

Clustering Search Results with Carrot²

Stanisław Osiński¹ Dawid Weiss²

Poznan Supercomputing and Networking Center,
ul. Noskowskiego 10, 61-704, Poznan, Poland
`stanislaw.osinski@man.poznan.pl`

Institute of Computing Science, Poznan University of Technology,
ul. Piotrowo 3A, 60-965 Poznań, Poland
`dawid.weiss@cs.put.poznan.pl`

January 2007

About the authors

Dawid Weiss

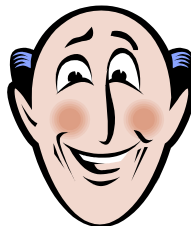
- entertainment experience
- MSc in Software Engineering
- industrial experience
- currently in academia
- PhD in Information Retrieval

Stanisław Osiński

- MSc in Design of IT Systems
- industrial/ research experience
- employed in a research institute
- PhD in progress. . .

- 1 Introduction to Search Results Clustering
- 2 Carrot² Framework
- 3 Lingo clustering algorithm
- 4 Summary

Ranked lists are not perfect



Clustering Search Results...

└ Introduction to Search Results Clustering

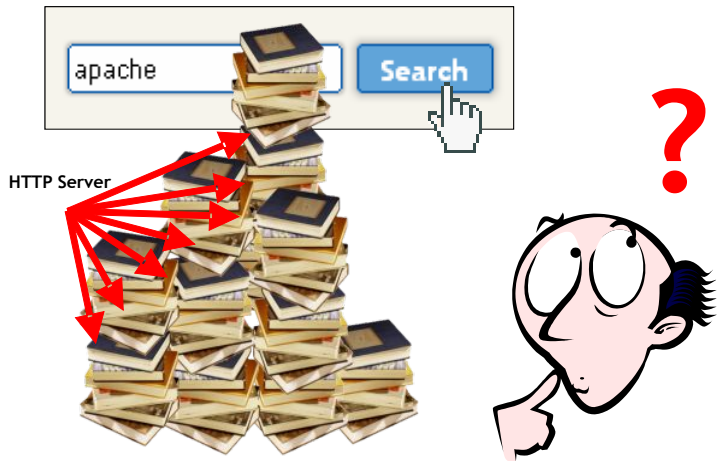
└ **Ranked lists** are not perfect

Ranked lists are not perfect

Search results clustering is one of many methods that can be used to **improve user experience while searching collections of text documents**, web pages for example. To illustrate the problems with conventional ranked list presentation, let's imagine a user wants to find web documents about “apache”. Obviously, this is a very general query, which can lead to...

Ranked lists are not perfect



2007-01-22

Clustering Search Results...

└ Introduction to Search Results Clustering

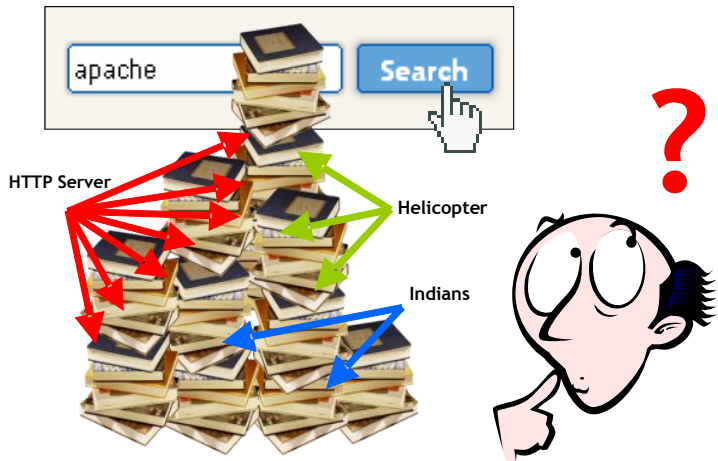
└ **Ranked lists** are not perfect

Ranked lists are not perfect



... large numbers of references being returned, the majority of which will be about the Apache Web Server.

Ranked lists are not perfect



Clustering Search Results. . .

└ Introduction to Search Results Clustering

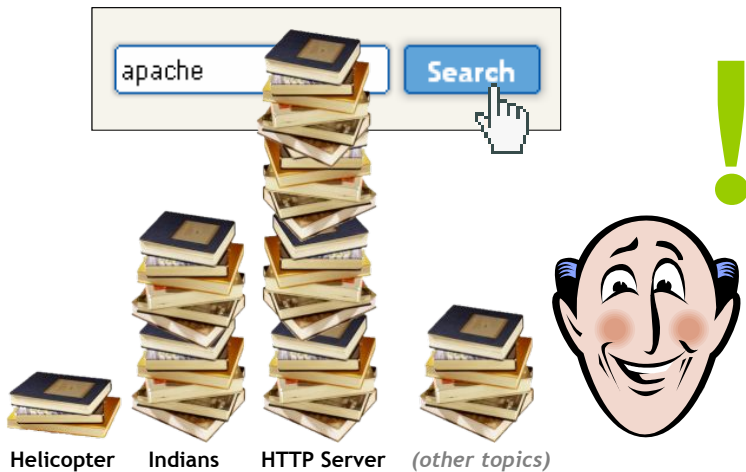
└ **Ranked lists** are not perfect

Ranked lists are not perfect



A more patient user, a user who is determined enough to look at results at rank 100, should be able to reach some scattered results about the Apache Helicopter or Apache Indians. As you can see, one problem with ranked lists is that **sometimes users must go through many irrelevant documents** in order to get to the ones they want.

