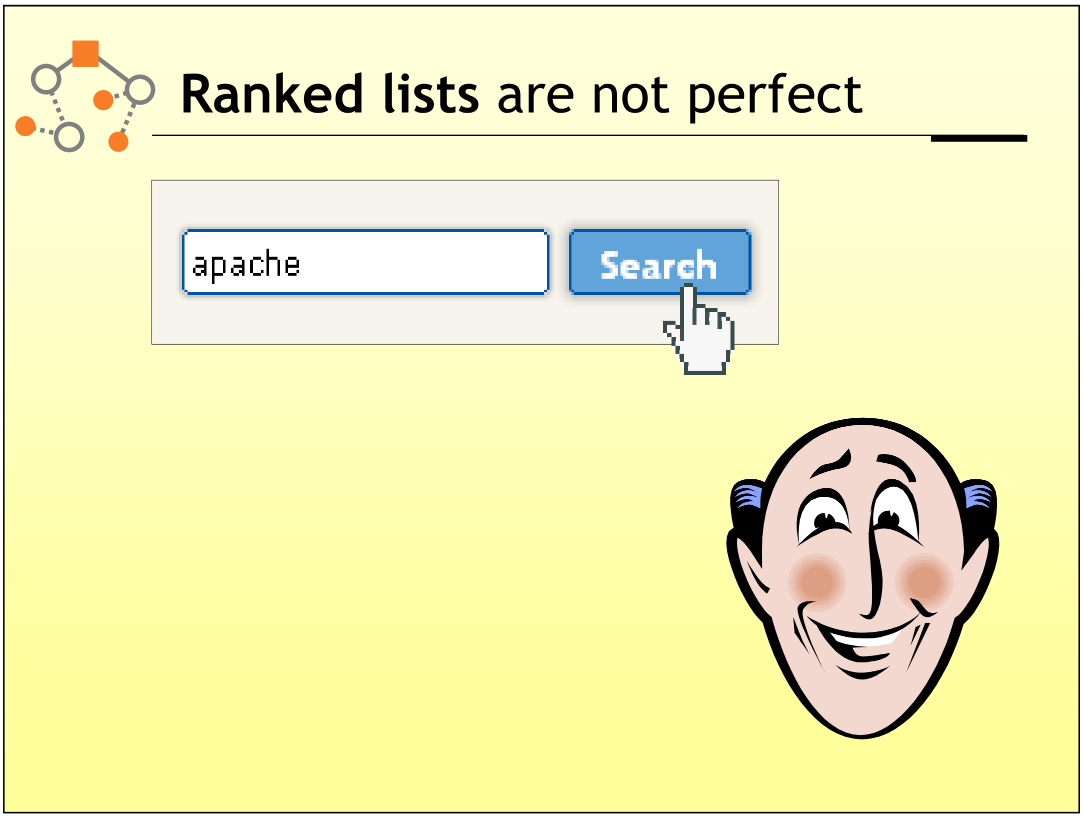
Improving Quality of **Search Results Clustering** with Approximate Matrix Factorisations

Stanisław Osiński

Poznan Supercomputing and Networking Center



Ranked lists are not perfect

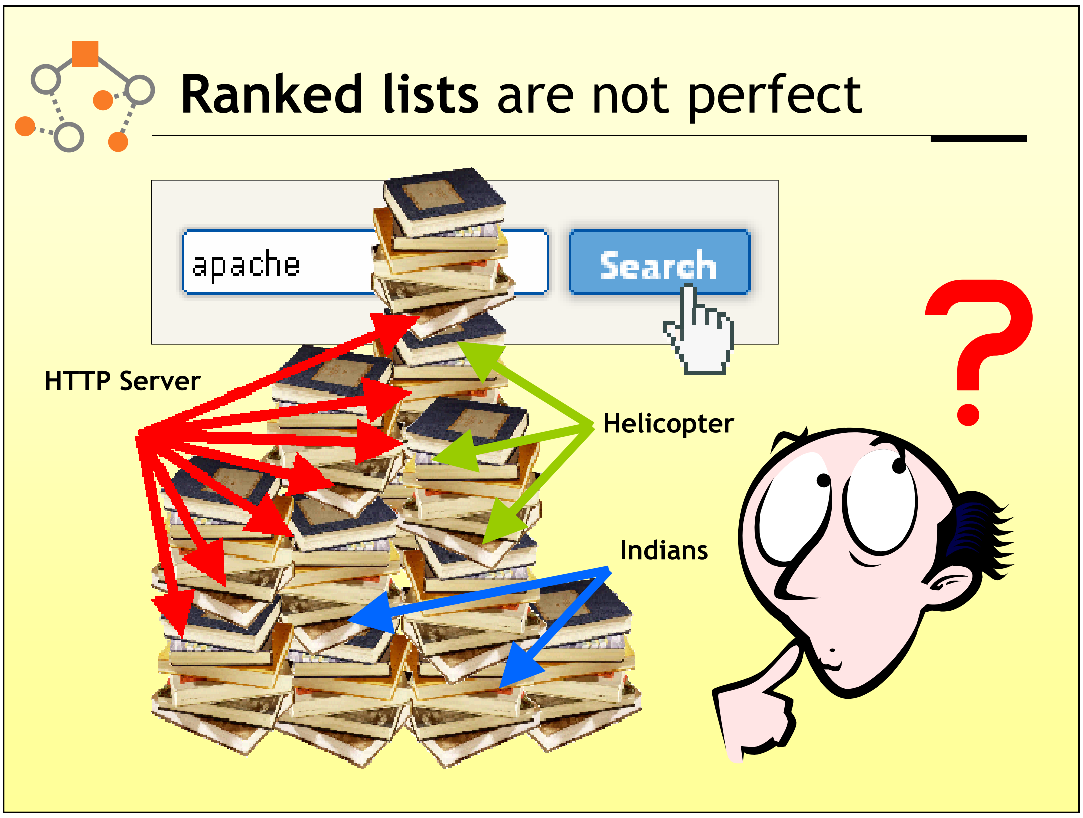
Search results clustering is one of many methods that can be used to **improve user experience while searching collections of text documents**, web pages for example.

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

apache
Search
HTTP Server
?

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



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.

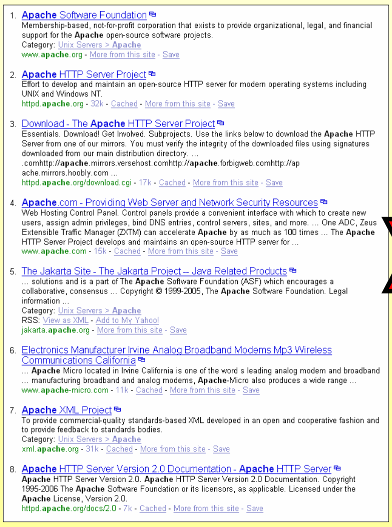
Search Results Clustering can help

apache
Search
!
Helicopter
Indians
HTTP Server
(other topics)

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 shuold 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.

SRC is based on **document snippets**



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

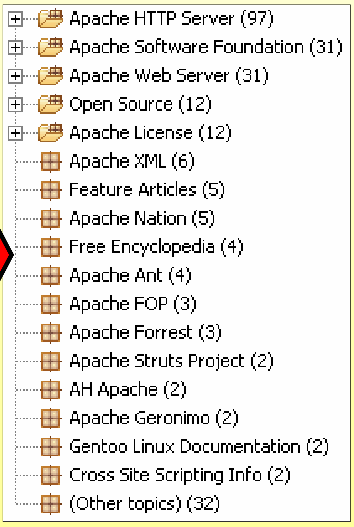
4. **Apache**.com - Providing Web Server and Network Security Resources
Web Hosting Control Panel. Control panels provide a convenient interface with which to create new users, assign admin privileges, bind DNS entries, control servers, sites, and more. ... One ADC, Zeus Extensible Traffic Manager (ZXTM) 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 word 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

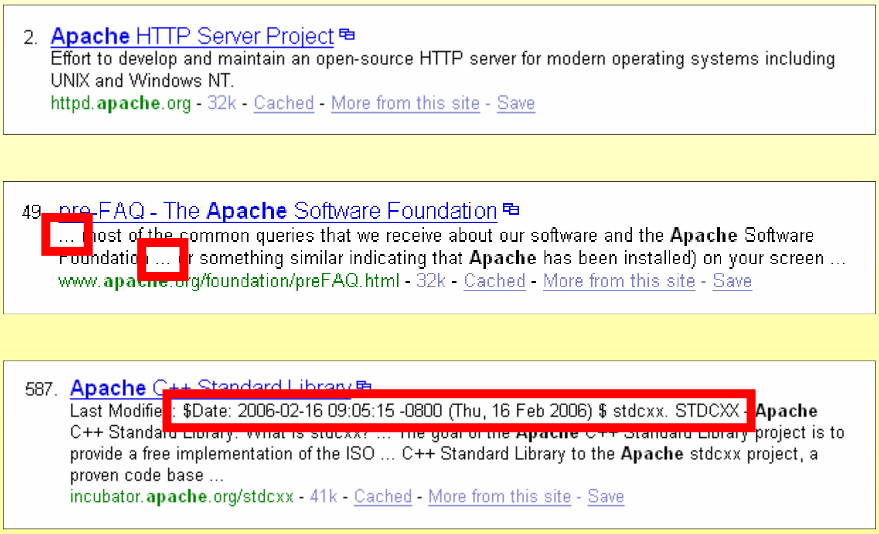
8. **Apache** HTTP Server Version 2.0 Documentation - **Apache** HTTP Server
Apache HTTP Server Version 2.0. **Apache** HTTP Server Version 2.0 Documentation. Copyright 1995-2006 The **Apache** Software Foundation or its licensors, as applicable. Licensed under the **Apache** License, Version 2.0.
httpd.**apache**.org/docs/2.0 - 7k - Cached - More from this site - Save



- Apache 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)

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 we have, we don't have full documents.

