

Introduction to Web Clustering

D. De Cao R. Basili

Corso di *Web Mining e Retrieval*
a.a. 2008-9

June 26, 2009

Outline

- ▶ Introduction to Web Clustering
- ▶ Some Web Clustering engines
- ▶ The KeySRC approach
- ▶ Some tools for build a Web Clustering engine
 - ▶ Yahoo Search API
 - ▶ CLUTO - Family of Data Clustering Software Tools

Web data clustering - Basics

- ▶ Organize data circulated over the Web into groups / collections in order to facilitate data availability & accessing, and at the same time meet user preferences.
- ▶ The initial idea was to define the correlation distance / similarity measure between any two “elements”.

Why use Web Clustering?

- ▶ *Increasing* Web information accessibility
- ▶ *Decreasing* lengths in Web navigation pathways
- ▶ *Improving* Web users requests servicing
- ▶ *Improving* information retrieval
- ▶ *Improving* content delivery on the Web
- ▶ *Understanding* users' navigation behavior
- ▶ *Integrating* various data representation standards
- ▶ *Extending* current Web information organizational practices

Web Directories vs. Web Clustering

Web Directory:

represent a widespread scenario where the most relevant web pages are classified with respect to a predefined set of categories organized into a hierarchy.

Google, Yahoo! are well known examples of such hierarchical organization of knowledge.

The Open Directory Project:

ODP, also known as **Dmoz** (from *directory.mozilla.org*, its original domain name), is a multilingual open content directory of *World Wide Web* links owned by Netscape that is constructed and maintained by a community of volunteer editors.

Web Directories vs. Web Clustering

Open Directory Project

ODP - Open Directory Project

dmoz open directory project

In partnership with
AOL search

[about dmoz](#) | [dmoz blog](#) | [suggest URL](#) | [help](#) | [link](#) | [editor login](#)

Search [advanced](#)

[Arts](#)

[Movies](#), [Television](#), [Music](#)...

[Games](#)

[Video Games](#), [RPGs](#), [Gambling](#)...

[Kids and Teens](#)

[Arts](#), [School Time](#), [Teen Life](#)...

[Reference](#)

[Maps](#), [Education](#), [Libraries](#)...

[Shopping](#)

[Clothing](#), [Food](#), [Gifts](#)...

[World](#)

[Català](#), [Dansk](#), [Deutsch](#), [Español](#), [Français](#), [Italiano](#), [日本語](#), [Nederlands](#), [Polski](#), [Русский](#), [Svenska](#)...

[Business](#)

[Jobs](#), [Real Estate](#), [Investing](#)...

[Health](#)

[Fitness](#), [Medicine](#), [Alternative](#)...

[News](#)

[Media](#), [Newspapers](#), [Weather](#)...

[Regional](#)

[US](#), [Canada](#), [UK](#), [Europe](#)...

[Society](#)

[People](#), [Religion](#), [Issues](#)...

[Computers](#)

[Internet](#), [Software](#), [Hardware](#)...

[Home](#)

[Family](#), [Consumers](#), [Cooking](#)...

[Recreation](#)

[Travel](#), [Food](#), [Outdoors](#), [Humor](#)...

[Science](#)

[Biology](#), [Psychology](#), [Physics](#)...

[Sports](#)

[Baseball](#), [Soccer](#), [Basketball](#)...

[Become an Editor](#) Help build the largest human-edited directory of the web



Copyright © 1998-2009 Netscape

4,616,780 sites - 83,449 editors - over 590,000 categories

Web Directories vs. Web Clustering

Open Directory Project

Open Directory – Health: Fitness: ...

 open directory project In partnership with
AOL  search

[about dmoz](#) | [dmoz blog](#) | [suggest URL](#) | [update listing](#) | [become an editor](#) | [report abuse/spam](#) | [help](#)

the entire directory ▾

Top: Health: Fitness: News and Media (25)

[Description](#)

- [Magazines and E-zines \(19\)](#)

See also:

- [Health: News and Media \(225\)](#)