Search Results Clustering can help



Clustering Search Results. . .

- └ Introduction to Search Results Clustering

- └ **Search Results Clustering** can help








Search Results Clustering can help

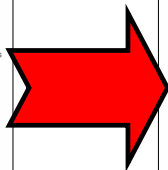

























So how about an interface that groups the search results into separate semantic topics, such as the Apache Web Server, Apache Indians, Apache Helicopter and so on? With such groups, the user will immediately get an overview of what is in the results and should be able to navigate to the interesting documents with less effort.

This kind of interface to search results can be implemented by applying a document clustering algorithm to the results returned by the search engine. This is something that is commonly called Search Results Clustering.

Search Results Clustering is an interesting problem

1. [Apache Software Foundation](#) 
Membership-based, not-for-profit corporation that exists to provide organizational, legal, and financial support for the **Apache** open-source software projects.
Category: [Unix Servers > Apache](#)
[www.apache.org](#) - [More from this site](#) - [Save](#)
2. [Apache HTTP Server Project](#) 
Effort to develop and maintain an open-source HTTP server for modern operating systems including UNIX and Windows NT.
[http://apache.org](#) - 32k - [Cached](#) - [More from this site](#) - [Save](#)
3. [Download - The Apache HTTP Server Project](#) 
Essentials. Download! Get Involved. Subprojects. Use the links below to download the **Apache** HTTP Server from one of our mirrors. You must verify the integrity of the downloaded files using signatures downloaded from our main distribution directory.
[comhttp://apache.mirrors.versehost.comhttp://apache.forbigweb.comhttp://apache.mirrors.hoobly.com](#) ...
[http://apache.org/download.cgi](#) - 17k - [Cached](#) - [More from this site](#) - [Save](#)
4. [Apache.com - Providing Web Server and Network Security Resources](#) 
Web Hosting Control Panel. Control panels provide a convenient interface with which to create new users, assign admin privileges, bind DNS entries, control servers, sites, and more. ... One ADC, Zeus Extensible Traffic Manager (ZXTM) can accelerate **Apache** by as much as 100 times ... The **Apache** HTTP Server Project develops and maintains an open-source HTTP server for ...
[www.apache.com](#) - 15k - [Cached](#) - [More from this site](#) - [Save](#)
5. [The Jakarta Site - The Jakarta Project -- Java Related Products](#) 
... solutions and is a part of The **Apache** Software Foundation (ASF) which encourages a collaborative, consensus ... Copyright © 1999-2005, The **Apache** Software Foundation. Legal information ...
Category: [Unix Servers > Apache](#)
RSS: [View as XML](#) - [Add to My Yahoo!](#)
[jakarta.apache.org](#) - [More from this site](#) - [Save](#)
6. [Electronics Manufacturer Irvine Analog Broadband Modems Mp3 Wireless Communications California](#) 
... **Apache** Micro located in Irvine California is one of the world's leading analog modem and broadband ... manufacturing broadband and analog modems, **Apache-Micro** also produces a wide range ...
[www.apache-micro.com](#) - 11k - [Cached](#) - [More from this site](#) - [Save](#)
7. [Apache XML Project](#) 
To provide commercial-quality standards-based XML developed in an open and cooperative fashion and to provide feedback to standards bodies.
Category: [Unix Servers > Apache](#)
[xml.apache.org](#) - 31k - [Cached](#) - [More from this site](#) - [Save](#)
8. [Apache HTTP Server Version 2.0 Documentation - Apache HTTP Server](#) 
Apache HTTP Server Version 2.0. **Apache** HTTP Server Version 2.0 Documentation. Copyright 1995-2005 The **Apache** Software Foundation or its licensors, as applicable. Licensed under the **Apache** License, Version 2.0.
[http://apache.org/docs/2.0](#) - 7k - [Cached](#) - [More from this site](#) - [Save](#)



-   **Apache HTTP Server (97)**
-   **Apache Software Foundation (31)**
-   **Apache Web Server (31)**
-   **Open Source (12)**
-   **Apache License (12)**
-  **Apache XML (6)**
-  **Feature Articles (5)**
-  **Apache Nation (5)**
-  **Free Encyclopedia (4)**
-  **Apache Ant (4)**
-  **Apache FOP (3)**
-  **Apache Forrest (3)**
-  **Apache Struts Project (2)**
-  **AH Apache (2)**
-  **Apache Geronimo (2)**
-  **Gentoo Linux Documentation (2)**
-  **Cross Site Scripting Info (2)**
-  **(Other topics) (32)**

Clustering Search Results. . .

- └ Introduction to Search Results Clustering

- └ Search Results Clustering is an **interesting problem**

Search Results Clustering has a few interesting characteristics and one of them is the fact that it is **based only on the fragments of documents returned by the search engine** (document snippets). This is the only input an algorithm has, full text of documents is not available.

Search Results Clustering is an interesting problem

2. [Apache HTTP Server Project](#)

Effort to develop and maintain an open-source HTTP server for modern operating systems including UNIX and Windows NT.

