Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition

Stanisław Osiński Jerzy Stefanowski Dawid Weiss

Institute of Computing Science
Poznań University of Technology

May 20th, 2004

For a good start. . .

For a good start. . .

Placeholder for a really good joke

Some background: textual information clustering

- Full text clustering
  - Goals: split documents into semantically related collections or recreate known partitioning
  - Input: usually full text of the input documents
  - Output: discovered structure of clusters

Some background: textual information clustering

- Full text clustering
  - Goals: split documents into semantically related collections or recreate known partitioning
  - Input: usually full text of the input documents
  - Output: discovered structure of clusters
- Classification ("clusters" known in advance)
  - Goals: establish most probable assignment of first-time-seen documents
  - Output: document-cluster assignments

Some background: textual information clustering

Question

What about cluster descriptions? $\longrightarrow$ **usually neglected**

Some background: textual information clustering

Clustering of search results

- Goals: organize search results into groups, *describe* them to shorten the inspection time
- Input: fragments of original documents ("snippets" and titles)
- Output: groups of semantically related results and their meaningful descriptions

Linear search result for “IIPWM” query

1 **New Trends in Intelligent Information Processing and Web Mining ...**
The conference will have special tracks on: Artificial Immune Systems, Search Engines, Con
http://iipwm.ipipan.waw.pl/

2 **DEADLINES**
DEADLINES: 15th October, 2003 - paper submission deadline Upon many requests we have
http://iipwm.ipipan.waw.pl/2004/dates.html

3 **EvoWeb - Conference record for IIPWM '04**
... You are here: Home, News & events, Conferences, Record, IIPWM '04. ... Dates: 17 - 20
http://evonet.lri.fr/evoweb/news_events/conferences/record.php?id=957

4 **unpublished{ iipwm -osinski2004-2, author = "Stanis{l{}}aw Osi ...**
@unpublished{ iipwm -osinski2004-2, author = "Stanis{l{}}aw Osi{'{n}}ski and Jerzy Stefa
http://www.cs.put.poznan.pl/dweiss/site/publications/bibtex/iipwm-osinski2004-2.bib

5 **IIPWM - Call for Papers**
File Format: Unrecognized - View as HTML ... IIPWM - Call for Papers. • From: IIPWM Conf. (
http://www.mail-archive.com/inductive@listserv.unb.ca/msg00641.html

6 **Machine Learning List: Vol. 15, No. 14**
File Format: Unrecognized - View as HTML ... Genetic and Evolutionary Computation COnfere
http://www.mail-archive.com/ml@isle.org/msg00020.html

7 **All books by International IIS: Iipwm '03 Conference / Wed Nov 26 6 ...**
All books by International IIS: Iipwm '03 Conference. All books by International IIS: Iipwm

Sample clusters for "IIPWM" query

Question

Is this a new research problem?

Question

Is this a new research problem?

It seems that the answer is *yes*; "conceptual clustering" (?) combines at least two difficult tasks:

Question

Is this a new research problem?

It seems that the answer is *yes*; "conceptual clustering" (?) combines at least two difficult tasks:

- text clustering (what are similar documents?)

Question

Is this a new research problem?

It seems that the answer is *yes*; "conceptual clustering" (?) combines at least two difficult tasks:

- text clustering (what are similar documents?)
- discovering quality description (what constitutes a correct, informative description?)

Question

Is this a new research problem?

It seems that the answer is *yes*; "conceptual clustering" (?) combines at least two difficult tasks:

- text clustering (what are similar documents?)
- discovering quality description (what constitutes a correct, informative description?)

Challenges:

- objective difficulties: language properties (syntax, inflection, text segmentation), definition of similarity between documents

Question

Is this a new research problem?

It seems that the answer is *yes*; "conceptual clustering" (?) combines at least two difficult tasks:

- text clustering (what are similar documents?)
- discovering quality description (what constitutes a correct, informative description?)