SRC is an **interesting problem**



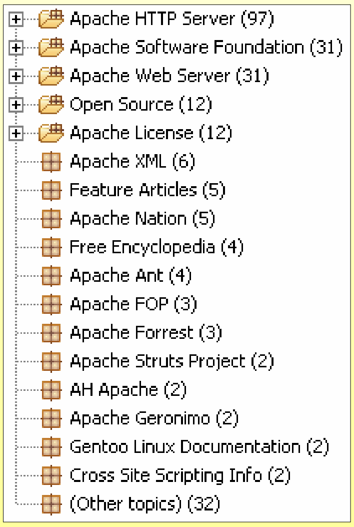
2. **Apache** HTTP Server Project
Effort to develop and maintain an open-source HTTP server for modern operating systems including UNIX and Windows NT.
httpd.**apache**.org - 32k - Cached - More from this site - Save

49. pre-FAQ - The **Apache** Software Foundation
... most of the common queries that we receive about our software and the **Apache** Software Foundation ... 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

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

SRC is an **interesting problem**



- 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)

Semantic clusters

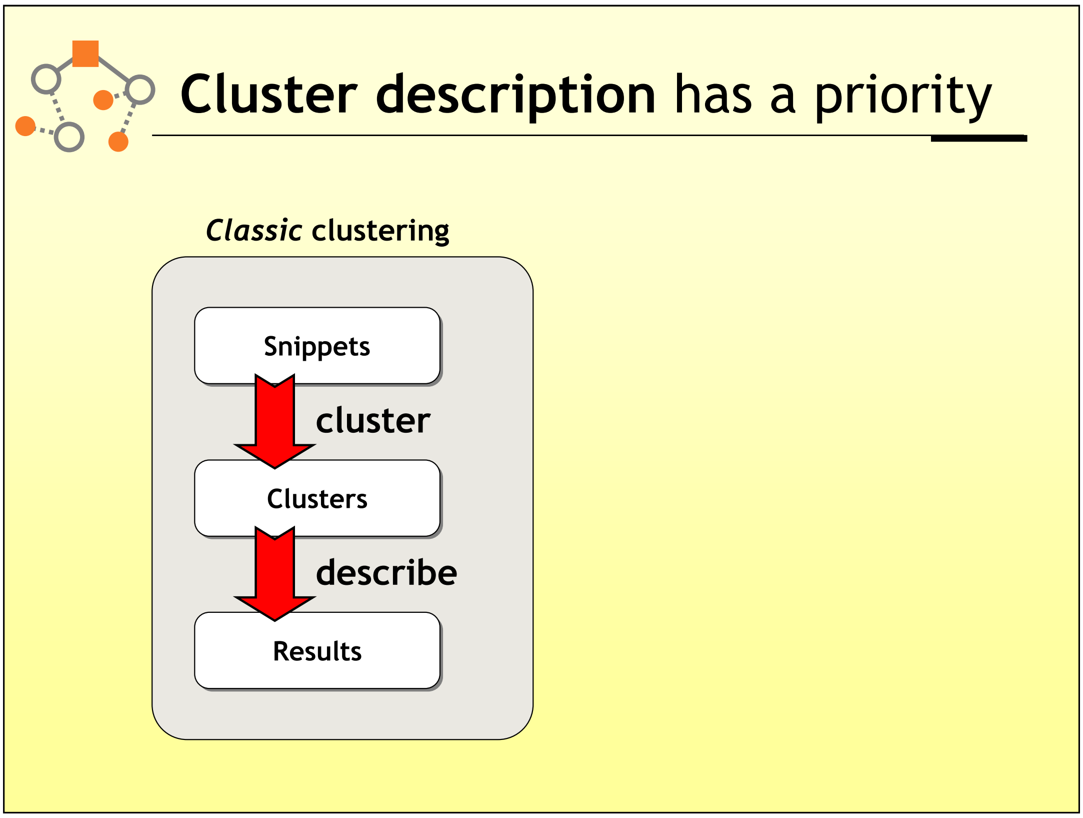
Meaningful cluster labels

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.



Cluster description has a priority

Classic clustering

Snippets

cluster

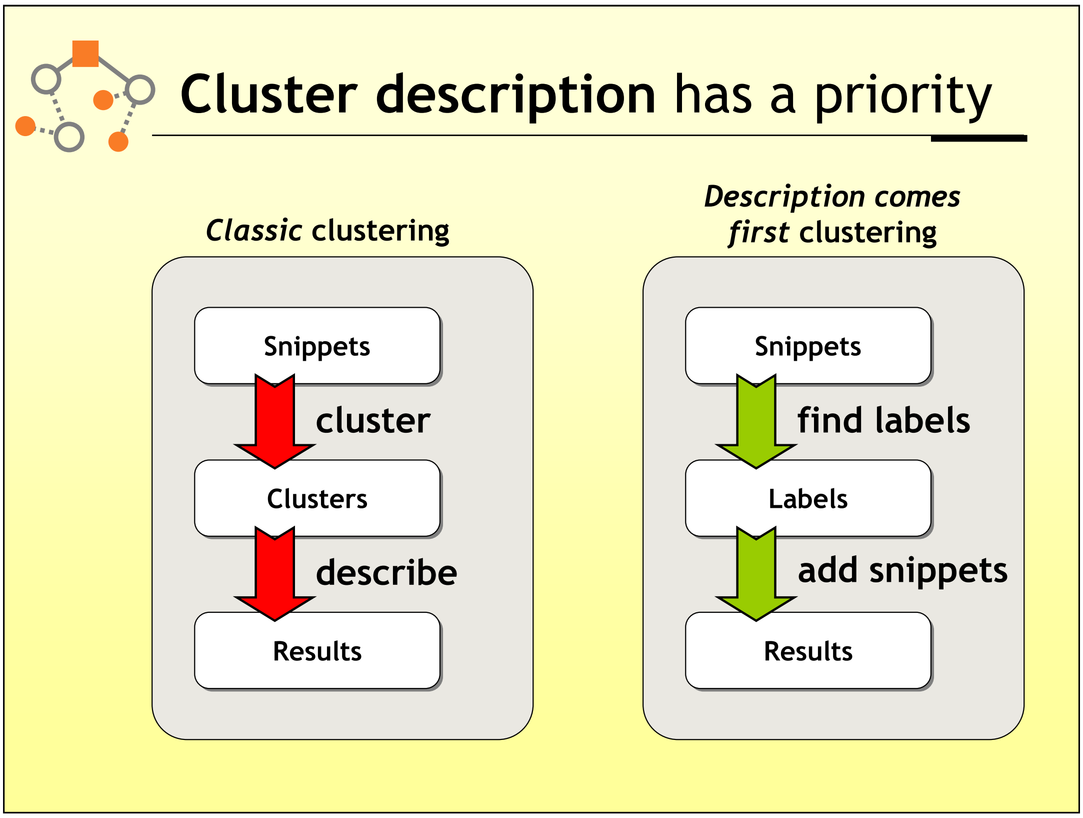
Clusters

describe

Results

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

Snippets

cluster

Clusters

describe

Results

Description comes first clustering

Snippets

find labels

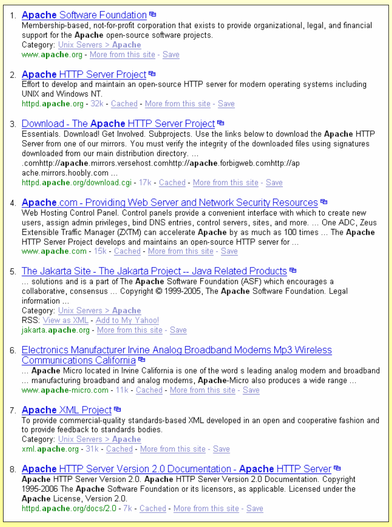
Labels

add snippets

Results

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 resuts 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://ap
   ache.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 (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 word 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

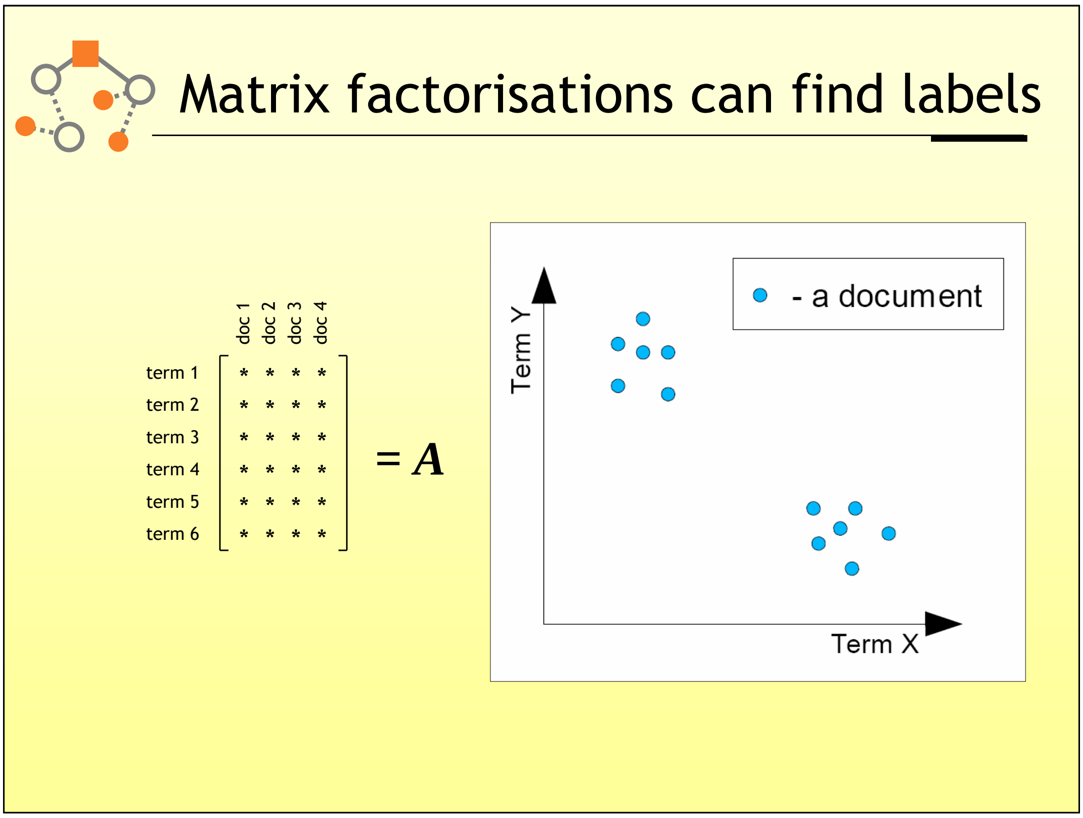
7. **Apache** XML Project
   To provide commercial-quality standards-based XML developed in an open and cooperative fashion and to provide feedback to standards bodies.
   Category: Unix Servers > **Apache**
   xml.**apache**.org - 31k - Cached - More from this site - Save