This category in other languages:

[Italian \(5\)](#)

- [AskMen.com](#) - A two part article with tips on choosing the right gym.
- [Body-Mind-Spirit Conference - Pilates, Yoga, Gyrotonic & NTA](#) - Welcome to the Body Mind Expo educational conference.
- [CNN's Health News: Diet and Fitness](#) - Offering current health news, tips, on-line encyclopedia and many fitness related links.
- [Fitness-Events.Com](#) - Fitness and figure news, reports, and photos.
- [Health & Fitness Business Expo and Conference](#) - Annual show focuses on new fitness products and industry trends. Site contains information on the event, sponsorship, registration, press releases.
- [Revolution Health: Fitness](#) - Provides articles, forums, medical advice, and consumer reviews.

- "News and Media" search on: [AltaVista](#) - [A9](#) - [AOL](#) - [Ask](#) - [Clusty](#) - [Gigablast](#) - [Google](#) - [Lycos](#) - [MSN](#) - [Yahoo](#)

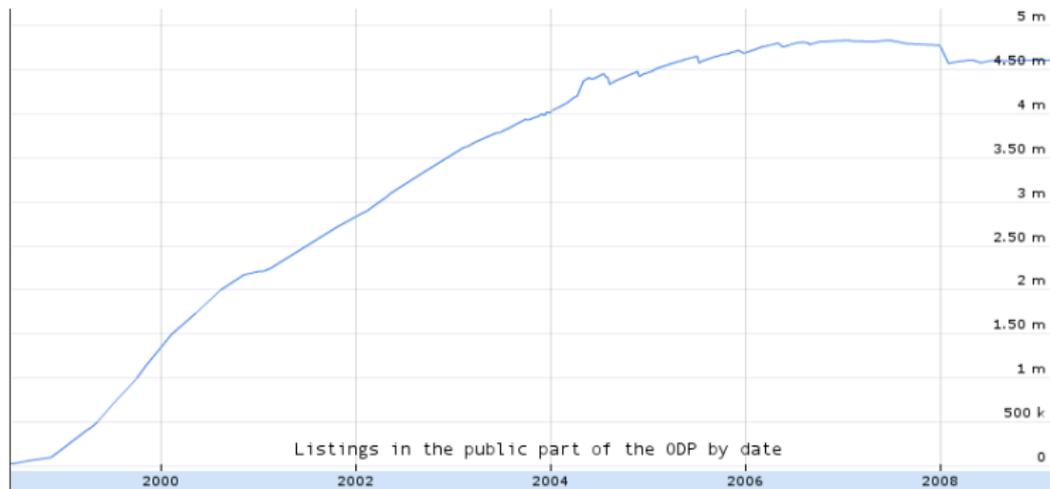


[Volunteer](#) to edit this category.



Web Directories vs. Web Clustering

Open Directory Project



Web Directories vs. Web Clustering

- ▶ Web Directories are based on taxonomies.
- ▶ Web Directories are *static* view of WWW.
- ▶ Extend Web Directories is a *classification* problem.

- ▶ Web Clustering is totally unsupervised.
- ▶ Clusters are dynamically generated on user needs.
- ▶ Filtering out irrelevant results.
- ▶ Need to define a label for each cluster.