Challenges:

- objective difficulties: language properties (syntax, inflection, text segmentation), definition of similarity between documents
- subjective difficulties: "good" cluster label choice

Solving the problem: the idea

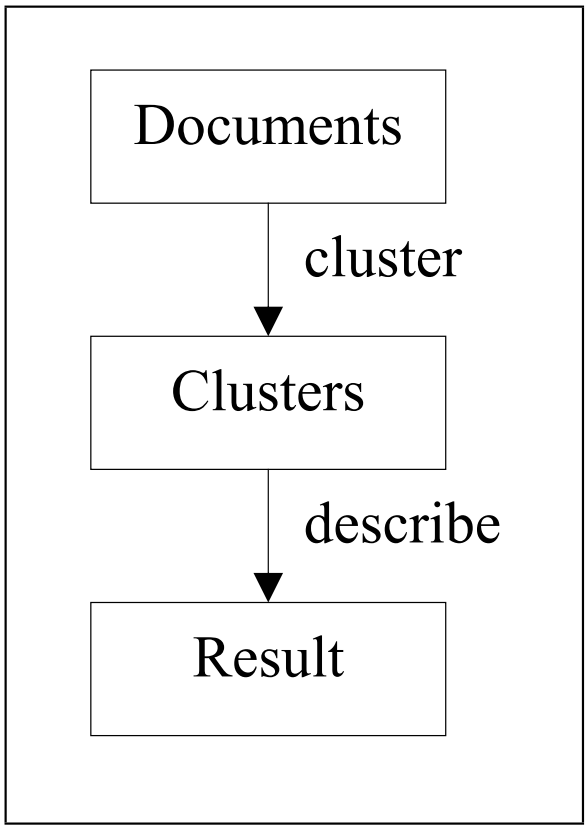
Split the process into *two independent phases*:

- cluster label candidate discovery,
- clusters discovery

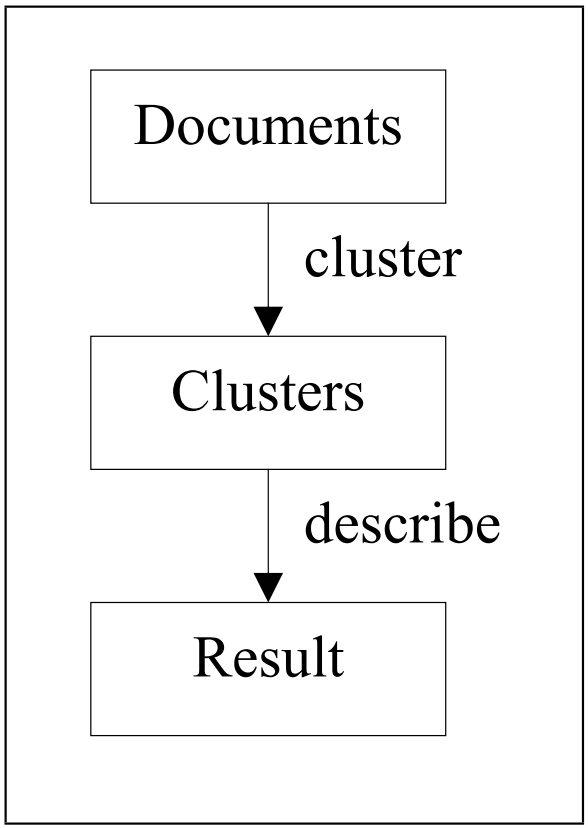
and *combine* them to produce the desired effect.