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

... and 300 more...

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...



Matrix factorisations can find labels

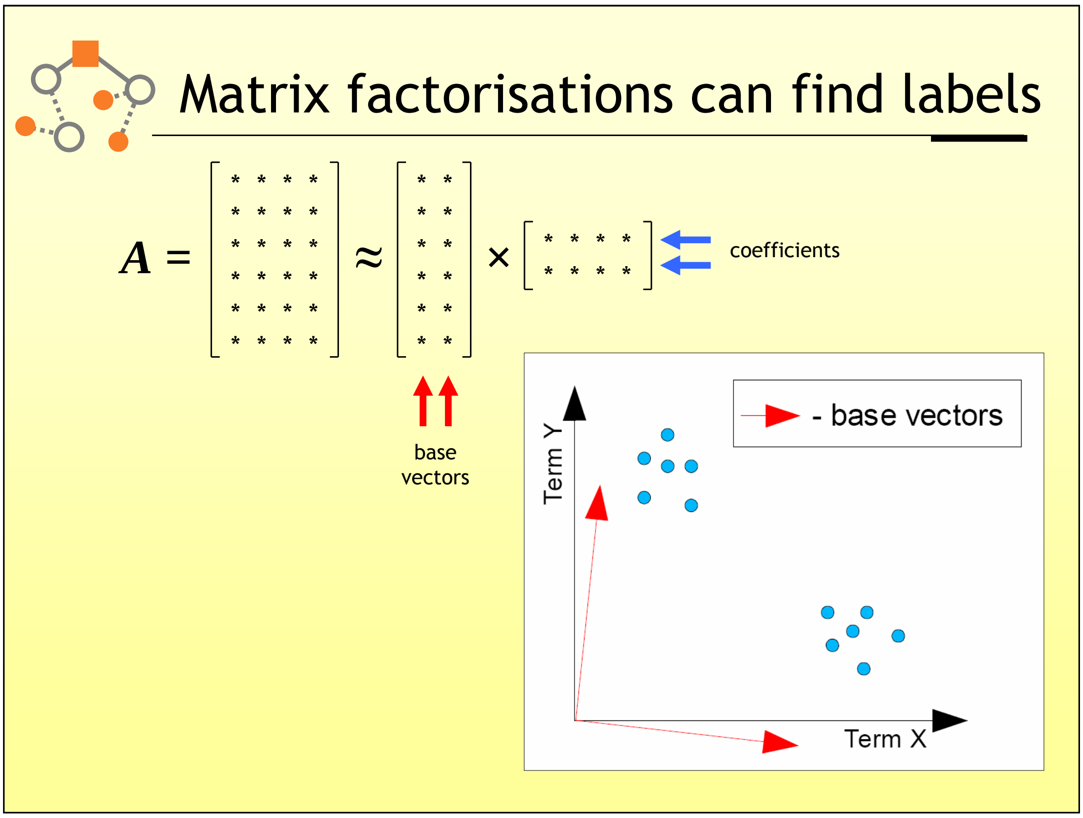
$$\begin{array}{c|cccc|} & \text{doc 1} & \text{doc 2} & \text{doc 3} & \text{doc 4} \\ \text{term 1} & * & * & * & * \\ \text{term 2} & * & * & * & * \\ \text{term 3} & * & * & * & * \\ \text{term 4} & * & * & * & * \\ \text{term 5} & * & * & * & * \\ \text{term 6} & * & * & * & * \end{array} = \boldsymbol{A}$$

● - a document
Term Y
Term X

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.



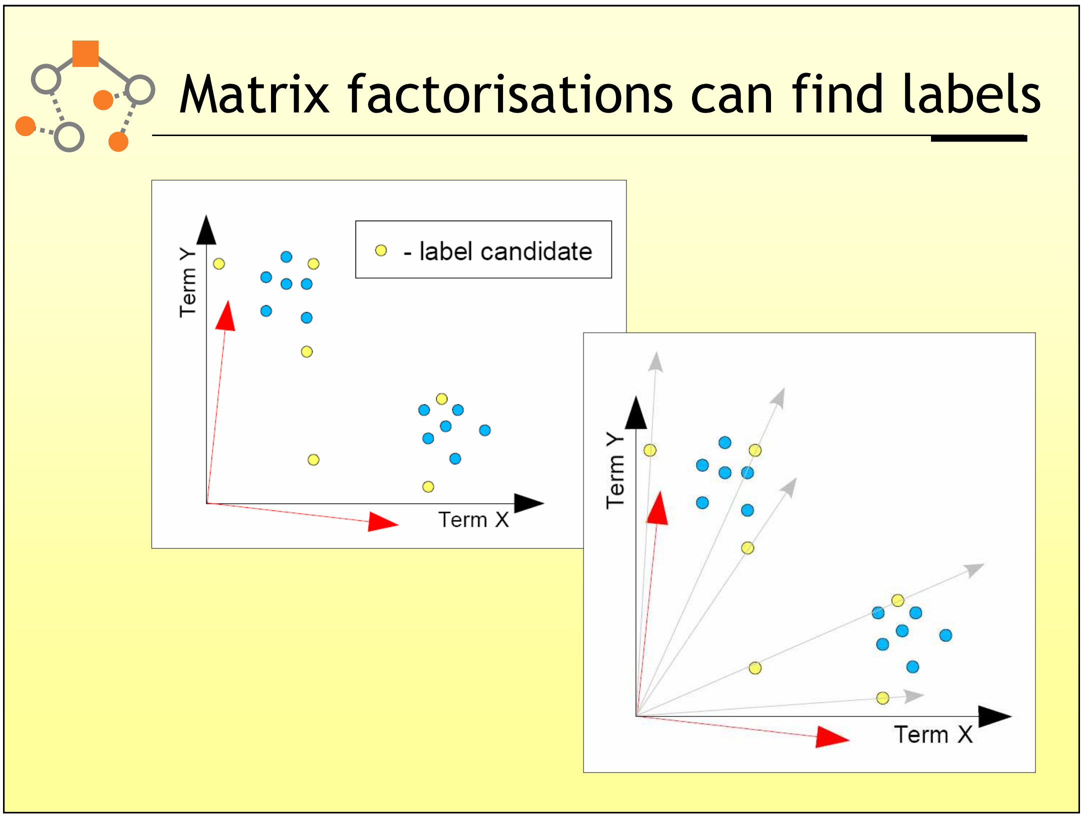
Matrix factorisations can find labels

$$A = \begin{bmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{bmatrix} \approx \underbrace{\begin{bmatrix} * & * \\ * & * \\ * & * \\ * & * \\ * & * \\ * & * \end{bmatrix}}_{\text{base vectors}} \times \begin{bmatrix} * & * & * & * \\ * & * & * & * \end{bmatrix} \leftleftarrows \text{coefficients}$$

The task of an **approximate matrix factorisation** 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.

