TECHNIQUES OF CLUSTERING

(a short review for students)

Mikhail Alexandrov^{1,2}, Pavel Makagonov^{3}

[1] Autonomous University of Barcelona, Spain
[2] Social Network Research Center with UCE, Slovakia
[3] Mixtec Technological University, Mexico
dyner1950@mail.ru, mpp@mixteco.utm.mx

Petersburg 2008

CONTENTS

Introduction ←

Definitions

Clustering

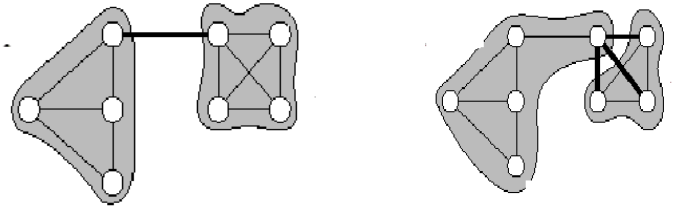
Discussion

Open Problems

Ideas, Materials, Collaboration

Prof. Dr. Benno Stein
Dr. Sven Meyer zu Eissen

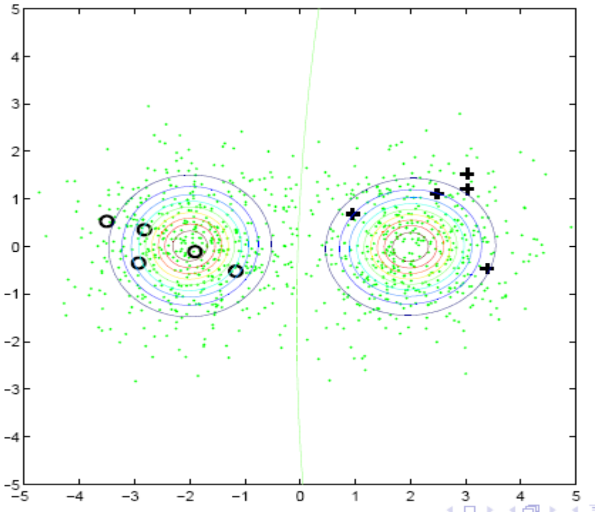
Weimar University, **Germany**



- **Structuring and Indexing**
- **AI search in IR**

Dr. Xiaojin Zhu

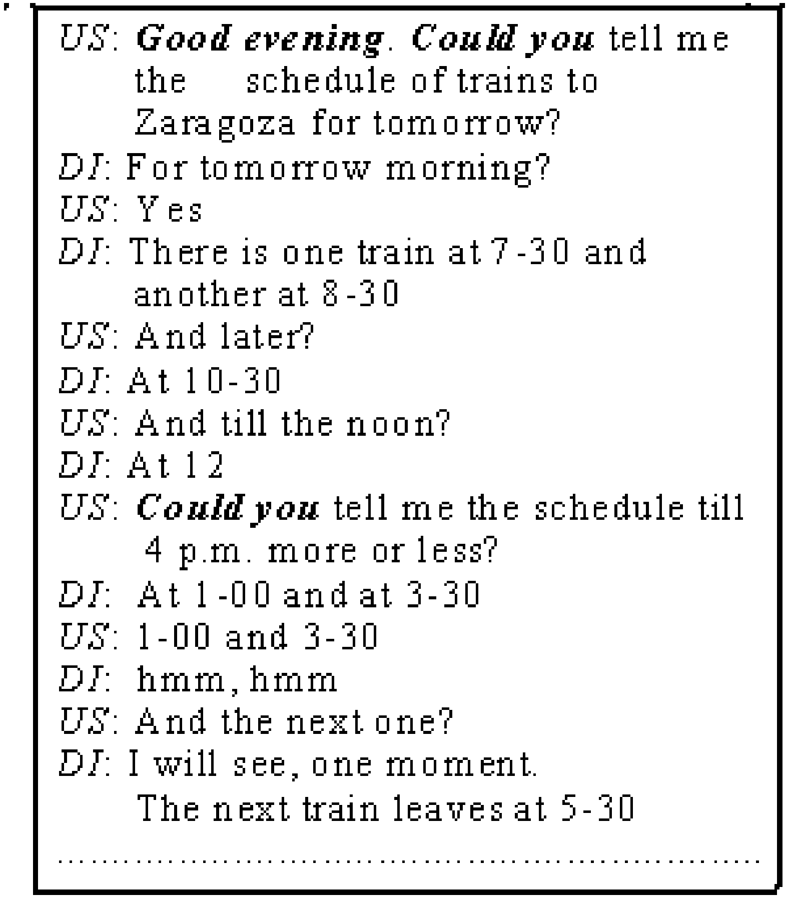
University of Wisconsin, Madison, **USA**