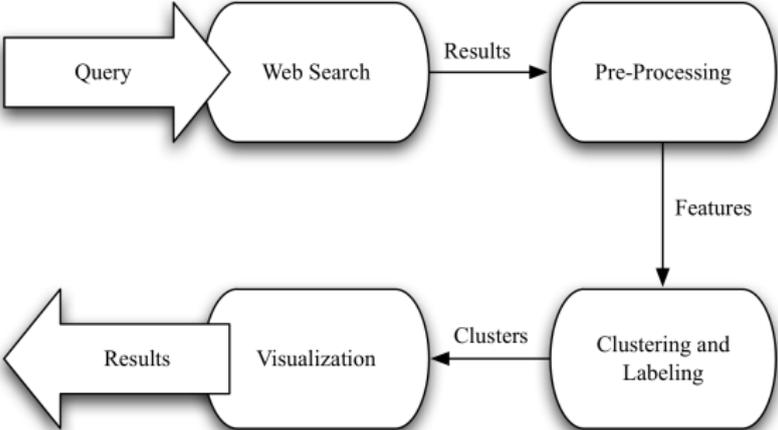
Issues for Web Clustering

- ▶ Representation for clustering
 - ▶ How represent Document?
 - ▶ Full documents or snapshot?
 - ▶ Need a notion of similarity/distance
- ▶ How many clusters?
 - ▶ Fixed a priori?
 - ▶ Completely data driven?
 - ▶ Avoid “trivial” clusters - too large or small

Classic Document Clustering vs. Web Clustering

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

Web Clustering Architecture



Web Search API

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).



[web](#) [news](#) [images](#) [wikipedia](#) [blogs](#) [jobs](#) [more »](#)

Tor Vergata

[advanced preferences](#)

clusters
sources
sites

All Results (224)
remix

- [Dipartimento, Sito ufficiale](#) (34)
- [Ibis, Hotel](#) (33)
- [University of Rome Tor Vergata](#) (28)
- [Università degli Studi](#) (14)
- [Mathematics Dept](#) (9)
- [Università di Tor Vergata](#) (11)
- [CEIS, Paper](#) (8)
- [Policlinico](#) (8)
- [Studenti](#) (5)
- [INFN Tor Vergata](#) (4)

[more](#) | [all clusters](#)

find in clusters:

Find

Font size: A A A A

Top 224 results of at least 521,000 retrieved for the query **Tor Vergata** ([details](#))

Did you mean: [Tor vergara](#)

Search Results

1. [Best Hotels - Εύ+οέα Ιοάεε Δεία](#)   
TOR VERGATA Αίηρόιέε00 Δεία ... Italy dips down out of Europe and into the Mediterranean like a womens leg firmly ...
[www.tor-vergata.com](#) - [cache] - Live, Open Directory, Ask, Gigablast
2. [Home page del sito dell'università degli Studi di Roma "Tor Vergata"](#)   
 Università degli Studi di Roma "**Tor Vergata**" L'Ateneo che costruisce il domani
[web.uniroma2.it](#) - [cache] - Live, Ask, Gigablast
3. [Università degli Studi di Roma Tor Vergata](#)   
 Sito principale dell'Ateneo.
[www.uniroma2.it](#) - [cache] - Open Directory, Gigablast
4. [CEIS - CEIS - CENTRE FOR ECONOMICS AND INTERNATIONAL STUDIES](#)   
Tor Vergata Economic Foundation and CEIS - Centre for Economic and International Studies
 University of Rome "**Tor Vergata**" - June 24 - 25, 2009
[www.ceistorvergata.it](#) - [cache] - Live, Open Directory, Ask, Gigablast
5. [University of Rome Tor Vergata - Wikipedia, the free encyclopedia](#)   
 The University of Rome **Tor Vergata** (Italian: Università degli Studi di Roma **Tor Vergata**) is a university located in Rome, Italy, and founded in 1982. The university occupy a large ...
[en.wikipedia.org/wiki/University_of_Rome_Tor_Vergata](#) - [cache] - Live, Ask
6. [Santiago Calatrava - Tor Vergata University \(Roma II\) :: arcSPACE.com](#)   
 Ground has been broken for a new Campus Master Plan, Sports City and Rectorate Tower.
[www.arcSPACE.com/architects/calatrava/tor_vergata/tor_vergata.html](#) - [cache] - Live, Ask
7. [Default Empty Site](#)   
 Empty Site
[www.torvergata.it](#) - [cache] - Live, Gigablast
8. [ART: Artificial Intelligence Research @ Tor Vergata](#)   