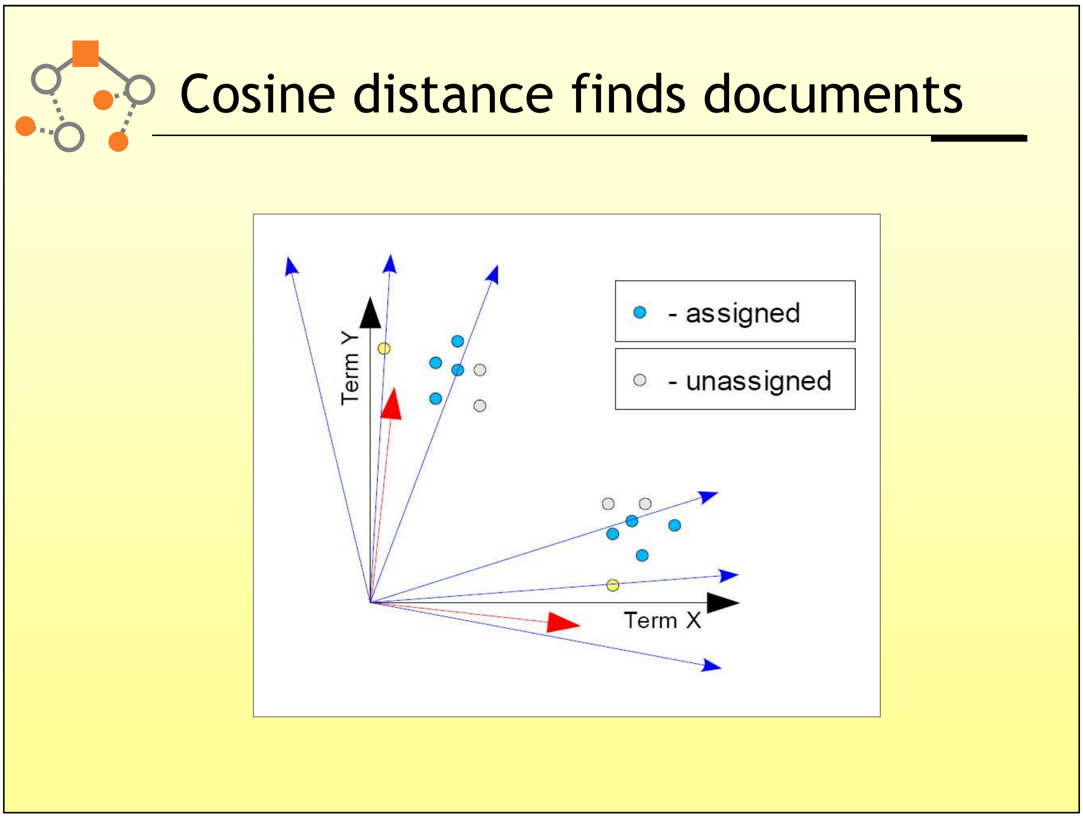
The left-hand matrix of the product can be tought of as a set of base vectors of the new low-dimensional space, while the other matrix contains the corresponding coefficients that enable us to reconstruct the original matrix.

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



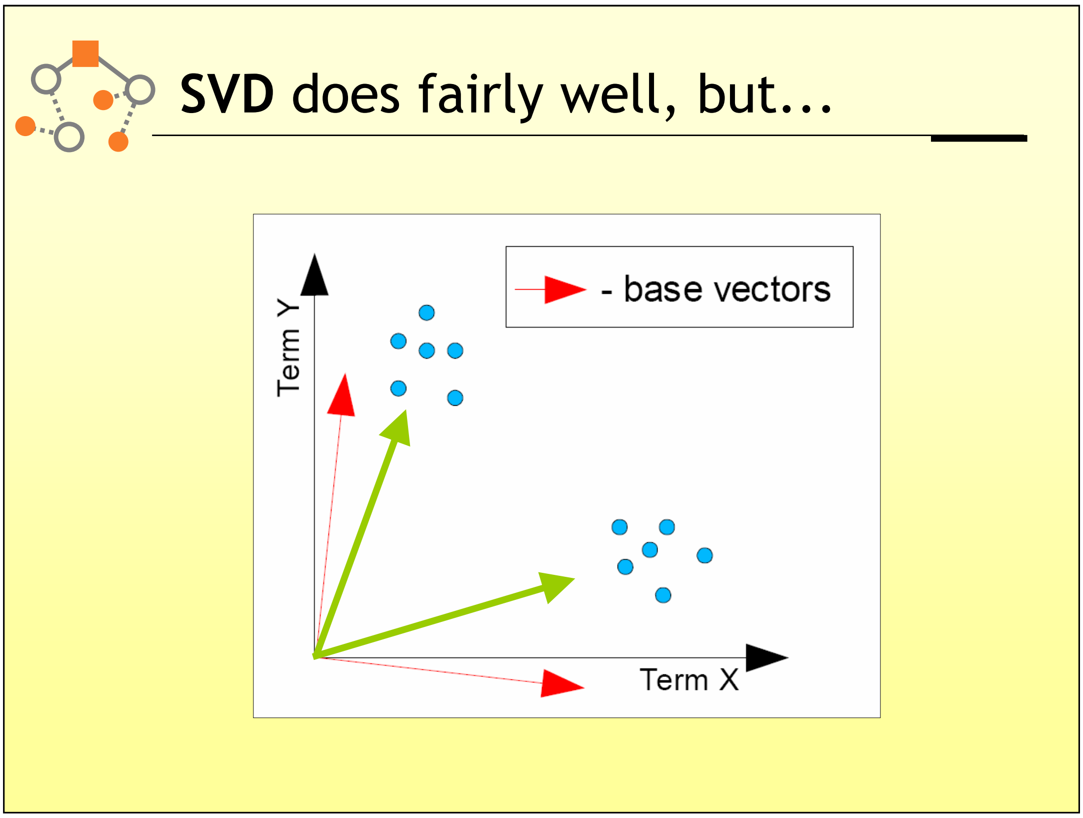
Matrix factorisations can find labels

Please notice that frequent phrases 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 finds documents

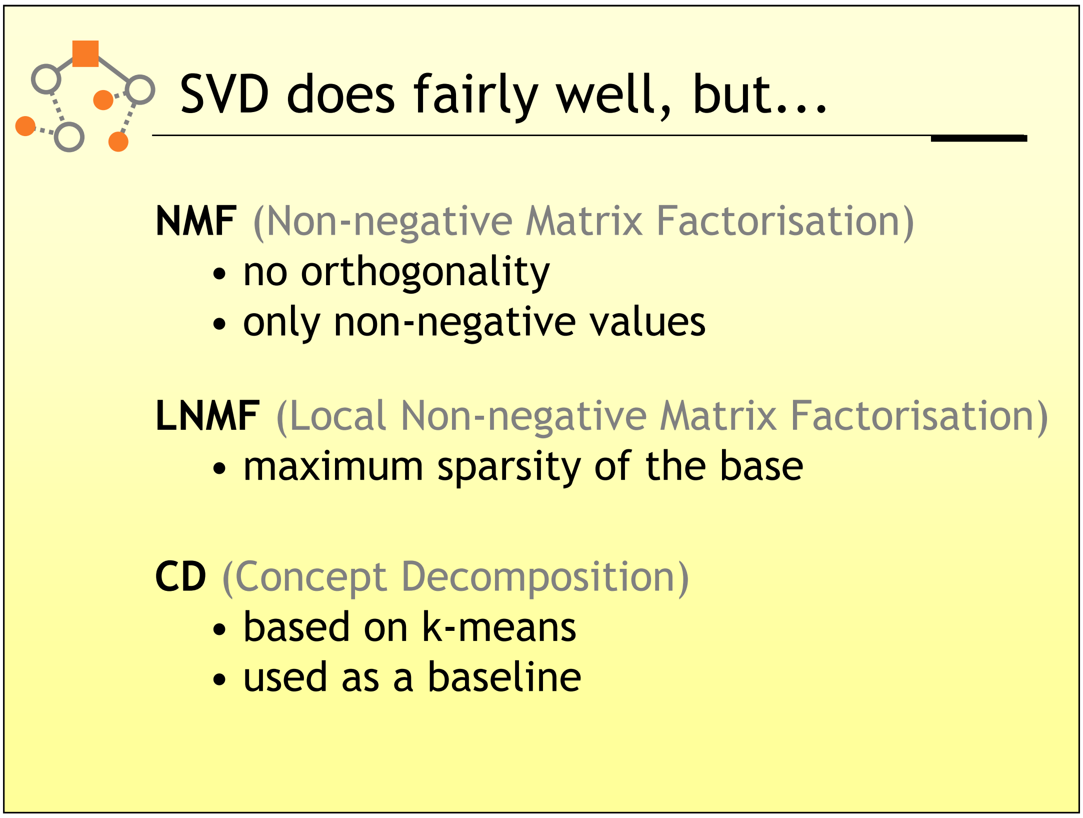
To form proper clusters, we can again use cosine similarity and assign to each label those documents whose similarity to that label is larger than some threshold.



SVD does fairly well, but...

In our initial experiments, we used SVD to obtain the base vectors.

And although it performed quite well, **the problem with SVD is that it generates strictly orthogonal bases**, which can lead to discovering not the best labels in some cases.



SVD does fairly well, but...

NMF (Non-negative Matrix Factorisation)

- no orthogonality
- only non-negative values

LNMF (Local Non-negative Matrix Factorisation)

- maximum sparsity of the base

CD (Concept Decomposition)

- based on k-means
- used as a baseline

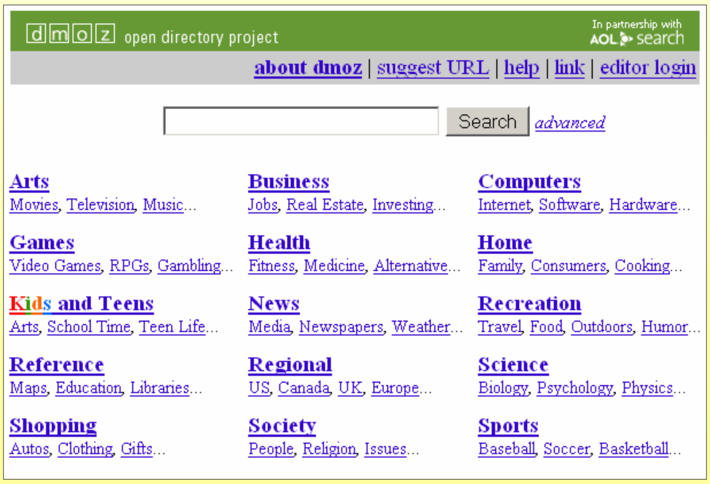
For this reason, **we tried how other matrix factorisations** would work as part of the description comes first clustering algorithm.

We tried **Nonnegative Matrix Factorisation**, which generates bases vectors with only positive values and doesn't impose orthogonality on the base vectors.