- **Semi-Supervised Learning**

Subject of Grouping

TEXTUAL DATA

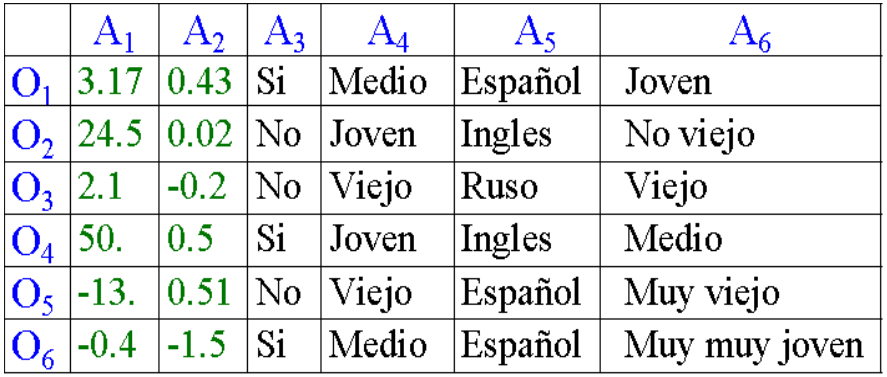


US: ***Good evening*. *Could you*** tell me
the schedule of trains to
Zaragoza for tomorrow?
DI: For tomorrow morning?
US: Yes
DI: There is one train at 7-30 and
another at 8-30
US: And later?
DI: At 10-30
US: And till the noon?
DI: At 12
US: ***Could you*** tell me the schedule till
4 p.m. more or less?
DI: At 1-00 and at 3-30
US: 1-00 and 3-30
DI: hmm, hmm
US: And the next one?
DI: I will see, one moment.
The next train leaves at 5-30

....

Example: typical dialog between passenger (**US**) and railway directory inquires (**DI**)

NON TEXTUAL DATA



	A_1	A_2	A_3	A_4	A_5	A_6
O_1	3.17	0.43	Si	Medio	Español	Joven
O_2	24.5	0.02	No	Joven	Ingles	No viejo
O_3	2.1	-0.2	No	Viejo	Ruso	Viejo
O_4	50.	0.5	Si	Joven	Ingles	Medio
O_5	-13.	0.51	No	Viejo	Español	Muy viejo
O_6	-0.4	-1.5	Si	Medio	Español	Muy muy joven

Local Terminology

It is not important what is the source of data: textual or non textual.

Data: work in the space of **numerical parameters**

Texts: work in the space of **words**

Presentation of Textual Data

TEXTS (‘indexed’)

*During four years we have researched methods helping the **healthcare** professional to continuosly monitor, intelligently retrieve, evaluate and manage **medical** documents available as **electronic texts** (e-texts). We developed tools for **semi-automatic** processing...*

Vector model (‘parameterized’)

healthcare	*medical*	*electronic texts*	*semi-automatic*
12	17	2	1

Local Terminology

Indexed texts are only parameterized texts in the **space of words**

Presentation of Textual Data

TEXTS (‘indexed’)

Themes	*Doc 1*	*Doc 2*
medicine, pensioners, privileges	4,3,0	2,4,0
transport, queue, time-table, privileges	2,4,3,1	0,0,0,0
police, security, corruption	2,1,0	0,0,0
environment, water, air, cleanline	0,0,0,0	2,1,3,2