<http://apache.org> - 32k - [Cached](#) - [More from this site](#) - [Save](#)

49. [pre-FAQ - The Apache Software Foundation](#)

... most of the common queries that we receive about our software and the **Apache** Software Foundation ... (or something similar indicating that **Apache** has been installed) on your screen ...

www.apache.org/foundation/preFAQ.html - 32k - [Cached](#) - [More from this site](#) - [Save](#)

587. [Apache C++ Standard Library](#)

Last Modified: \$Date: 2006-02-16 09:05:15 -0800 (Thu, 16 Feb 2006) \$ stdcxx. STDCXX: **Apache** C++ Standard Library. What is stdcxx? ... The goal of the **Apache** C++ Standard Library project is to provide a free implementation of the ISO ... C++ Standard Library to the **Apache** stdcxx project, a proven code base ...

incubator.apache.org/stdcxx - 41k - [Cached](#) - [More from this site](#) - [Save](#)

Clustering Search Results...

└ Introduction to Search Results Clustering

└ Search Results Clustering is an **interesting problem**

Search Results Clustering is an interesting problem

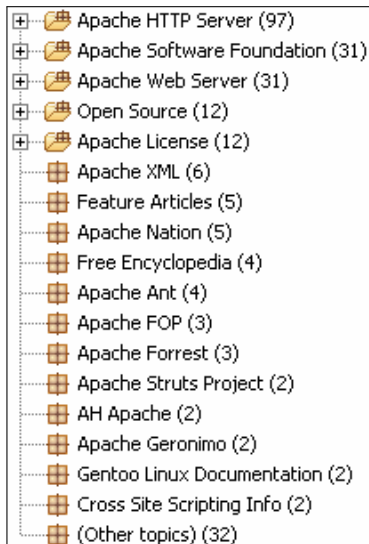
2. **Apache HTTP Server Project** 
 Client to develop and maintain an open-source HTTP server for modern operating systems including Unix and Windows NT.
<http://apache.org> - [FAQ](#) - [Contact](#) - [What's New](#) - [Site](#) - [Help](#)

45. **Apache/2.0.49 - The Apache Software Foundation** 
 Apache is a free and open-source HTTP server used by the vast majority of websites on the Internet. It was created by [Roberto Serot](#) and [Mark M. Beagrie](#) in 1995. Apache is the most popular web server in the world.
www.apache.org - [FAQ](#) - [Contact](#) - [What's New](#) - [Site](#) - [Help](#)

107. **Apache** 
 Get Apache [Source](#) - [FAQ](#) - [Contact](#) - [What's New](#) - [Site](#) - [Help](#)
 Apache is a free and open-source HTTP server used by the vast majority of websites on the Internet. It was created by [Roberto Serot](#) and [Mark M. Beagrie](#) in 1995. Apache is the most popular web server in the world.
www.apache.org - [FAQ](#) - [Contact](#) - [What's New](#) - [Site](#) - [Help](#)

Document snippets returned by search engines are usually very short and noisy. So we can get broken sentences or useless symbols, numbers or dates in the input.

Search Results Clustering is an **interesting** problem



- **Semantic** clusters
- **Meaningful** cluster labels
- **Small** input

Clustering Search Results...

└ Introduction to Search Results Clustering

└ Search Results Clustering is an **interesting problem**

Search Results Clustering is an interesting problem

11	Apache HTTP Server (37)
11	Apache Software Foundation (31)
11	Apache Web Server (31)
11	Open Source (12)
11	Apache License (12)
11	Apache 1.0 (6)
11	Feature Articles (5)
11	Apache Nation (5)
11	Free Encyclopedia (4)
11	Apache Ant (4)
11	Apache FOP (2)
11	Apache Forrest (2)
11	Apache Struts Project (2)
11	4th Apache (2)
11	Apache Geomime (2)
11	Genio Linux Documentation (2)
11	Cross Site Scripting (XSS) (2)
11	(Other topics) (2)

- Semantic clusters
- Meaningful cluster labels
- Small input

In order to be helpful for the users, search results clustering must put results that deal with the same topic into one group. This is the primary requirement for all document clustering algorithms.

But in search results clustering very important are also the labels of clusters. We must **accurately and concisely describe the contents of the cluster**, so that the user can quickly decide if the cluster is interesting or not. This aspect of document clustering is sometimes neglected.

Finally, because the total size of input in search results clustering is small (e.g. 200 snippets), **we can afford some more complex processing**, which can possibly let us achieve better results.

Search Results Clustering is an **interesting** problem

...and that's why we created



- 1 Introduction to Search Results Clustering
- 2 Carrot² Framework
- 3 Lingo clustering algorithm
- 4 Summary

Carrot² is about **search results clustering**

Carrot² is a...