We also tried **Local Non-negative Matrix Factorisation**, which is similar to NMF, but also maximises the sparsity of base vectors.

Finally, as a baseline we used a matrix facorisation called **Concept Decomposition**, which is based on the k-means clustering algorithm (to be precise: centroid vectors obtained from k-means are taken as base vectors).

Open Directory good for evaluation



dmoz open directory project

In partnership with AOL search

about dmoz | suggest URL | help | link | editor login

Arts
Movies, Television, Music...

Business
Jobs, Real Estate, Investing...

Computers
Internet, Software, Hardware...

Games
Video Games, RPGs, Gambling...

Health
Fitness, Medicine, Alternative...

Home
Family, Consumers, Cooking...

Kids and Teens
Arts, School Time, Teen Life...

News
Media, Newspapers, Weather...

Recreation
Travel, Food, Outdoors, Humor...

Reference
Maps, Education, Libraries...

Regional
US, Canada, UK, Europe...

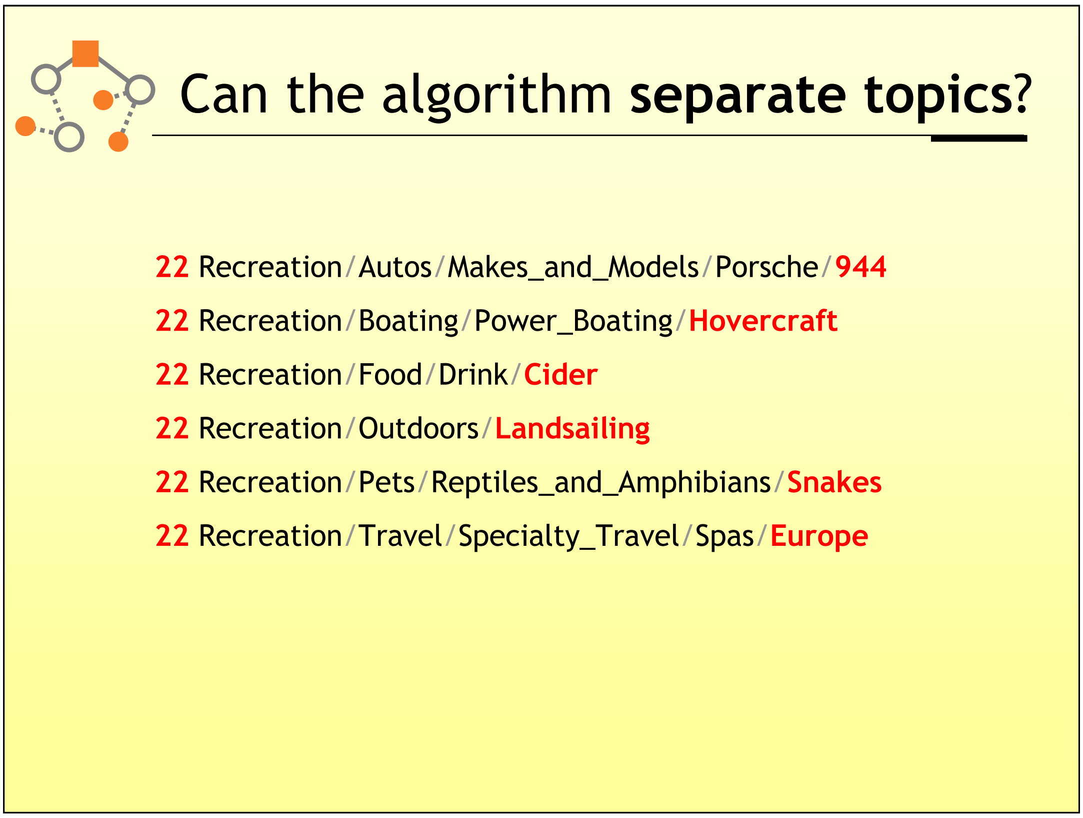
Science
Biology, Psychology, Physics...

Shopping
Autos, Clothing, Gifts...

Society
People, Religion, Issues...

Sports
Baseball, Soccer, Basketball...

Because there are no standard test collections for search results clustering, we decided to evaluate our approach using data collected by the **Open Directory Project**, which is a large, hierarchical, human-edited directory of the Web. Each branch in this directory corresponds to a distinct topic such as Games or Business, and each entry in the directory is accompanied by a short description, which to some extent emulates the document snippets returned by search engines.



Can the algorithm **separate topics?**

22 Recreation/Autos/Makes_and_Models/Porsche/**944**

22 Recreation/Boating/Power_Boating/**Hovercraft**

22 Recreation/Food/Drink/**Cider**

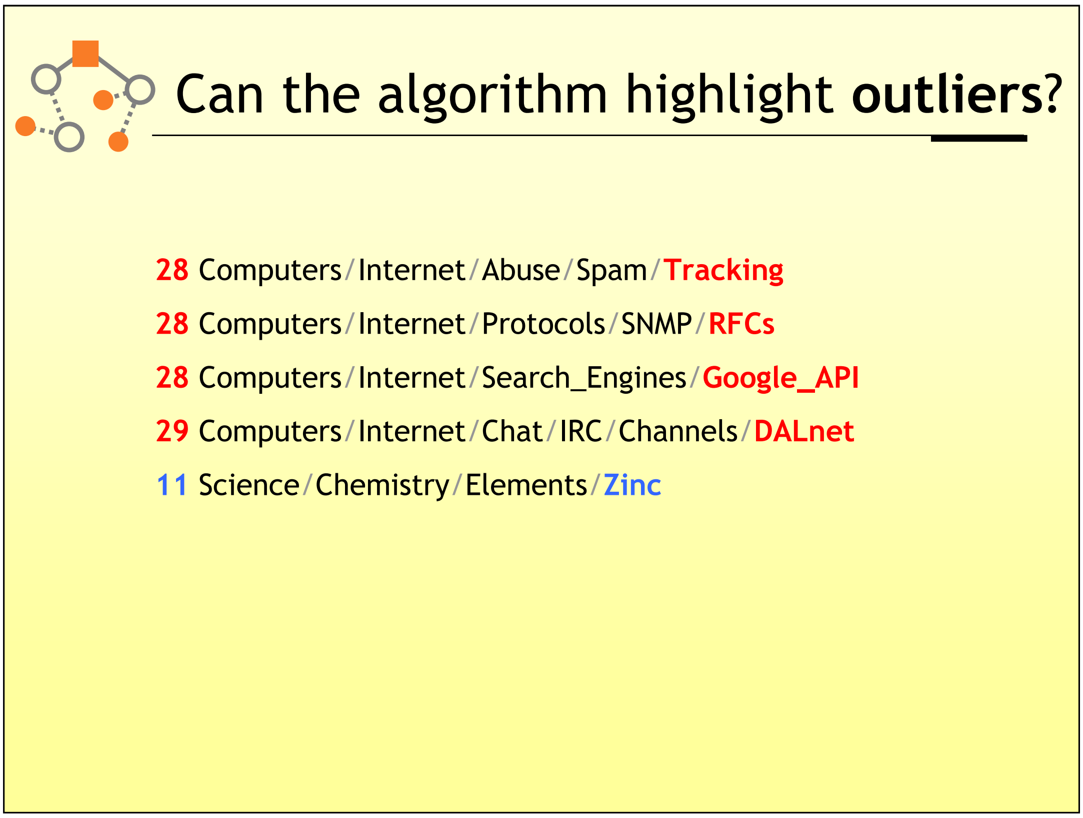
22 Recreation/Outdoors/**Landsailing**

22 Recreation/Pets/Reptiles_and_Amphibians/**Snakes**

22 Recreation/Travel/Specialty_Travel/Spas/**Europe**

During the evaluation we tried to answer two basic questions.

One question was whether the algorithm could effectively separate topics. To get this question answered, we prepared 63 data sets, each of which contained a mix of some manually selected Open Directory categories. One of these data sets you can see on the slide: we’ve got 22 snippets about Porshe, ... Obviously, we would expect the clustering algorithm to create clusters that somehow correspond to the original categories.



Can the algorithm highlight **outliers**?

28 Computers/Internet/Abuse/Spam/**Tracking**

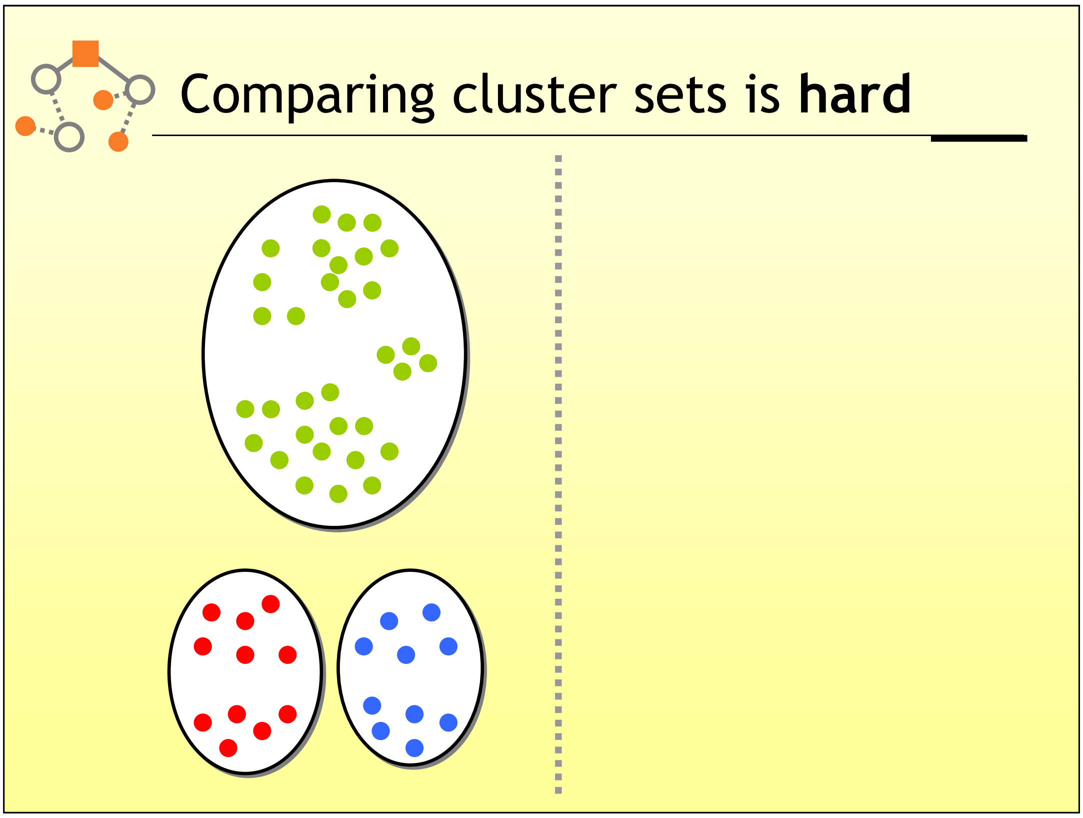
28 Computers/Internet/Protocols/SNMP/**RFCs**