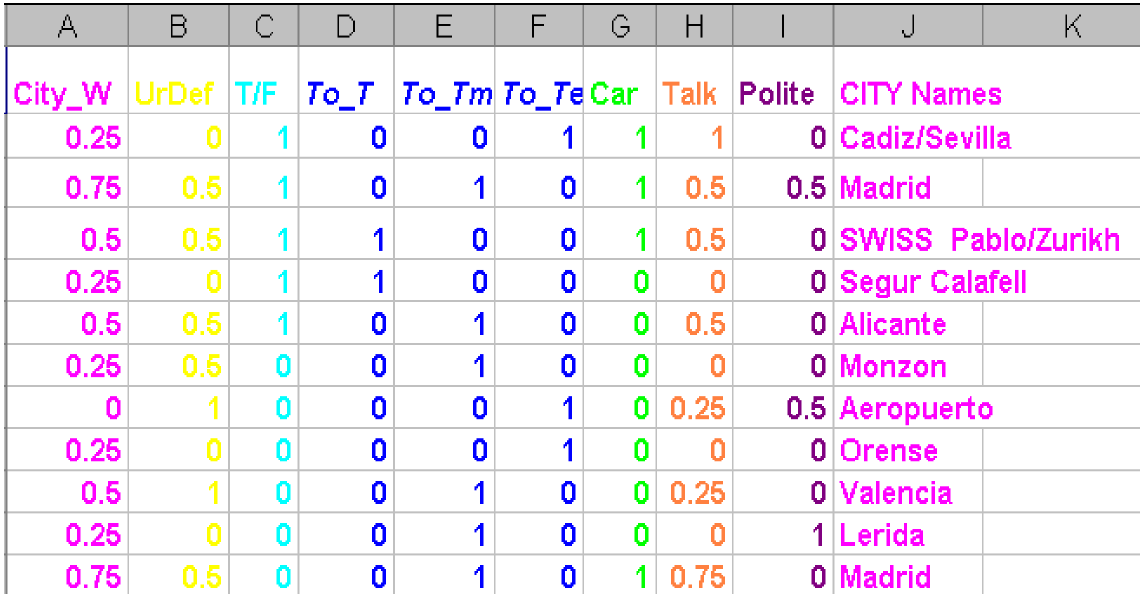
Local Terminology

Indexed texts are only parameterized texts in the **space of themes**

=> **Category/Context Vector Models...**

TEXTS (‘parameterized’)

Example: manually parameterized dialogs in the **space of parameters** (transport service and passenger needs)



A	B	C	D	E	F	G	H	I	J	K
City_W	UrDef	T/F	*To_T*	*To_Tm*	*To_Te*	Car	Talk	Polite	CITY Names	
0.25	0	1	0	0	1	1	1	0	Cadiz/Sevilla	
0.75	0.5	1	0	1	0	1	0.5	0.5	Madrid	
0.5	0.5	1	1	0	0	1	0.5	0	SWISS Pablo/Zurikh	
0.25	0	1	1	0	0	0	0	0	Segur Calafell	
0.5	0.5	1	0	1	0	0	0.5	0	Alicante	
0.25	0.5	0	0	1	0	0	0	0	Monzon	
0	1	0	0	0	1	0	0.25	0.5	Aeropuerto	
0.25	0	0	0	0	1	0	0	0	Orense	
0.5	1	0	0	1	0	0	0.25	0	Valencia	
0.25	0	0	0	1	0	0	0	1	Lerida	
0.75	0.5	0	0	1	0	1	0.75	0	Madrid	

CONTENTS

Introduction

Definitions ←

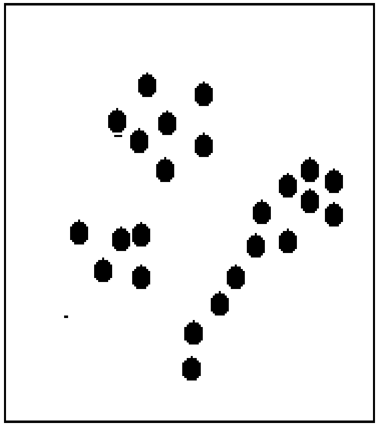
Clustering

Discussion

Open Problems

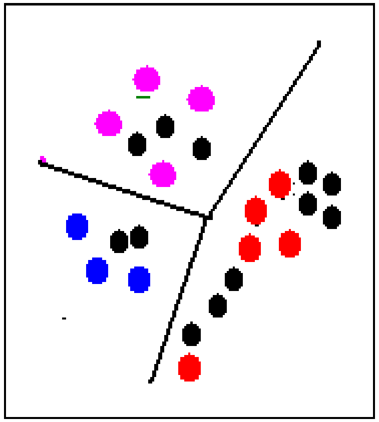
Types of Grouping

Unsupervised Learning



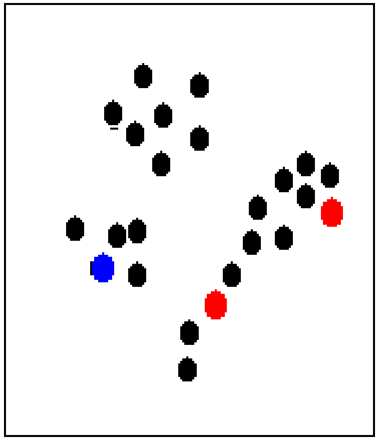
We know nothing about data sructure

Supervised Learning



We know well data sructure

Semi-Supervised Learning



We know something about data sructure

Types of Grouping

Clustering

Characteristics:

Absence of **patterns** or **descriptions** of classes, so the results are defined by the **nature** of the data themselves ($N > 1$)

Synonyms:

Classification **without teacher**

Unsupervised learning

Number of clusters

[ ] is known **exactly**

[x] is known **approximately**

[ ] **is not** known => searching

Classification

***Characteristics*:**

Presence of **patterns** or **descriptiones** of classes, so the results are defined by the **user** ($N \geq 1$)

***Synonyms*:**

Classification **with teacher**

Supervised learning

***Specials terms*:**

Categorization (of documents)

Diagnostics (technics, medicine)

Recognition (technics, science)

Types of Grouping

“Semi Clustering/Classification”

Characteristics:

Presence of **limited number** of patterns, so the results are defined both by the **user** or by the **data** themselves (N >1)

Synonyms:

Semi-Classification
Semi Supervised learning

Number of clusters/categories

[ ] is known **exactly**
[x] is known **approximately**
[ ] **is not** known => searching

Classification

***Characteristics*:**

Presence of **patterns** or **descriptiones** of classes, so the results are defined by the **user** (N>=1)

***Synonyms*:**

Classification **with teacher**
Supervised learning

***Specials terms*:**

Categorization (of documents)
Diagnostics (technics, medicine)
Recognition (technics, science)

Objectives of Grouping

1. Organization (structuring) of an object set

Process is named **data structuring**

2. Searching interesting patterns

Process is named **navigation**

3. Grouping for other applications:

- **Knowledge** discovery (clustering)
- **Summarization** of documents

Note:

Do not mix the **type** of grouping and its **objective**

Classification of methods

Based on belonging to cluster/category

Exclusive methods

Every object belongs only to one cluster/category. Methods are named **hard** grouping methods

Non-exclusive methods

Every object can belong to several clusters/categories. Methods are named **soft** grouping methods.