Originally proposed by Vivisimo

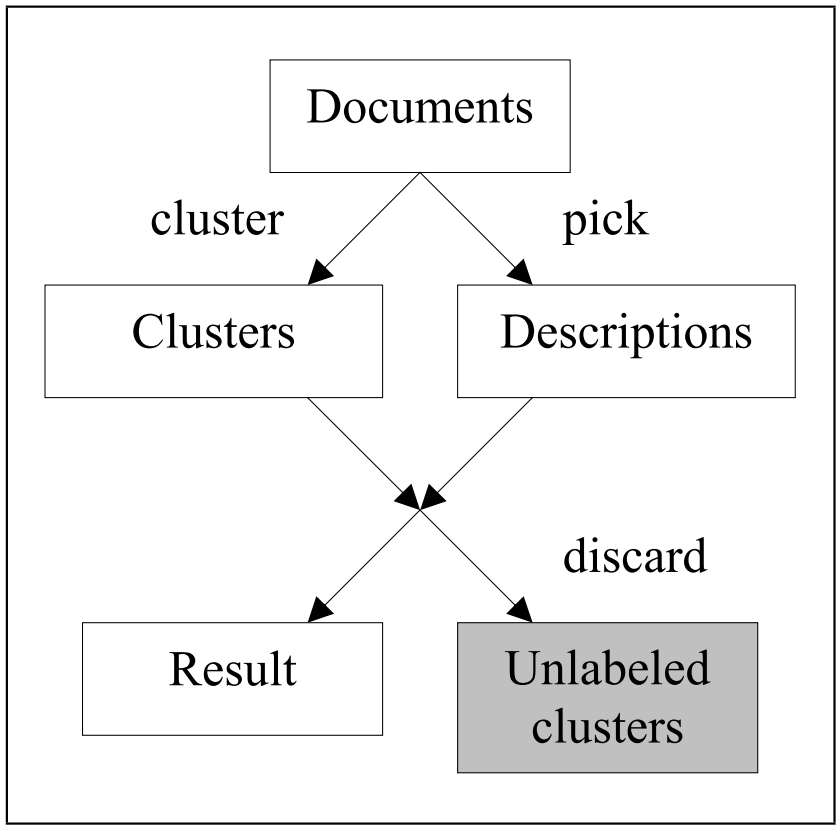
"classic" clustering



"classic" clustering



"conceptual" clustering

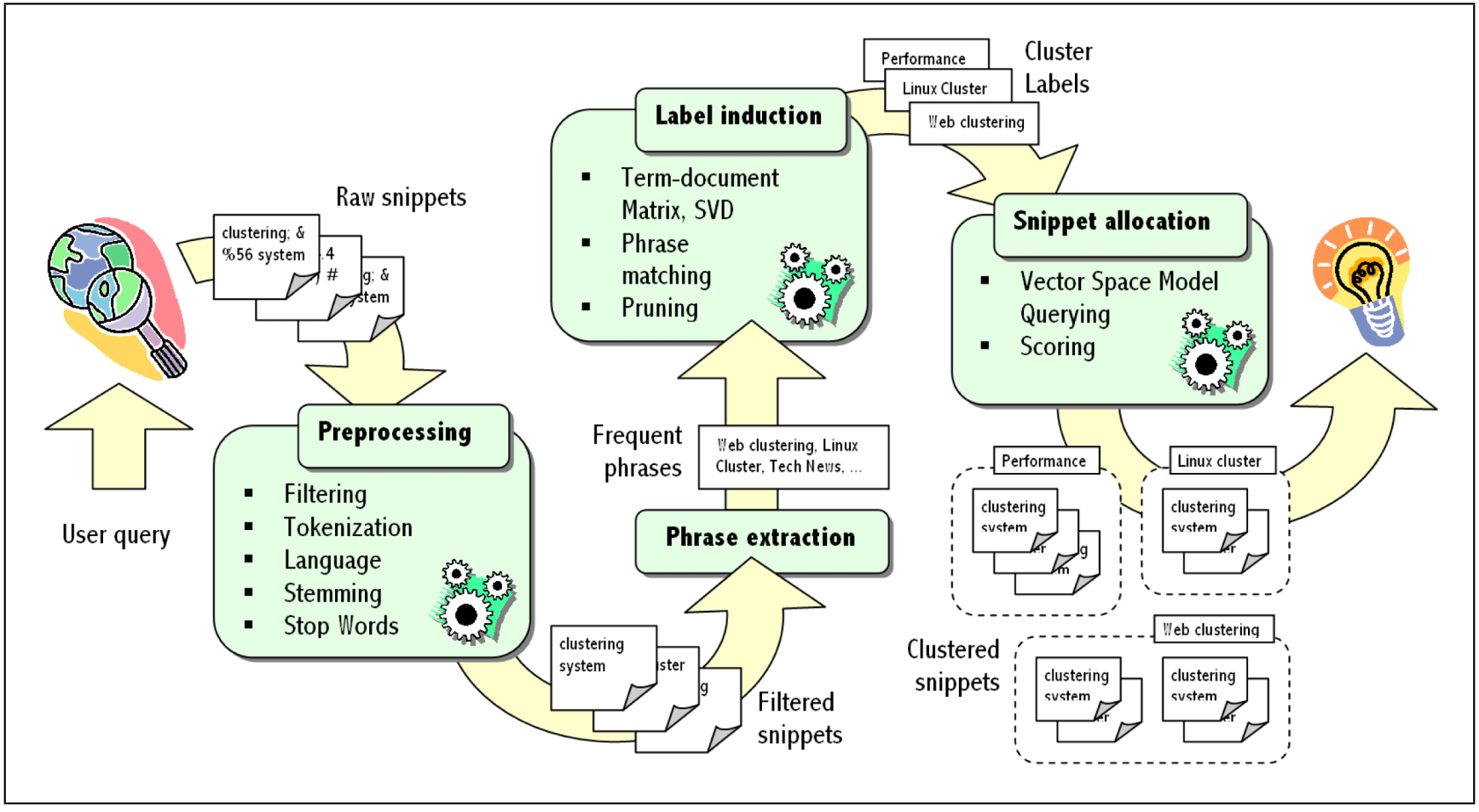


The Lingo algorithm

Instantiation of the idea

Lingo: description-comes-first

Lingo: the idea's concrete example



Lingo combines:

- SVD — successful at finding clusters
- phrases discovery — usually good label indicators

Lingo: data preprocessing and frequent phrase extraction

- Step 1: Data preprocessing
  The usual: stemming (Porter stemmer, Lametyzator), stop words marking, text segmentation heuristic
- Step 2: Frequent phrase extraction and cluster label induction
  Discover *complete phrases* in the input text
  - maximum length term subsequences
  - occurring at least TERM FREQUENCY THRESHOLD times
  - do not cross sentence boundaries
  - no stop words at ends