28 Computers/Internet/Search_Engines/**Google_API**

29 Computers/Internet/Chat/IRC/Channels/**DALnet**

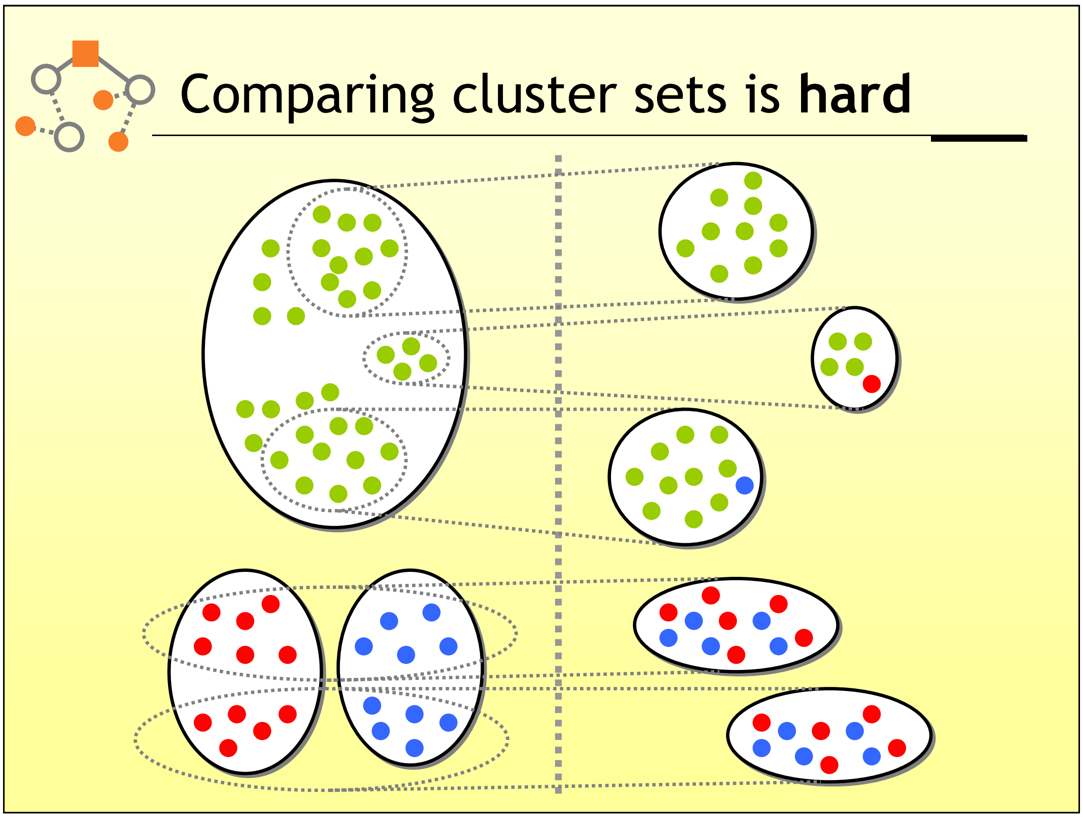
11 Science/Chemistry/Elements/**Zinc**

Another question was whether the clustering algorithm could highlight small outlier topics which are unrelated to the general theme of the data set. To answer this question we prepared 14 test sets similar to the one you can see on the slide: the majority of its snippets are related to the Internet, and there is just one, comparatively small category dealing with chemistry. We would expect an effective clustering algorithm to create at least one cluster corresponding to the outlier topic.



Comparing cluster sets is **hard**

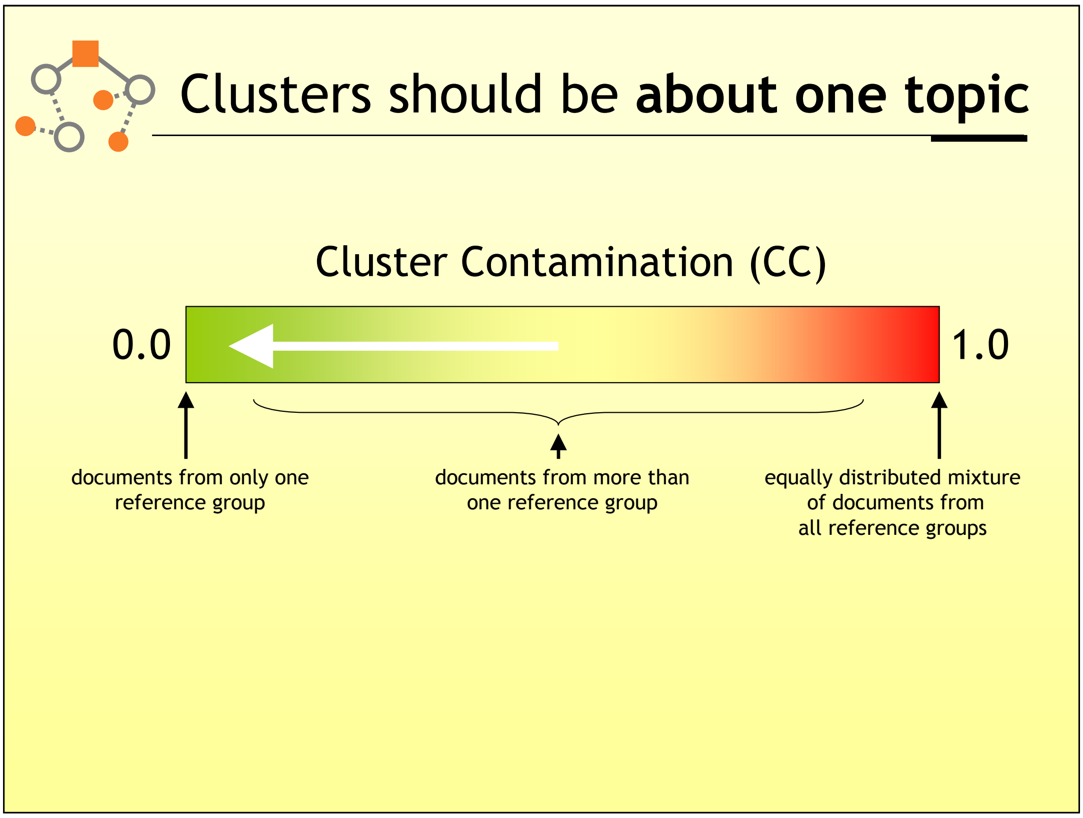
One problem with evaluation of clustering algorithms is that **there is no one definitely right answer here**, especially when the results are to be shown to a human user. Given an input data set consisting of three topics, the algorithm can come up with a different, but equally good clustering.



Comparing cluster sets is **hard**

For instance, the algorithm can decide to **split a large reference group into three smaller ones**. Or it could **cross-cut two reference groups**.

To avoid penalising the algorithm for making different, but equally justified choices, we decided to define three alternative cluster quality measures.



Clusters should be **about one topic**

Cluster Contamination (CC)

