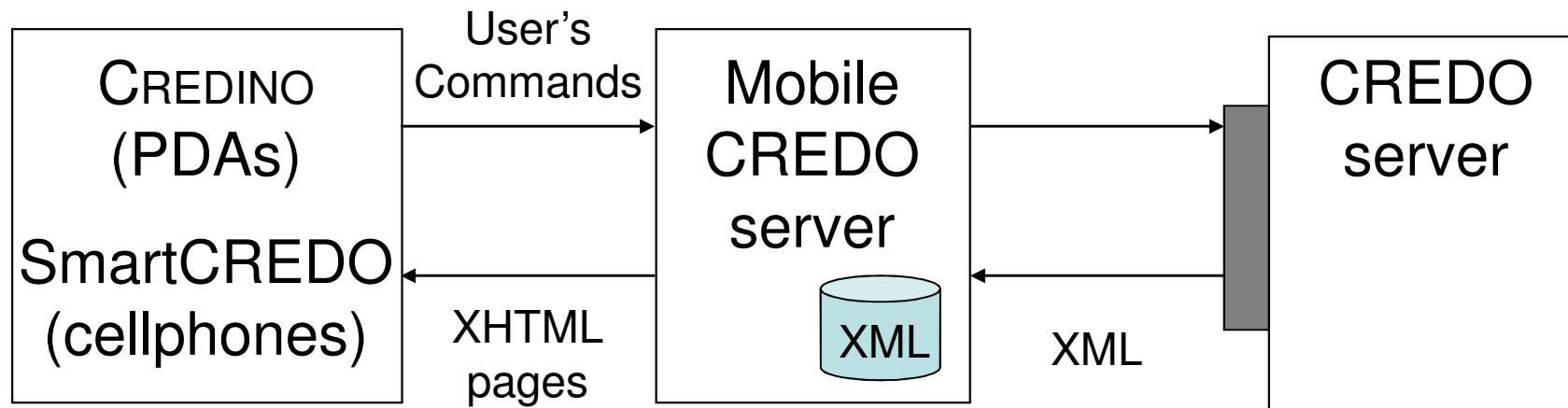
Research directions

- More powerful search results indexing
- Multiple clustering methods combination
- Generation of more informative cluster labels
- Personalized clustering creation/reorganization
- Integration with ontologies

Potentials of mobile Web clustering engines

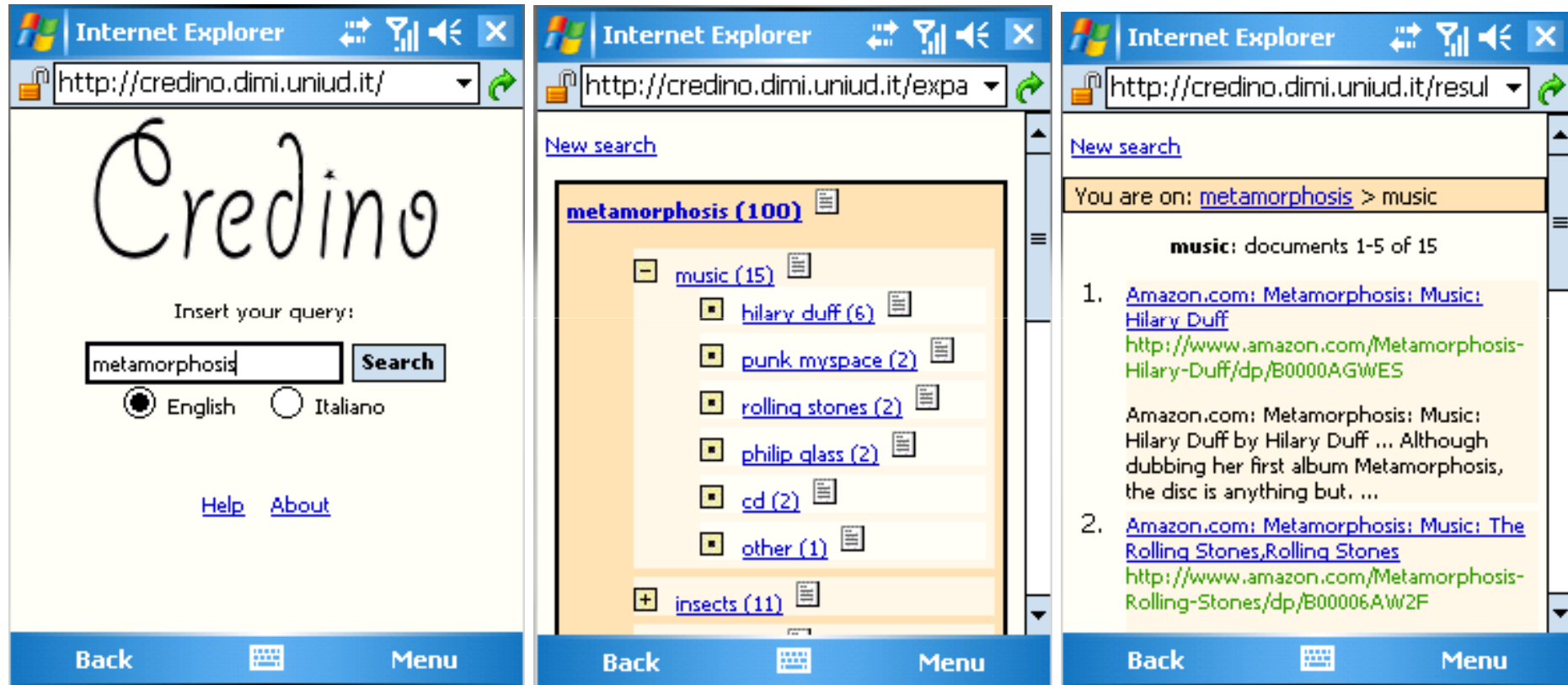
- Reduction of scrolling
- Reduction of typing for query refinement
- Reduction of downloaded data
- Extending mobile usage patterns to domain exploration