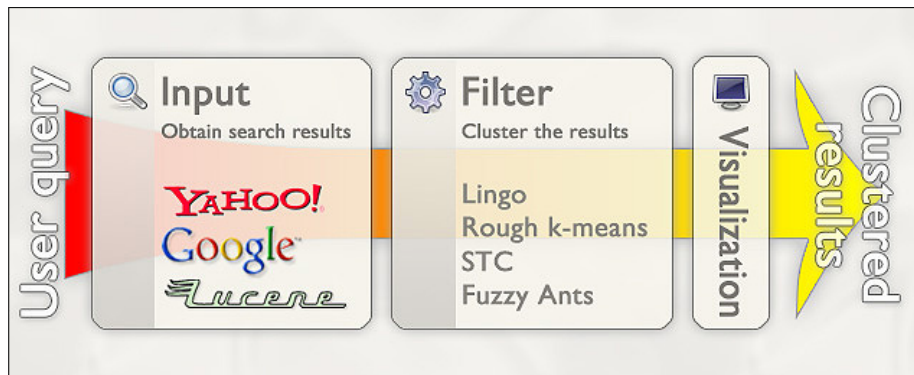
- framework for **experimenting** with processing and presentation of search results
- framework for building **real-world** production-quality applications
- BSD-licensed **open source** project



Carrot² targets **researchers, developers** and **end-users**



Carrot² is based on **processing pipelines**



Carrot² offers **ready-to-use components**: input

Fetching search results from:



Google (API)



Yahoo (API)



MSN (API)



Open Search



Lucene



ODP Project

Carrot² offers **ready-to-use components**: clustering

5 search results clustering algorithms:

- Lingo
(Stanisław Osiński)
- STC
(Oren Zamir, Oren Etzioni)
- Rough-KMeans
(Ngo Chi Lang)
- HAOG
(Karol Gołembniak, Irmina Maśłowska)
- FuzzyAnts
(Steven Schockaert)

Carrot² offers **ready-to-use components**: other

Other utilities:

- language tokenizers, stemmers and stop word lists
- very fast matrix computations library
- desktop browser application for tuning and rapid experiments

Query: data mining

Search



Process: Yahoo Search API -- Suffix Tree Clusterer

Settings

Results: 150

data mining

[Yahoo Search API -- Suffix Tree Clusterer] data mining



- Data Mining (135)
- Knowledge Discovery (17)
- Data Mining Tools (8)
- Data Mining and Knowledge Discovery (4)
- Data Mining Software (8)
- Data Mining Techniques (6)
- Business Intelligence (9)
- Software (19)
- Tools (18)
- Data Mining Applications (5)
- Data Mining : Concepts and Techniques (3)
- Machine Learning (6)
- Information (16)
- Knowledge Discovery in Databases (4)
- Data Mining Products (4)

[KDnuggets: Data Mining, Web Mining, and Knowledge Discovery Guide](#) [1] [en]

Newsletter on the data mining and knowledge industries, offering information on data mining, knowledge discovery, text mining, and web mining software, courses, jobs, publications, and meetings.
<http://www.kdnuggets.com/>

[Data Mining and Analytic Technologies \(Kurt Thearling\)](#) [3] [en]

Kurt Thearling's site dedicated to sharing information about data mining, the automated extraction of hidden predictive information from databases, and other analytic technologies.
<http://www.thearling.com/>

Process settings

Input preprocessing

Ignore word if in fewer docs:

2

Ignore word if in more docs (%):

0,9

Base clusters

Max base clusters:

300

Min base cluster score:

2

Min base cluster size:

2

Merging and output

Merge threshold:

Update settings

☒ Live update

Refresh

Query: data mining

Search



Yahoo Search API -- Suffix Tree Clusterer benchmark

Benchmark settings

JVM warm-up cycles

25

Benchmark cycles

75

Benchmark progress & results

Benchmarking...

Stop

Average time 234,08 ms Time std dev 97,1

Min time 109,00 ms Max time 438

Categories

All results	142
Data Mining	135
Knowledge Discov...	17
Data Mining Tools	8
Data Mining and Kno...	4
Data Mining Software	8
Data Mining Techni...	6
Business Intelligen...	9
Software	19
Tools	16
Data Mining Applicab...	5
Data Mining - Concept...	3
Machine Learning	6
Information	16
Knowledge Discovery...	4
Data Mining Products	4

Data Mining,
and
discovery

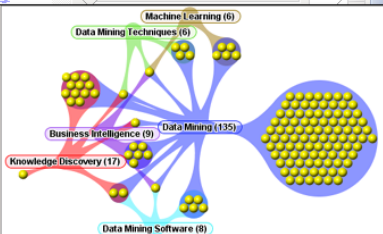
Process settings

Input preprocessing

Ignore word if in fewer docs:



2



Process settings

Input preprocessing

Ignore word if in fewer docs:



2

Ignore word if in more docs (%):



0,9

Thearing's site
ated to sharing
nation about data
g, the automated
ction of hidden
ctive information from
ases, and other
tic technologies.

<http://www.thearing.com/>

0 1 2 3 4 5 6 7 8 9 10

Min base cluster size:



2

2 6 10 14 18

Merging and output

Merge threshold:



Update settings

☒ Live update

Refresh

The desktop application allows detailed tuning of each algorithm. In the query panel we have options for:

- input/ algorithm selection,
- number of search results to fetch,
- default algorithm configuration settings.

After a query is performed, a **result tab** appears on screen allowing:

- benchmarking,
- visualization,
- on-line modification of algorithm parameters, reflected in the clusters panel.