Lingo: cluster label induction

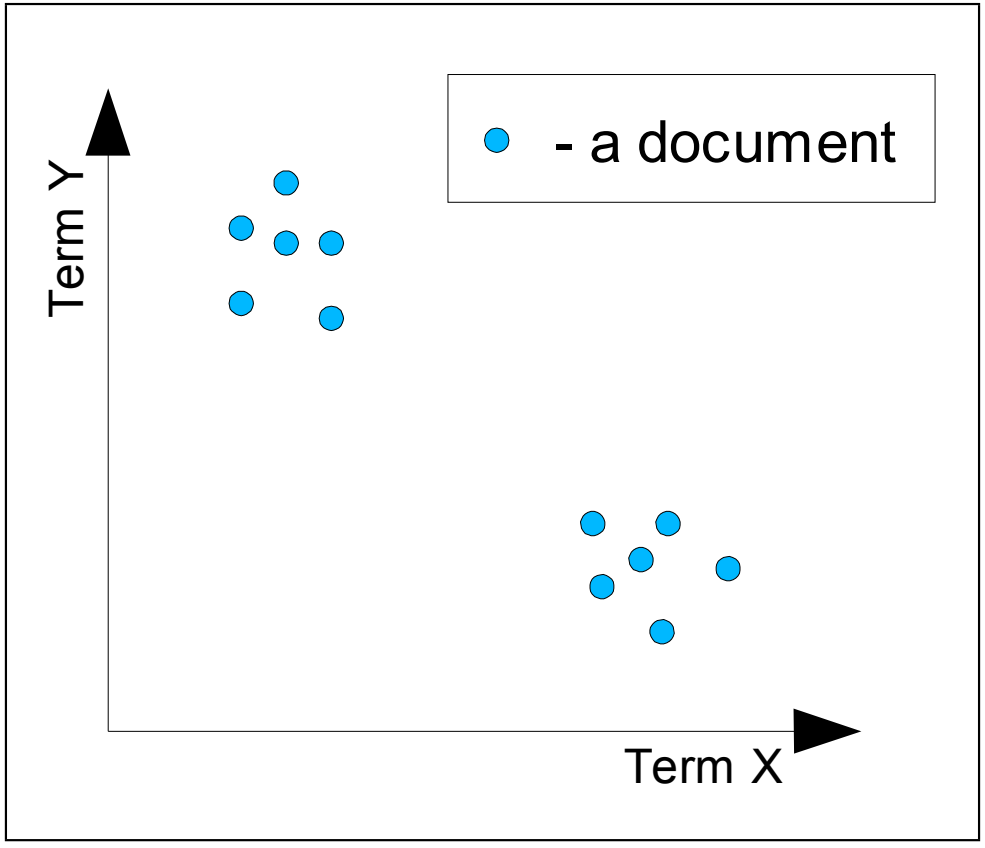
Data representation

Term-document matrix A: representation of documents as *vectors* of feature weights.

Lingo: cluster label induction

Data representation

Term-document matrix A: representation of documents as *vectors* of feature weights.



Latent semantic structure discovery

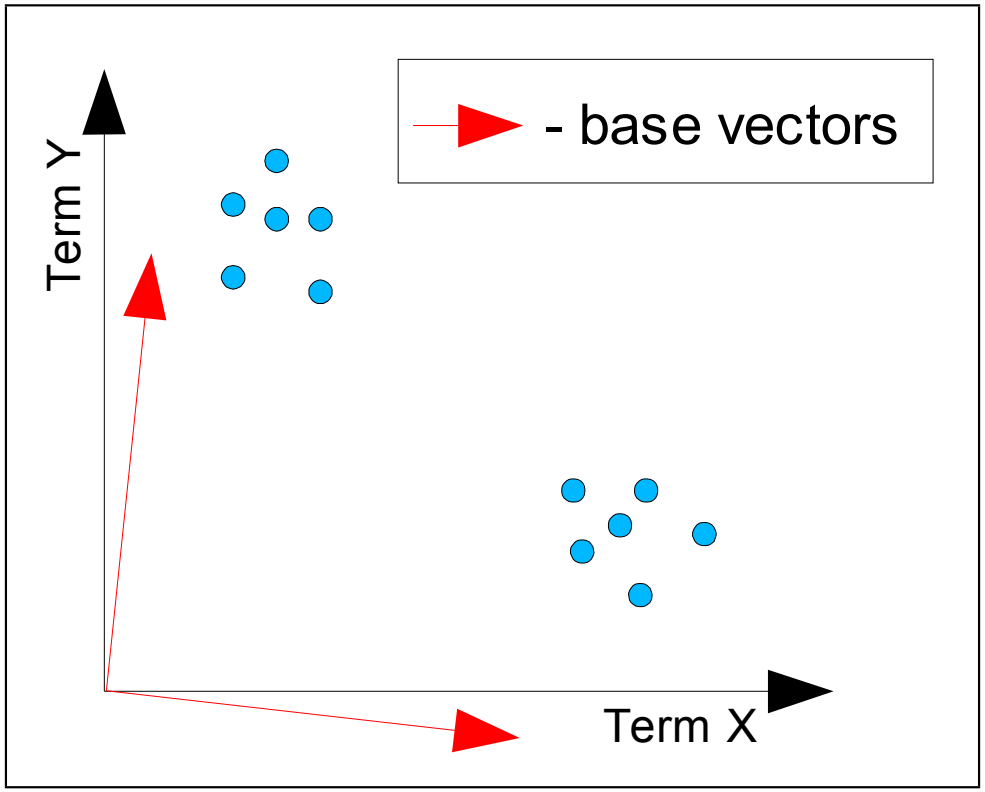
Matrix A is decomposed using SVD to aquire an orthogonal base in the multidimensional feature space:

$$\underbrace{\begin{bmatrix} * & * \\ * & * \\ * & * \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} * & \\ & * \end{bmatrix}}_{S} \underbrace{\begin{bmatrix} * & * \\ * & * \end{bmatrix}}_{V^T}$$