0.0

1.0

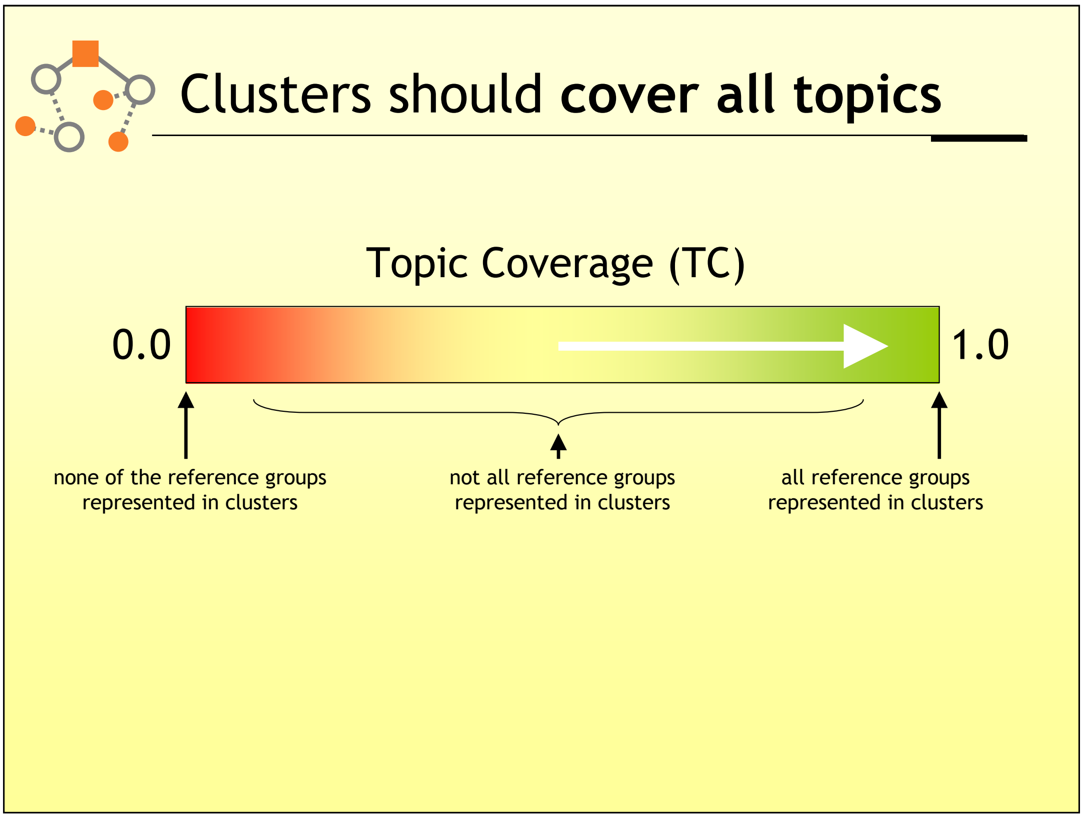
documents from only one reference group

documents from more than one reference group

equally distributed mixture of documents from all reference groups

The first measure says that a good cluster should be about one topic. To quantify that, we defined the **Cluster Contamination** measure, which is 0 for a cluster that contains documents from only one reference group. Cluster Contamination is greater than 0 for clusters containing documents from more than one reference group. Finally, in the worst case, for a cluster that contains an equally distributed mixture of documents from all reference groups, Cluster Contamination is 1.0.

Please notice that a cluster which is a subset of one reference group is not contaminated. Obviously, the closer we get to 0, the better.



Clusters should cover all topics

Topic Coverage (TC)

0.0

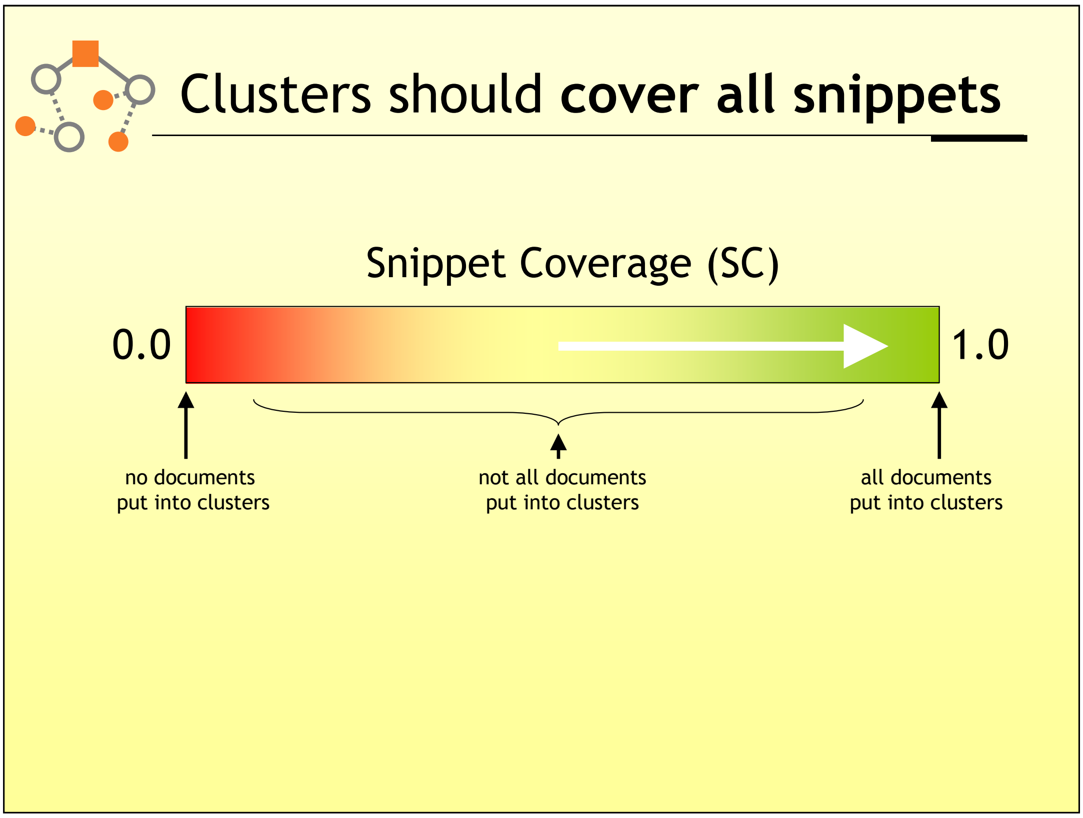
1.0

none of the reference groups represented in clusters

not all reference groups represented in clusters

all reference groups represented in clusters

The cluster set as a whole should cover all input topics. To quantify this aspect, we defined the **Topic Coverage** measure, which is 0 if none of the reference groups is represented in the clusters, greater than 0 if not all reference groups are represented. And finally, Topic Coverage is 1, when each reference group is represented by at least one cluster. The closer we get to 1 with this measure the better.



Clusters should **cover all snippets**

Snippet Coverage (SC)