[Y Yahoo!](#)[G Google](#)[MSN](#)[PP PUT](#)[W Wikipedia](#)[Q ODP](#)[i Jobs](#)[About](#) | [More demos](#) | [Download](#) | [Carrot2 @ sf.net](#) | [Carrot Search](#)[Search](#)[Hide options](#)Download Cluster with **All results (400)**

- [Climate Change \(66\)](#)
- [Greenhouse Effects \(46\)](#)
- [Increase in the Average Temperature \(35\)](#)
- [Global Warming News \(33\)](#)
- [Cause Global \(37\)](#)
- [Beings \(24\)](#)
- [Information on Global Warming \(23\)](#)
- [Global Warming Debate \(22\)](#)
- [Earth's Atmosphere \(30\)](#)
- [New Scientists \(30\)](#)
- [Time Global Warming \(24\)](#)
- [World \(21\)](#)
- [Heat is Online \(22\)](#)
- [Doing \(16\)](#)
- [Scientific \(18\)](#)
- [Answers to Frequently Asked Questions \(14\)](#)

1 [Climate Change | U.S. EPA](#)

... the issue of climate change and global warming in a way that is accessible and ... An archive of the Global Warming Site is available. ...

<http://www.epa.gov/climatechange/index.html>

2 [Global Warming](#)

Find answers about global warming, climate change, and the world's weather. From the National Oceanic and Atmospheric Administration.

<http://www.ncdc.noaa.gov/oa/climate/globalwarming.html>

3 [Global warming - Wikipedia, the free encyclopedia](#)

Read about the global warming and climate crisis debate, with information about warming's causes and its effects on life on Earth. Wikipedia's user-written overview also discusses climate models, statistics, and predictions.

http://en.wikipedia.org/wiki/Global_warming

4 [Global Warming International Center - Home](#)

Disseminates information on global warming science and policy, serving governmental and non-governmental organizations, and industries in more than 120 countries.

<http://www.globalwarming.net/>

Query: global warming -- Input: Yahoo! (400 results) -- Clusterer: Lingo

<http://www.carrot2.org>

© 2002-2000 Alan H. Wright, Jr. All rights reserved.

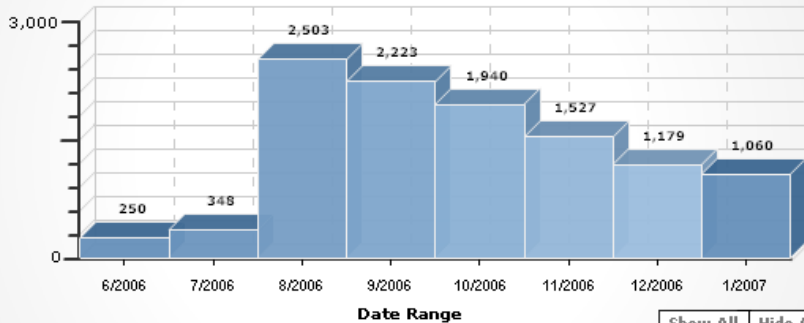
The on-line demo is a playground for users, but also a demonstration of the technology really used by quite a number of people.

Visitors Graph



Totals: 11,030 Average: 1,378.75

Visitors



Show All

Hide All



Carrot² has a number of **real-world applications**



Commercial spin-off: Carrot Search s.c.



CARROT
SEARCH

- A different, improved clustering algorithm – **Lingo3G**
- Consulting and support for the open source project
- Text-mining consultancy

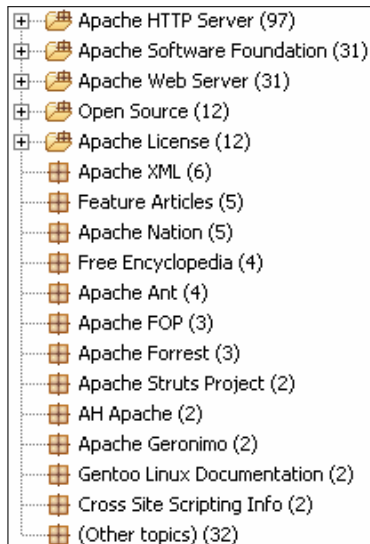
Lingo3G introduces many improvements

- hierarchical results,
- a number of customization options,
- much faster and robust,
- better cluster labels.



- 1 Introduction to Search Results Clustering
- 2 Carrot² Framework
- 3 Lingo clustering algorithm
- 4 Summary

Lingo is designed specifically for search results clustering



- **Semantic** clusters
- **Meaningful** cluster labels
- **Small** input

Clustering Search Results...

└ Lingo clustering algorithm

└ **Lingo** is designed specifically for search results clustering

Lingo is designed specifically for search results clustering

```

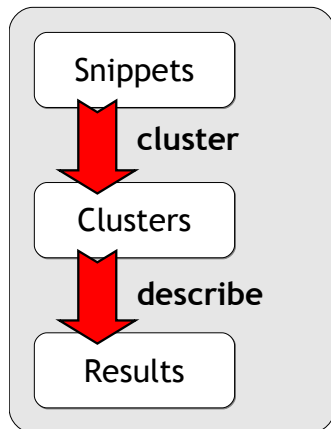
11  Apache HTTP Server (34)
12  Apache Software Foundation (31)
13  Apache Web Server (31)
14  Open Source (12)
15  Apache License (12)
16  Apache 1.0 (6)
17  Feature Articles (5)
18  Apache Nation (5)
19  Free Encyclopedia (4)
20  Apache Ant (4)
21  Apache POP (3)
22  Apache Project (3)
23  All Apache (2)
24  Apache Geomarine (2)
25  Gentoo Linux Documentation (2)
26  Cross Site Scripting (XSS) (2)
27  (Other topics) (2)
  
```

- Semantic clusters
- Meaningful cluster labels
- Small input

The primary assumption we made when working Lingo was that it should be an algorithm specifically designed to handle search results clustering. Therefore our main focus was the quality of cluster label. We also were aware that, due to the small size of input, we could afford more complex processing.