The screenshot shows the Carrot search engine interface. At the top, there is a navigation bar with links for 'About', 'New featured!', 'Search feeds', 'Search plugins', 'Download', 'Carrot Search', and 'Contact'. Below this is a search bar containing the text 'rome' and a 'Search' button with a 'More options' link. The left sidebar features a 'Tree' view with a 'Visualization' tab. The tree lists various categories: 'All Topics (100)', 'Rome Italy (16)', 'Hotels Hotels (13)', 'Rome Apartments (7)', 'Ancient Rome (6)', 'History (6)', 'Travel Guide (6)', 'Rome (5)', 'Pope (4)', 'Empire (3)', and 'Images (3)'. A 'more | show all' link is at the bottom of the tree. The main content area displays 'Top 100 results of about 106000000 for rome'. The first five results are listed:

- Rome - Wikipedia, the free encyclopedia**
The metropolitan area of **Rome** is estimated by OECD to have a population of 3.7 million. It is located in the central-western portion of the Italian ...
<http://en.wikipedia.org/wiki/Rome> [Ask, Bing, Exalead, Google, Wikipedia, Yahoo]
- Rome (TV series) - Wikipedia, the free encyclopedia**
Rome is an American-British-Italian historical drama television series created by Bruno Heller, John Milius, and William J. MacDonald. ...
[http://en.wikipedia.org/wiki/Rome_\(TV_series\)](http://en.wikipedia.org/wiki/Rome_(TV_series)) [Ask, Wikipedia, Google]
- ItalyGuides.it: Virtual tour of Rome, travel information and city ...**
Rome tourism and travel information: transport, attractions, maps, travel advice , pictures, audio guides, airport information, activities, hotels and more ...
http://www.italyguides.it/us/roma/rome_italy_travel.htm [Ask, Google]
- Rome Travel Information and Travel Guide - Italy - Lonely Planet**
Rome tourism and travel information such as accommodation, festivals, transport, maps, activities and attractions in **Rome**, Italy - Lonely Planet.
<http://www.lonelyplanet.com/italy/rome> [Ask, Google]
- Romaturismo**
<http://www.romaturismo.it/v2/en/main.asp> [Ask, Google]

The sixth result is partially visible:

- CATHOLIC ENCYCLOPEDIA: Rome**
The significance of **Rome** lies primarily in the fact that it is the city of the pope.
<http://www.newadvent.org/cathen/13164a.htm> [Ask, Bing, Google]

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

GROK

[Search Options](#)

Refine Search

by keyword

 exclude

by date

 to

by source

by domain

 Outline View

 Map View

142 total results

 Working List (0 items)

 Email

 Export

[Expand Outline](#) | [Collapse Outline](#)

 Results: [1-10 of 142] << < > >> [Show All](#)

 Detail: [Less](#) [Medium](#) [More](#)
clustering (142 results)

- Server (11)
- Clustering Algorithms (12)
- Data Clustering (11)
- Application (9)
 - Hierarchical Clustering (8)
 - Clustering Algorithm (5)
 - High Availability (5)
 - Document Clustering (5)
 - Linux (5)
 - Words (5)
 - Article (5)
 - Dictionary (4)
 - Storage Clustering (4)
 - Books (4)
 - Tutorial (4)
 - Method (4)
 - Performance (4)
 - Clustering Basics (3)
 - Review (3)
 - Spectral Clustering (3)
 - Process (3)
 - Problem (3)
 - Clustering Architecture (3)

clustering

[Cluster \(computing\) - Wikipedia, the free encyclopedia](#)
[Add to Working List](#) | [Post to del.icio.us](#) | [Email](#)

Clustering can provide significant performance benefits versus price. ... The first commercial clustering product was ARCnet, developed by Datapoint in 1977. ... http://en.wikipedia.org/wiki/Computer_cluster - 11 giugno 2009
Source: Yahoo!

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

...Cluster analysis, a method for statistical data analysis ... http://en.wikipedia.org/wiki/Cluster_analysis - 24 giugno 2009
Source: Wikipedia

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

Scott Schnolf's Microsoft Cluster Center is the ultimate source for Microsoft Clustering information and answers. ... Clustering can help reduce both planned ... <http://www.nwnetworks.com/basics.htm> - 17 ottobre 2005
Source: Yahoo!