Mobile CREDO Architecture

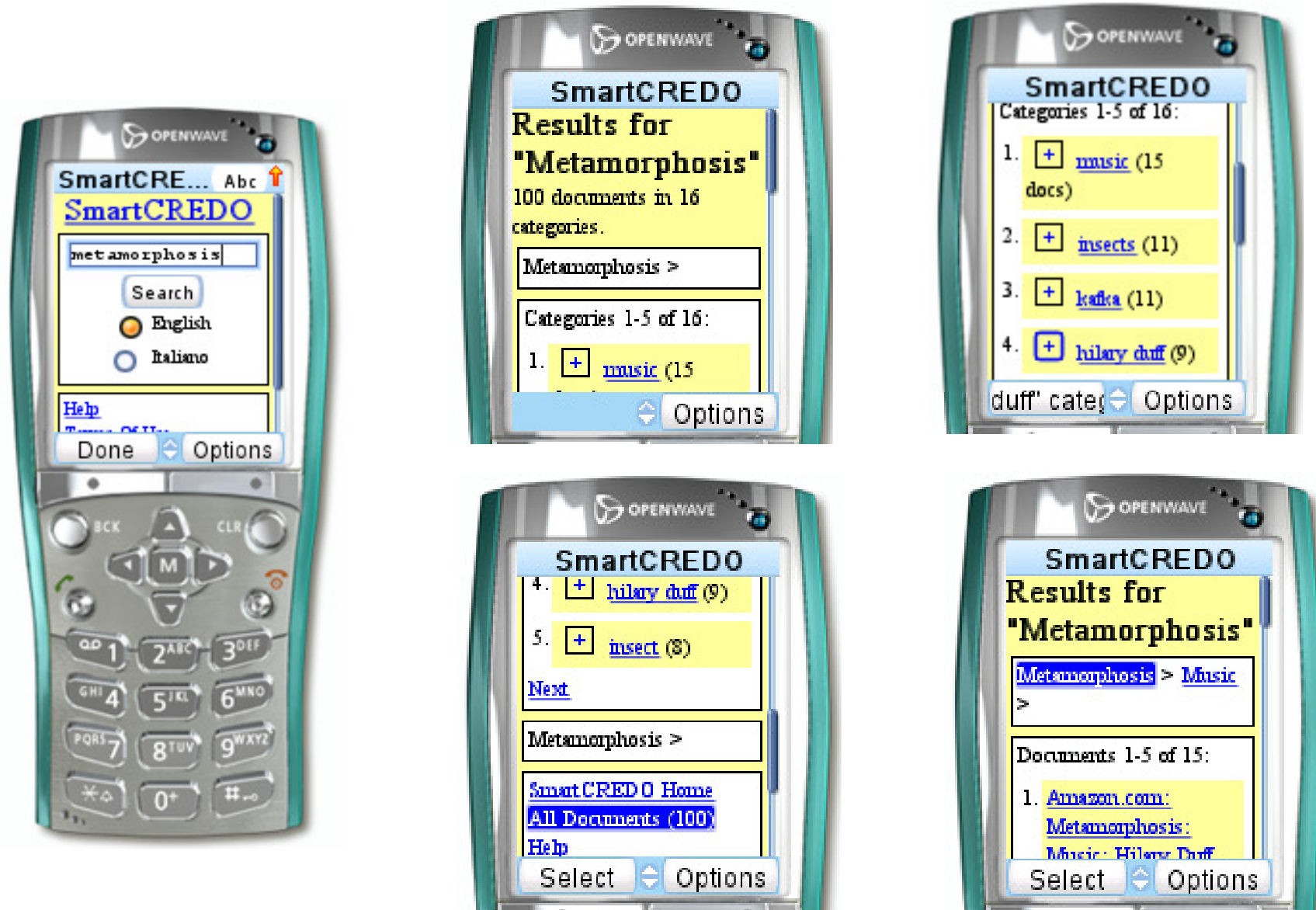


- Built on CREDO
- Bandwidth saving
- GUI for small screen

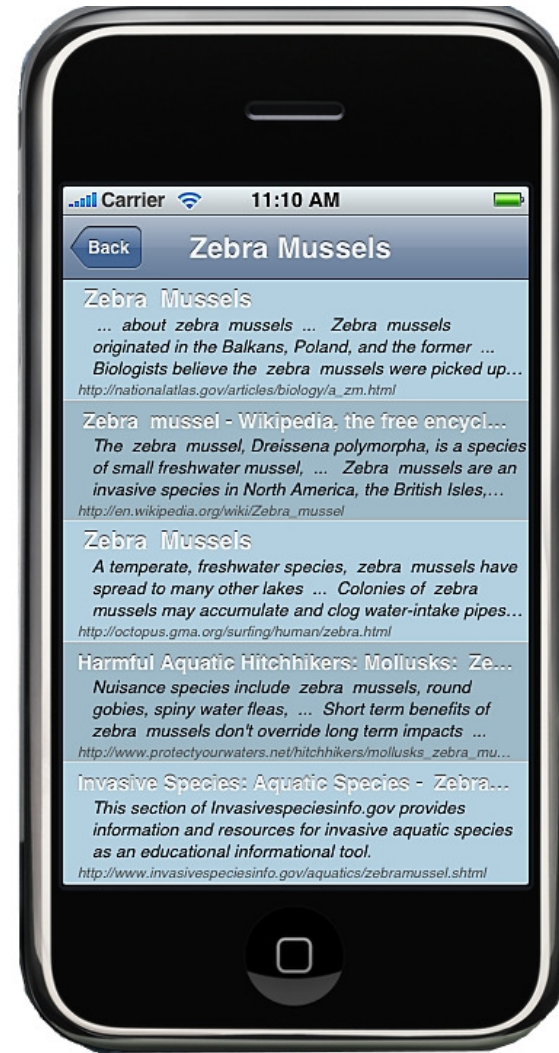
Credino (for PDAs)



SmartCREDO (for cellphones)



KeySRC on iPhone



Web search results clustering at Fondazione Ugo Bordoni: papers and relevant URLs

- C. Carpineto, S. Osinski, G. Romano, G. Weiss (to appear). A Survey of Web Clustering Engines. To appear in ACM Computing Surveys.
- Carpineto, S. Mizzaro, G. Romano, M. Snidero (2009). Mobile Information Retrieval with Search Results Clustering: Prototypes and Evaluations. JASIST, 60(5), 877-895.
- A. Bernardini, C. Carpineto, M. D'Amico (submitted). Full-Subtopic Retrieval with Keyphrase-based Search Results Clustering. Submitted.
- Carpineto, C., Della Pietra, A., Mizzaro, S., and Romano, G. (2006). Mobile Clustering Engine. Proceedings of ECIR 2006.
- Carpineto, C., Romano, G.. (2004). Concept Data Analysis: Theory and Applications. John Wiley & Sons.
- CREDO <http://credo.fub.it>
- Credino <http://credino.dimi.uniud.it>
- SmartCREDO <http://credino.dimi.uniud.it>
- AMBIENT <http://credo.fub.it/ambient>
- KeySRC <http://keysrc.fub.it>