Cluster description has a priority

Classic clustering

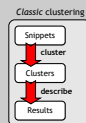


Clustering Search Results. . .

└ Lingo clustering algorithm

└ **Cluster description** has a priority

Cluster description has a priority

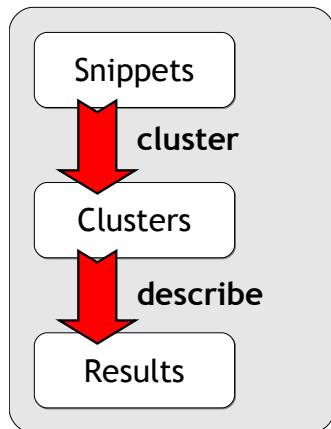


Having in mind the requirement for high quality of cluster labels, we experimented with **reversing the normal clustering order** and giving the cluster description a priority.

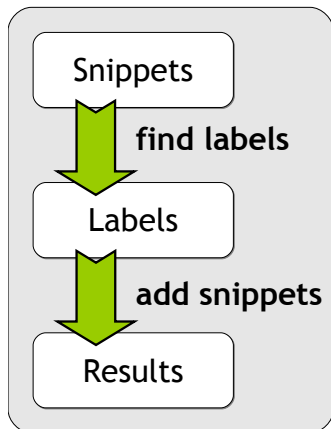
In the classic clustering scheme, in which the algorithm starts with finding document groups and then tries to label these groups, we can have situations where the algorithm knows that certain documents should be clustered together, but at the same time the algorithm is unable to explain to the user what these documents have in common.

Cluster description has a priority

Classic clustering



Description comes first clustering

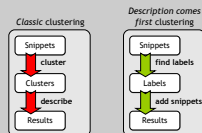


Clustering Search Results...

└ Lingo clustering algorithm









└ **Cluster description** has a priority

Cluster description has a priority



We can try to avoid these problems by starting with finding a set of meaningful and diverse cluster labels and then assigning documents to these labels to form proper clusters. This kind of general clustering procedure we called “**description comes first clustering**” and implemented in a search results clustering algorithm called LINGO.

Phrases are good label candidates

1. [Apache Software Foundation](#) 
Membership-based, not-for-profit corporation that exists to provide organizational, legal, and financial support for the **Apache** open-source software projects.
Category: [Unix Servers > Apache](#)
[www.apache.org](#) - [More from this site](#) - [Save](#)
2. [Apache HTTP Server Project](#) 
Effort to develop and maintain an open-source HTTP server for modern operating systems including UNIX and Windows NT.
[httpd.apache.org](#) - 32k - [Cached](#) - [More from this site](#) - [Save](#)
3. [Download - The Apache HTTP Server Project](#) 
Essentials. Download! Get Involved. Subprojects. Use the links below to download the **Apache** HTTP Server from one of our mirrors. You must verify the integrity of the downloaded files using signatures downloaded from our main distribution directory. ...
[comhttp://apache.mirrors.versehost.comhttp://apache.forbigweb.comhttp://apache.mirrors.hoobly.com](#) ...
[httpd.apache.org/download.cgi](#) - 17k - [Cached](#) - [More from this site](#) - [Save](#)
4. [Apache.com - Providing Web Server and Network Security Resources](#) 
Web Hosting Control Panel. Control panels provide a convenient interface with which to create new users, assign admin privileges, bind DNS entries, control servers, sites, and more. ... One ADC, Zeus Extensible Traffic Manager (ZTM) can accelerate **Apache** by as much as 100 times ... The **Apache** HTTP Server Project develops and maintains an open-source HTTP server for ...
[www.apache.com](#) - 15k - [Cached](#) - [More from this site](#) - [Save](#)
5. [The Jakarta Site - The Jakarta Project -- Java Related Products](#) 
... solutions and is a part of The **Apache** Software Foundation (ASF) which encourages a collaborative, consensus ... Copyright © 1999-2005, The **Apache** Software Foundation. Legal information ...
Category: [Unix Servers > Apache](#)
RSS: [View as XML](#) - [Add to My Yahoo!](#)
[jakarta.apache.org](#) - [More from this site](#) - [Save](#)
6. [Electronics Manufacturer Irvine Analog Broadband Modems Mp3 Wireless Communications California](#) 
... **Apache** Micro located in Irvine California is one of the world's leading analog modem and broadband ... manufacturing broadband and analog modems, **Apache**-Micro also produces a wide range ...
[www.apache-micro.com](#) - 11k - [Cached](#) - [More from this site](#) - [Save](#)
7. [Apache XML Project](#) 
To provide commercial-quality standards-based XML developed in an open and cooperative fashion and to provide feedback to standards bodies.
Category: [Unix Servers > Apache](#)
[xml.apache.org](#) - 31k - [Cached](#) - [More from this site](#) - [Save](#)
8. [Apache HTTP Server Version 2.0 Documentation - Apache HTTP Server](#) 
Apache HTTP Server Version 2.0. **Apache** HTTP Server Version 2.0 Documentation. Copyright 1995-2006 The **Apache** Software Foundation or its licensors, as applicable. Licensed under the **Apache** License, Version 2.0.
[httpd.apache.org/docs/2.0](#) - 7k - [Cached](#) - [More from this site](#) - [Save](#)

Apache Cocoon
 Apache Ant
 Apache HTTP Server
 XML
 Apache HTTP
 Apache Server
 Server HTTP
 Web Server
 Apache Tomcat
 Apache Web Server
 Apache Incubator
 Native Americans
 Apache Software Foundation
 Software Foundation
 Apache County
 Apache Geronimo
 Apache Indians
 Apache Junction

... and 300 more...