0.0

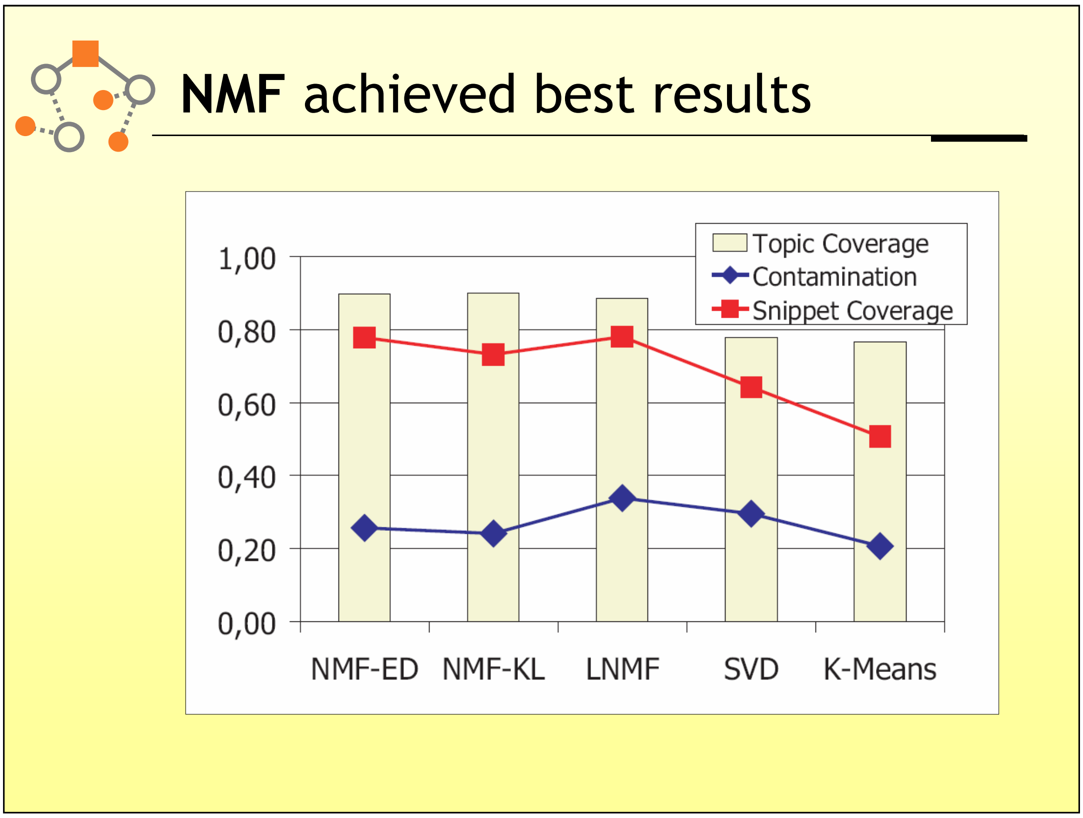
1.0

no documents put into clusters

not all documents put into clusters

all documents put into clusters

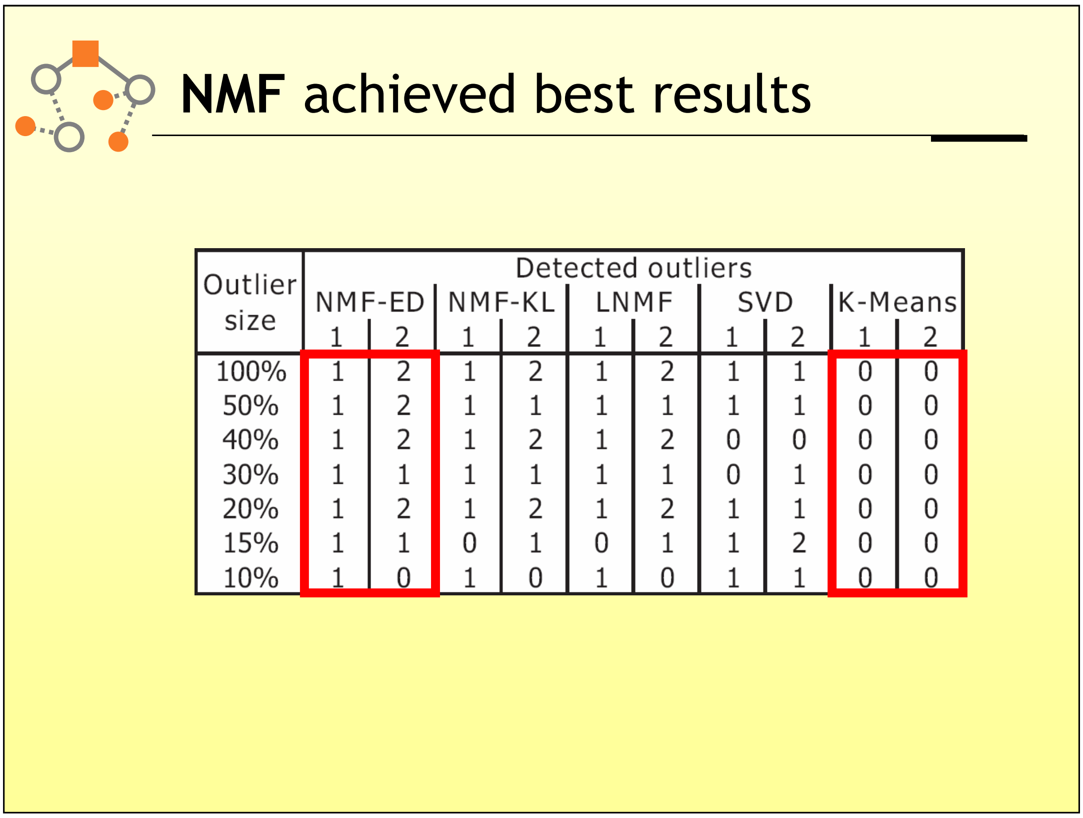
Finally, we'd like the clustering algorithm to put all input snippets into clusters, so that there are no unclustered content. So we define the **Snippet Coverage** measure, which is the percentage of the input snippets put into clusters. Obviously, the closer we get to 1, the better.



NMF achieved best results

Here you can see the average Cluster Contamination, Topic and Snippet Coverage for the description comes first approach using different matrix factorisations.

The smallest Cluster Contamination was achieved by the Nonnegative Matrix Factorizations (two slightly different variants of them) and the k-means based factorisation. Also, NMF-based clustering produced best Topic Coverage (about 90%) and Snippet Coverage (almost 80%).

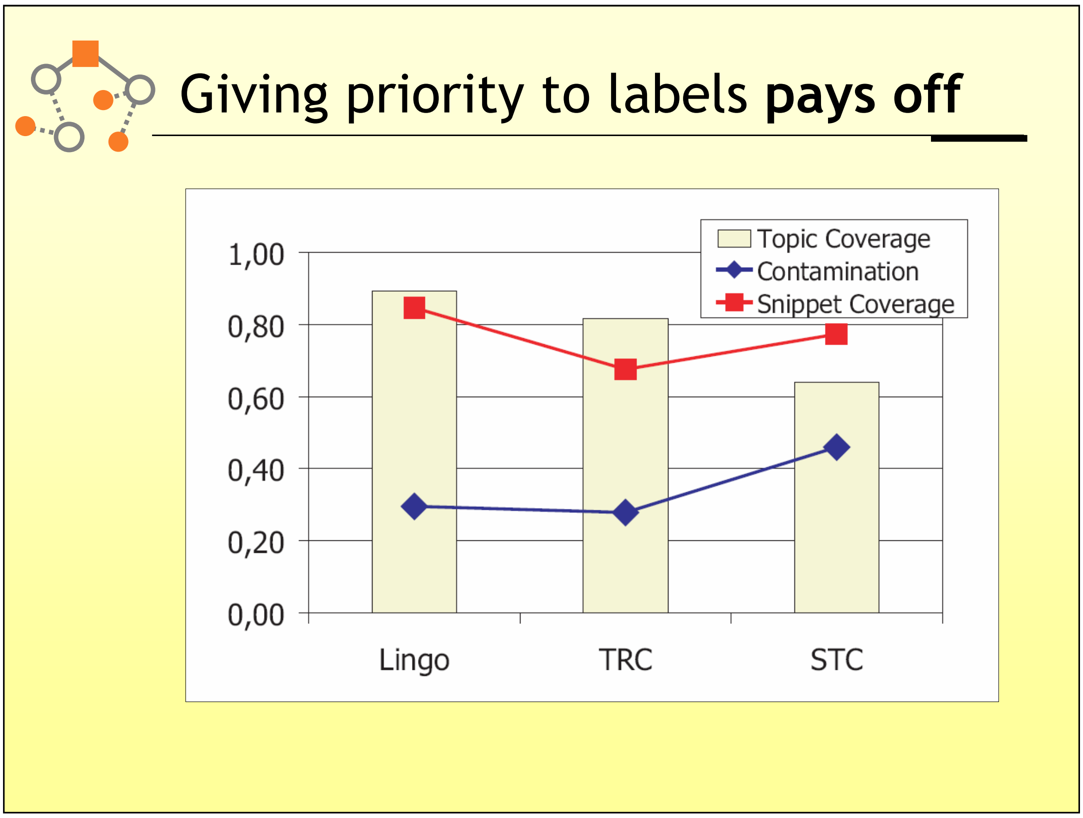


NMF achieved best results

Outlier size	Detected outliers									
	NMF-ED		NMF-KL		LNMF		SVD		K-Means	
	1	2	1	2	1	2	1	2	1	2
100%	1	2	1	2	1	2	1	1	0	0
50%	1	2	1	1	1	1	1	1	0	0
40%	1	2	1	2	1	2	0	0	0	0
30%	1	1	1	1	1	1	0	1	0	0
20%	1	2	1	2	1	2	1	1	0	0
15%	1	1	0	1	0	1	1	2	0	0
10%	1	0	1	0	1	0	1	1	0	0

Very interesting are the results of the outlier detection test. Here again **NMF proved the best** and successfuly dealt with data sets containing one outlier, but also handled quite well the test sets containing two outliers.

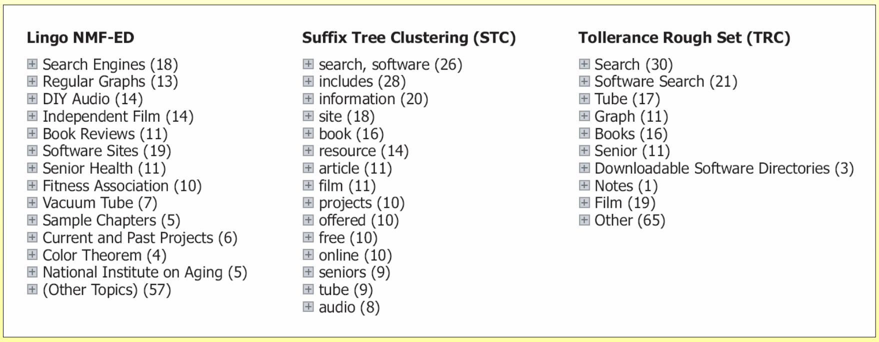
Interestingly, the **k-means based factorisation did not highlight any outliers**. The reason is that k-means locates the centroids in the most „dense” areas of the term space, and this is not where the outliers are.



Giving priority to labels **pays off**

We also compared the description comes first approach to clustering (implemented by Lingo) with two other algorithms designed specifically for search results clustering: Tollerance Rough Set Clustering and Suffix Tree Clustering. **Lingo seems to be quite a good competition** for them, at least with respect to the Contamination and Coverage measures.

Giving priority to labels **pays off**



Lingo NMF-ED

- ⊞ Search Engines (18)
- ⊞ Regular Graphs (13)
- ⊞ DIY Audio (14)
- ⊞ Independent Film (14)
- ⊞ Book Reviews (11)
- ⊞ Software Sites (19)
- ⊞ Senior Health (11)
- ⊞ Fitness Association (10)
- ⊞ Vacuum Tube (7)
- ⊞ Sample Chapters (5)
- ⊞ Current and Past Projects (6)
- ⊞ Color Theorem (4)
- ⊞ National Institute on Aging (5)
- ⊞ (Other Topics) (57)

Suffix Tree Clustering (STC)

- ⊞ search, software (26)
- ⊞ includes (28)
- ⊞ information (20)
- ⊞ site (18)
- ⊞ book (16)
- ⊞ resource (14)
- ⊞ article (11)
- ⊞ film (11)
- ⊞ projects (10)
- ⊞ offered (10)
- ⊞ free (10)
- ⊞ online (10)
- ⊞ seniors (9)
- ⊞ tube (9)
- ⊞ audio (8)

Tollerance Rough Set (TRC)

- ⊞ Search (30)
- ⊞ Software Search (21)
- ⊞ Tube (17)
- ⊞ Graph (11)
- ⊞ Books (16)
- ⊞ Senior (11)
- ⊞ Downloadable Software Directories (3)
- ⊞ Notes (1)
- ⊞ Film (19)
- ⊞ Other (65)

Finally, let’s have a look at the cluster labels generated by Lingo and the more conventional algorithms. As you can see, the labels generated by Lingo are quite descriptive and concise. Interestingly, Lingo avoided creating very generic and usually not terribly useful clusters like „Free” or „Online”.

Advertisement :-)

Carrot^2

Search Results Clustering Framework

http://www.carrot2.org

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

Implementations of all the algorithms I presented here are Open Source and available as part of the **Carrot2 Search Results Clustering Framework**. You can download and experiment with them free of charge, you can also try the online live demo.

Thank you

Improving Quality of Search Results Clustering with Approximate Matrix Factorisations

Stanisław Osiński

Poznan Supercomputing and Networking Center