Latent semantic structure discovery

Matrix A is decomposed using SVD to aquire an orthogonal base in the multidimensional feature space:

$$\underbrace{\begin{bmatrix} * & * \\ * & * \\ * & * \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} * & * & * \\ * & * & * \\ * & * & * \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} * & \\ & * \end{bmatrix}}_{S} \underbrace{\begin{bmatrix} * & * \\ * & * \end{bmatrix}}_{V^T}$$



Cluster label induction

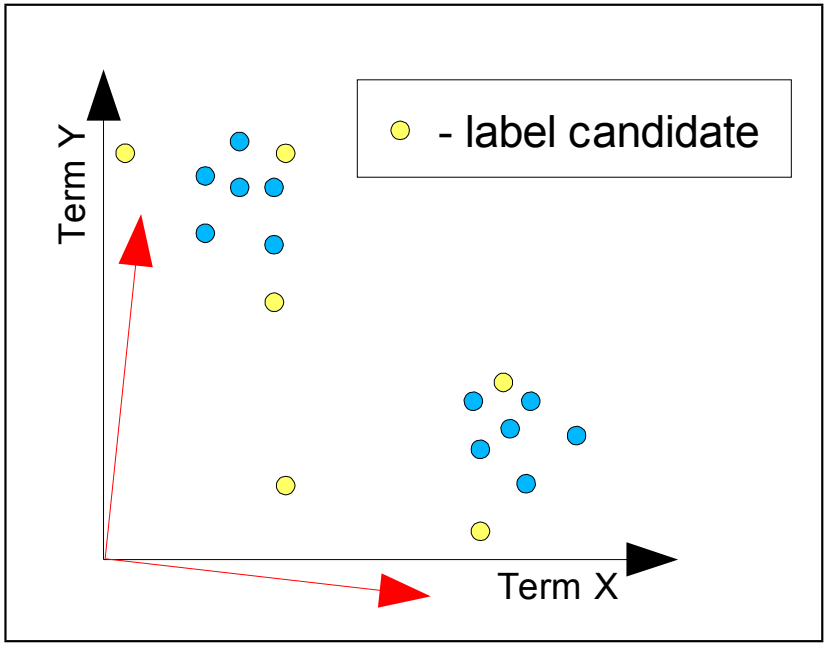
Observation

Cluster label candidates are expressed in the same vector space as documents in matrix A. One can easily calculate their similarity to the first k vectors of the base: $M = U_k^T P$.

Cluster label induction

Observation

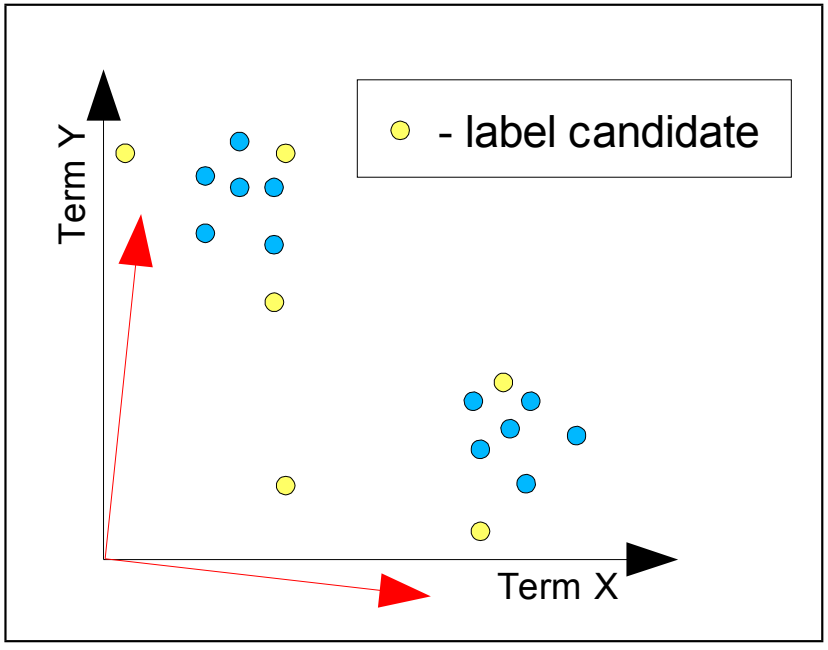
Cluster label candidates are expressed in the same vector space as documents in matrix A. One can easily calculate their similarity to the first k vectors of the base: $M = U_k^T P$.