Clustering Search Results. . .

Phrases are good label candidates

└ Lingo clustering algorithm

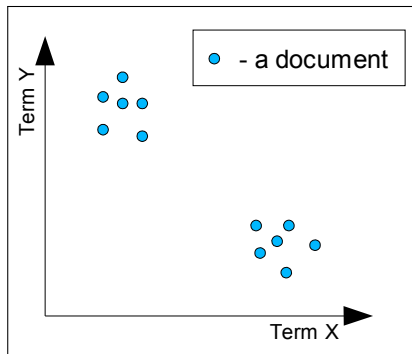
└ **Phrases** are good label candidates

So how do we go about finding good cluster labels? One of the first approaches to search results clustering called Suffix Tree Clustering would group documents according to the common phrase they shared.

Frequent phrases are very often collocations (such as Web Server or Apache County), which increases their descriptive power. But how do we select the best and most diverse set of cluster labels? We've got quite a lot of label candidates. . .

Approximate matrix factorizations can find labels

$$\begin{array}{c}
 \text{term 1} \\
 \text{term 2} \\
 \text{term 3} \\
 \text{term 4} \\
 \text{term 5} \\
 \text{term 6}
 \end{array}
 \begin{array}{c}
 \text{doc 1} \\
 \text{doc 2} \\
 \text{doc 3} \\
 \text{doc 4}
 \end{array}
 \begin{bmatrix}
 * & * & * & * \\
 * & * & * & * \\
 * & * & * & * \\
 * & * & * & * \\
 * & * & * & * \\
 * & * & * & *
 \end{bmatrix}
 = A$$



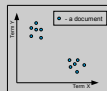
Clustering Search Results...

└ Lingo clustering algorithm

└ Approximate matrix factorizations can **find labels**

Approximate matrix factorizations can find labels

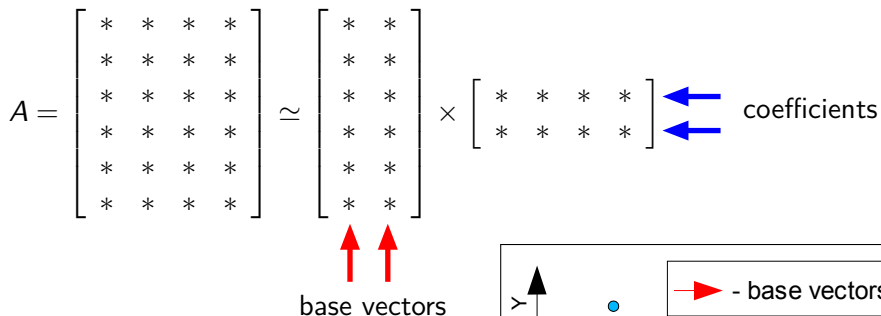
$$\begin{array}{c}
 \text{term 1} \\
 \text{term 2} \\
 \text{term 3} \\
 \text{term 4} \\
 \text{term 5} \\
 \text{term 6}
 \end{array}
 \begin{bmatrix}
 \text{doc 1} & + & + & + & + & + \\
 \text{doc 2} & + & + & + & + & + \\
 \text{doc 3} & + & + & + & + & + \\
 \text{doc 4} & + & + & + & + & +
 \end{bmatrix}
 = A$$



We can do that using **Vector Space Model** and matrix factorizations. To build the Vector Space Model we need to create a so called term-document matrix: a matrix containing frequencies of all terms across all input documents. If we had just two terms – term X and Y – we could visualise the Vector Space Model as a plane with two axes corresponding to the terms and points on that plane corresponding to the actual documents.

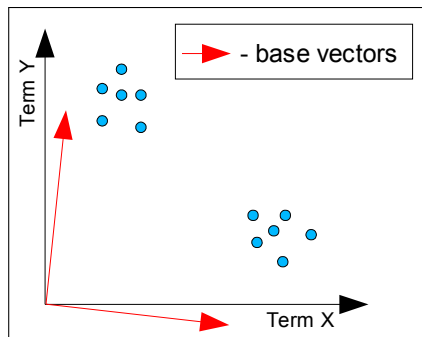
Approximate matrix factorizations can find labels

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



base vectors

coefficients

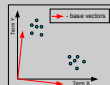


Clustering Search Results...