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

...Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a U.S. registered 501(c)(3) tax-deductible nonprofit charity. ... http://en.wikipedia.org/wiki/Non-profit_organization - 24 giugno 2009
Source: Wikipedia

[Category:All disambiguation pages](#)
[Add to Working List](#) | [Post to del.icio.us](#) | [Email](#)

All disambiguation pages http://en.wikipedia.org/wiki/Category:All_disambiguation_pages - 24 giugno 2009
Source: Wikipedia





zebra

Search

All results (100)

- » plains zebra (8)
- » zebra technologies (6)
- » zebra species (5)
- » african mammals of the genus equus (4)
- » thermal barcode label printers (5)
- » grevy's zebra (5)
- » zebra mussels (4)
- » zebra print (3)
- » zebra printers (5)
- » burchell s zebra (3)

[More clusters](#)

YOU ARE IN "ZEBRA SPECIES" CLUSTER WITH 5 DOCUMENTS

AWF: WILDLIFE: ZEBRA

Three species of **zebra** still occur in Africa, two of which are found in East Africa. ... The other is the Grevy's **zebra**, named for Jules Grevy, a president of ...

<http://www.awf.org/content/wildlife/detail/zebra>

AFRICAN SPECIES GUIDE - MAMMALS - ZEBRA

Characteristic horse of the African plains, the Burchell_s **Zebra** is the only **zebra** species where the black stripes extend onto the stomach.

<http://www.africanwildlifeguide.com/species-guide/mammals/large-mammals/zebra>

ZEBRA | AFRICAM

Zebra are grigarious animals who are vocal and ready to move off with much speed ... **Zebra** species can also be distinguished from one another by virtue of ...

<http://www.africam.com/wildlife/zebra>

ZEBRA MUSSELS

A temperate, freshwater species, **zebra** mussels have spread to many other lakes ... Colonies of **zebra** mussels may accumulate and clog water-intake pipes and screens ...

<http://www.gma.org/surfing/human/zebra.html>

ZEBRA INFORMATION AT ANIMALS ON RUGS

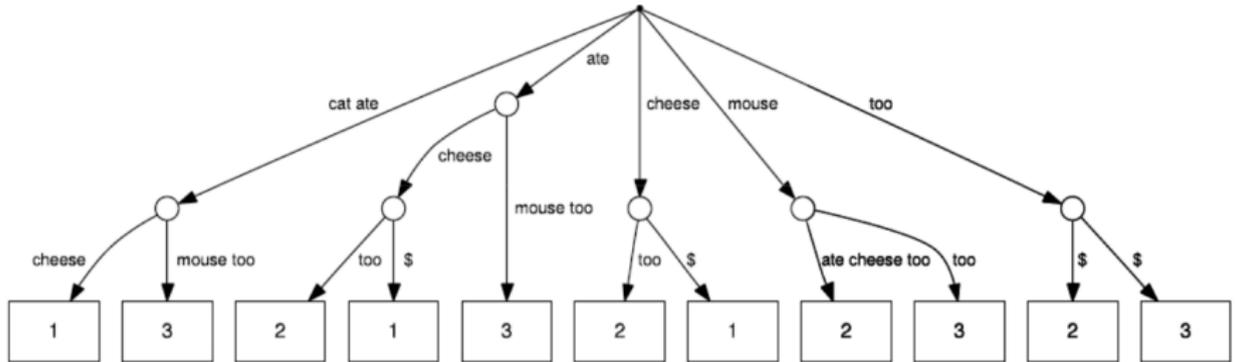
A short summary of information about: **zebra** species from Animals on Rugs, where you can get top quality **zebra** hide rugs for your own home decor. ... **Zebra** Rugs ...