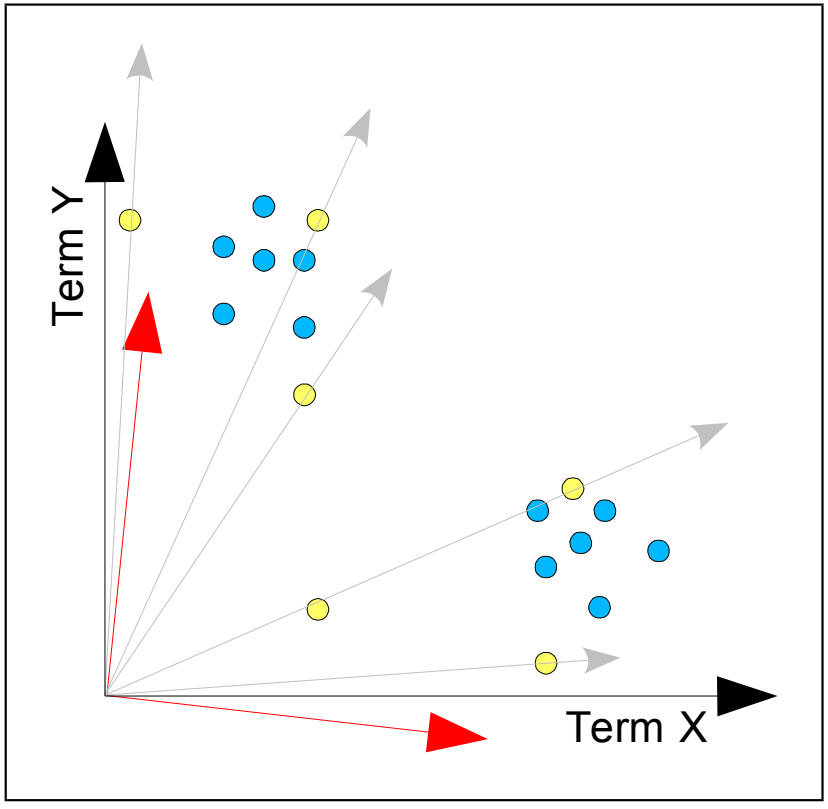


Cluster label induction

Observation

Cluster label candidates are expressed in the same vector space as documents in matrix A. One can easily calculate their similarity to the first k vectors of the base: $M = U_k^T P$.