└ Lingo clustering algorithm

└ Approximate matrix factorizations can **find labels**

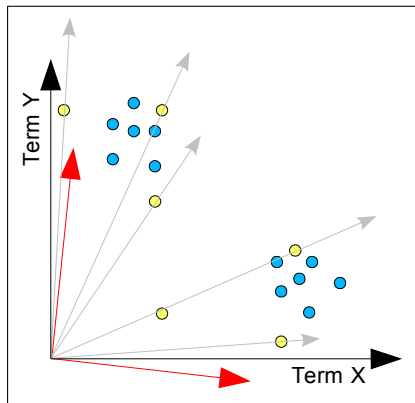
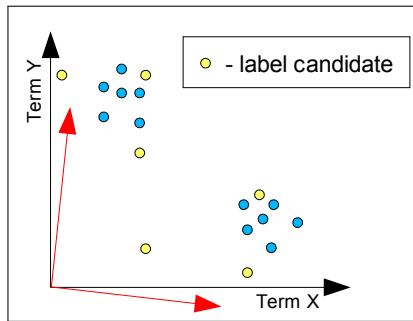
$$A = \begin{bmatrix} + & + & + & + & + \\ + & + & + & + & + \\ + & + & + & + & + \\ + & + & + & + & + \\ + & + & + & + & + \end{bmatrix} \approx \begin{bmatrix} + & + \\ + & + \\ + & + \\ + & + \\ + & + \end{bmatrix} \times \begin{bmatrix} + & + & + & + \\ + & + & + & + \end{bmatrix}$$



The task of an **approximate matrix factorization** is to break a matrix into a product of usually two matrices in such a way that the product is as close to the original matrix as possible and has much lower rank. The left-hand matrix of the product can be thought of as a set of base vectors of the new low-dimensional space, while the other matrix contains the corresponding coefficients that enable us to reconstruct the original matrix.

In the context of our simplified graphical example, base vectors show the general directions or trends in the input collection.

Approximate matrix factorizations can find labels

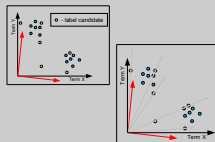


Clustering Search Results...

└ Lingo clustering algorithm

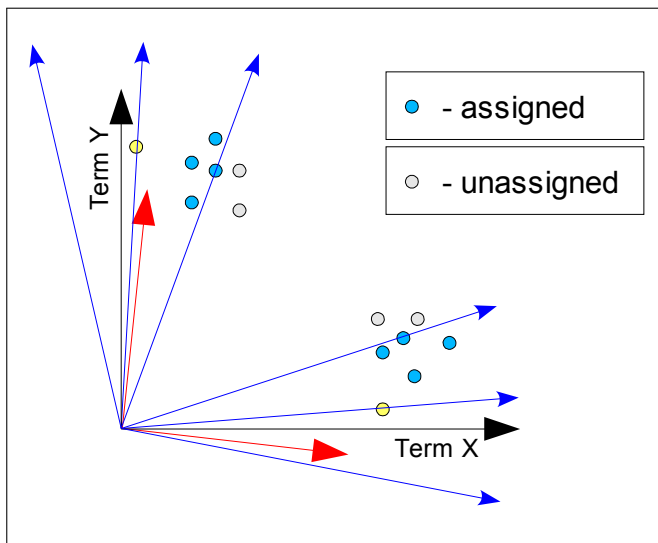
└ Approximate matrix factorizations can **find labels**

Approximate matrix factorizations can find labels



Please notice that both frequent phrases and base vectors are expressed in the same space as the input documents (think of the phrases as tiny documents). With this assumption we can use e.g. cosine distance to find the best matching phrase for each base vector. In this way, each base vector will lead to selecting one cluster label.

Cosine distance can **find** documents

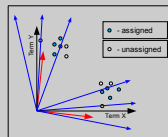


Clustering Search Results...

└ Lingo clustering algorithm

└ Cosine distance can **find documents**

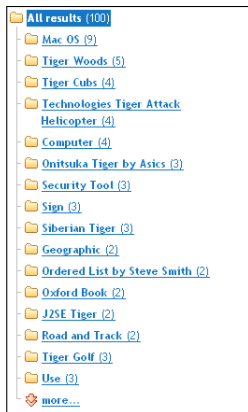
Cosine distance can find documents



To form proper clusters, we can again use cosine similarity and assign to each label those documents whose similarity to that label is larger than some threshold.

Giving priority to labels **pays off**

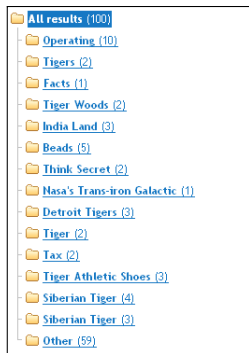
Lingo



STC



Rough K-Means



- 1 Introduction to Search Results Clustering
- 2 Carrot² Framework
- 3 Lingo clustering algorithm
- 4 Summary**

Summary

- Exploit the potential of existing ontologies?
- Investigate support for more languages.
- Investigate more data sources.

References

- Osinski, S. and Weiss, D. (2005). A Concept-Driven Algorithm for Clustering Search Results. *IEEE Intelligent Systems*, 20(3):48–54
- Osiński, S., Stefanowski, J., and Weiss, D. (2004). Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition. In *Proceedings of the International Intelligent Information Processing and Web Mining Conference, Zakopane, Poland*, Advances in Soft Computing, pages 359–368. Springer
- Weiss, D. (2006). *Descriptive Clustering as a Method for Exploring Text Collections*. PhD thesis, Poznan University of Technology, Poznań, Poland

Carrot² links

- On-line demo:
<http://www.carrot2.org>
- Open source project:
<http://project.carrot2.org>
- SourceForge (repository etc.):
<http://sourceforge.net/projects/carrot2>
- Carrot Search:
<http://www.carrot-search.com>