<http://www.animalsonrugs.com/site/890202/page/467392>

Some 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

Generalized suffix tree (from Zamir and Etzioni, 1998)

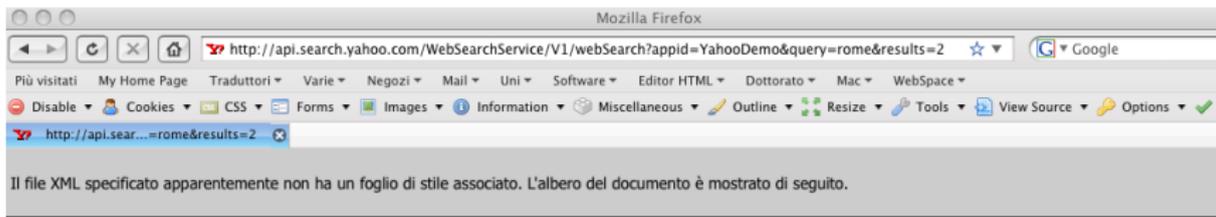


- 1) Cat ate cheese
- 2) Mouse ate cheese too
- 3) Cat ate mouse too

The KeySRC algorithm

1. Search results preprocessing
2. Construction of Generalized Suffix Tree (GST)
3. Extraction of keyphrases from GST Extraction of keyphrases from GST
(internal nodes of GST + ≤ 4 words + POS tagging)
4. Keyphrases clustering and Label assignment
5. Cluster ranking

Yahoo! search apis Example



```
- <ResultSet xsi:schemaLocation="urn:yahoo:srch http://api.search.yahoo.com/WebSearchService/V1/WebSearchResponse.xsd" type="web"
totalResultsAvailable="37500000" totalResultsReturned="2" firstResultPosition="1" moreSearch="/WebSearchService/V1/webSearch?query=rome&
appid=YahooDemo&region=us">
- <Result>
  <Title>Rome, Italy - Wikipedia</Title>
- <Summary>
  Includes history, geography, climate, economy, demographics, religion, culture, transportation, events, sister cities, and references about the Italian capital, Rome.
</Summary>
<Url>http://en.wikipedia.org/wiki/Rome</Url>
<ClickUrl>http://en.wikipedia.org/wiki/Rome</ClickUrl>
<DisplayUrl>en.wikipedia.org/wiki/Rome</DisplayUrl>
<ModificationDate>1244876400</ModificationDate>
<MimeType>text/html</MimeType>
- <Cache>
- <Url>
  http://uk.wrs.yahoo.com/_ylt=A0WTecwZKENkww8wABzPdmMwF;_ylu=X3oDMTBwZTdwbWtkBGNvbG8DZQRwb3MMDMQRzZWMDc3IEdnRpZAM-/SIG=15lbig26q
/EXP=1246001561/**http%3A//74.6.239.67/search/cache%3Ffei=UTF-8%26appid=YahooDemo%26query=rome%26results=2%26u=en.wikipedia.org
/wiki/Rome%26w=rome%26d=aB1snRIMS-4f%26icp=1%26.intl=us
</Url>
<Size>268140</Size>
</Cache>
</Result>
```

CLUTO: Clustering High-Dimensional Datasets

About CLUTO

It is a software package for clustering low- and high-dimensional datasets and for analyzing the characteristics of the various clusters.

Consists of both stand-alone programs and a library via which an application program can access directly the various clustering and analysis algorithms implemented in CLUTO.

- ▶ Multiple classes of clustering algorithms:
 - ▶ partitional, agglomerative and graph-partitioning based.
- ▶ Multiple similarity/distance functions:
 - ▶ Euclidean distance, cosine, correlation coefficient, extended Jaccard, user-defined.
- ▶ Numerous novel clustering criterion functions and agglomerative merging schemes.
- ▶ Traditional agglomerative merging schemes:
 - ▶ single-link, complete-link, UPGMA
- ▶ Extensive cluster visualization capabilities and output options:
 - ▶ postscript, SVG, gif, xfig, etc.
- ▶ Multiple methods for effectively summarizing the clusters:
 - ▶ most descriptive and discriminating dimensions, cliques, and frequent itemsets.
- ▶ Can scale to very large datasets containing hundreds of thousands of objects and tens of thousands of dimensions.