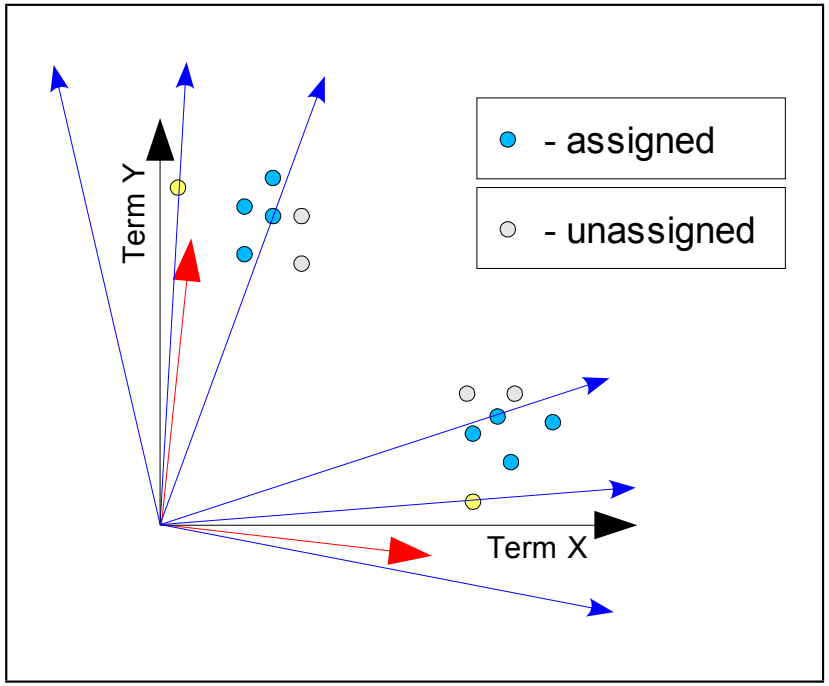


Lingo: cluster content discovery and cluster formation

- Step 3: Cluster content discovery
  Apply Vector Space Model technique to look for documents in proximity of search labels

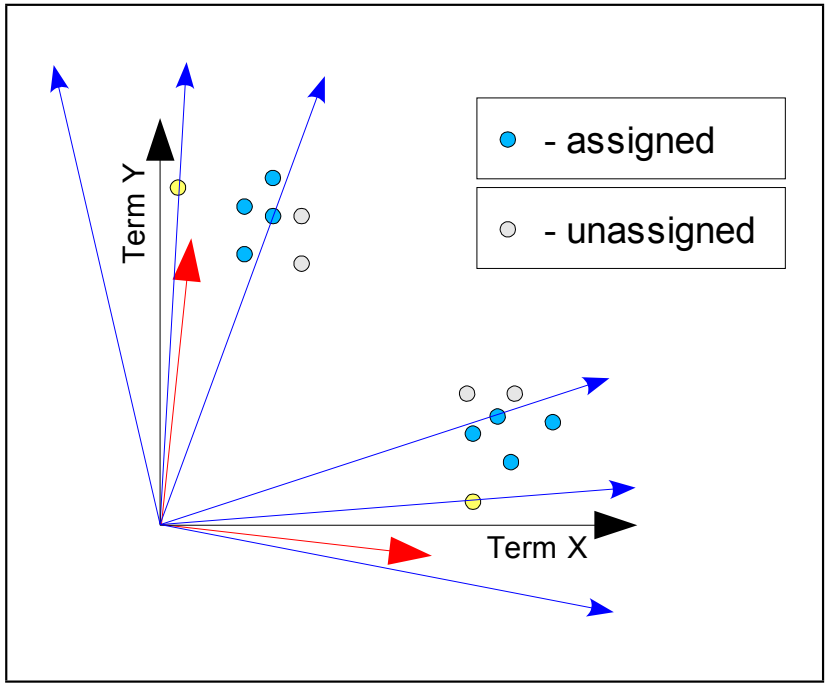
Lingo: cluster content discovery and cluster formation

- Step 3: Cluster content discovery
  Apply Vector Space Model technique to look for documents in proximity of search labels



Lingo: cluster content discovery and cluster formation

- Step 3: Cluster content discovery
  Apply Vector Space Model technique to look for documents in proximity of search labels



- Step 4: Final cluster formation
  Count the scores for each cluster, sort them, display them

Algorithm's evaluation

- Common evaluation difficulties
  - Quality of cluster labels subjective and hard to measure
  - Many different "good" cluster sets for the same data
  - ...
- Empirical evaluation experiment
  - 7 human evaluators (mixed background)
  - 4 clustered search results, 2 Polish and 2 English queries
  - Questions asked:
    - Are cluster labels meaningful?
    - Are document assignments within a group sensible?

Results of the experiment

- 70-70% clusters marked as useful
- 80-95% documents (snippets) matched a cluster's topic

Results of the experiment

- 70-70% clusters marked as useful
- 80-95% documents (snippets) matched a cluster's topic

Questions...

- Was this experiment informative?

Results of the experiment

- 70-70% clusters marked as useful
- 80-95% documents (snippets) matched a cluster's topic

Questions...

- Was this experiment informative?
- Was it helpful?

Results of the experiment

- 70-70% clusters marked as useful
- 80-95% documents (snippets) matched a cluster's topic

Questions...

- Was this experiment informative?
- Was it helpful?
- Did it prove anything?

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

Thank you. questions?