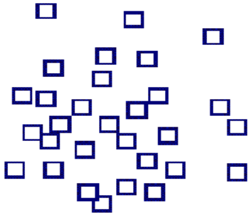
Based on data presentation

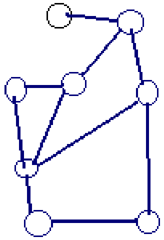
Methods oriented on free metric space

Every object is presented as a point in a free space

Methods oriented on graphs

Every object is presented as an element on graph





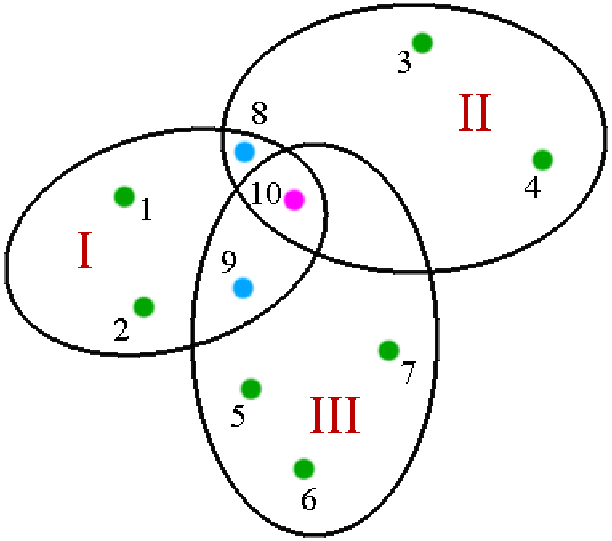
Fuzzy Grouping

Hard grouping

- **Hard** clustering
- **Hard** categorization

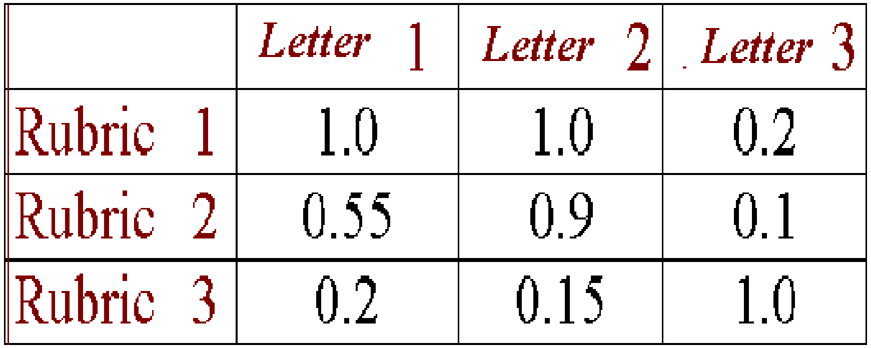
Soft grouping

- **Soft** clustering
- **Soft** categorization



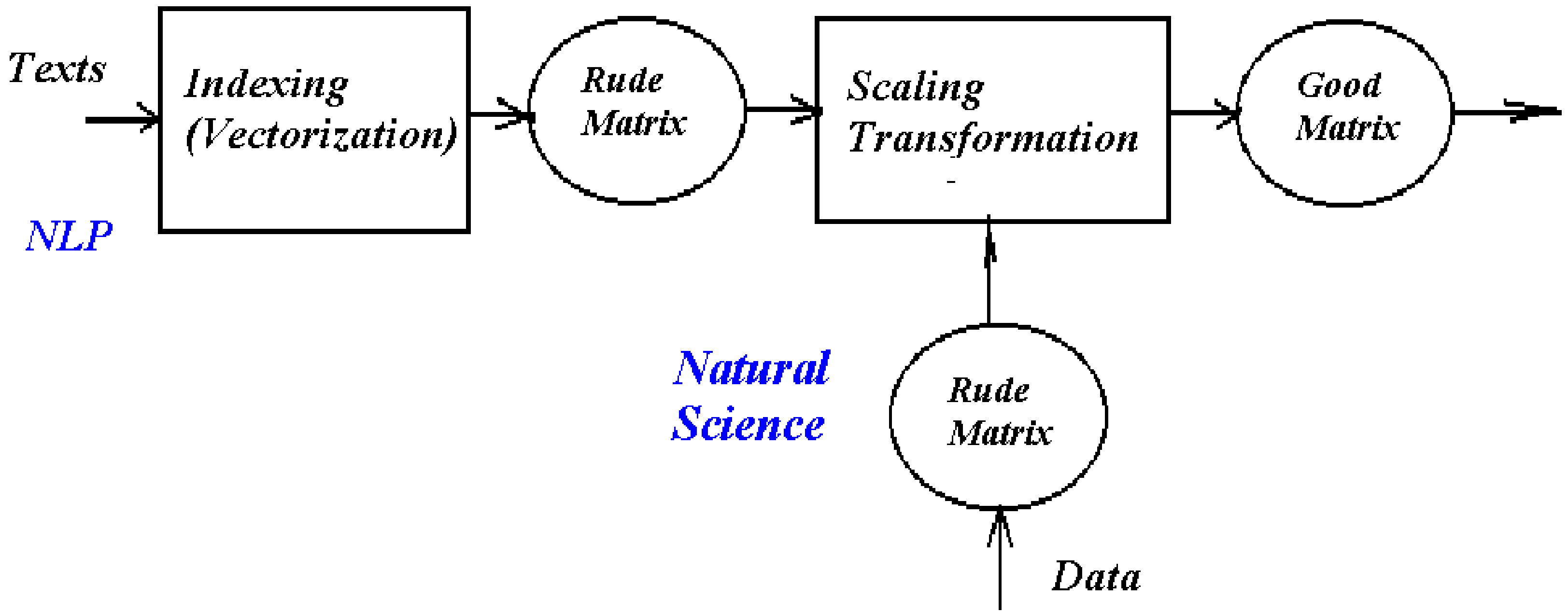
Example

The distribution of letters of Moscovites to the Government is **soft categorization** (numbers in the table reflect the relative weight of each theme)



	Letter 1	*Letter* 2	*Letter* 3
Rubric 1	1.0	1.0	0.2
Rubric 2	0.55	0.9	0.1
Rubric 3	0.2	0.15	1.0

General Scheme of Clustering Process - I



Here:

Both rude and good matrixes are matrix **Object/Attributes**

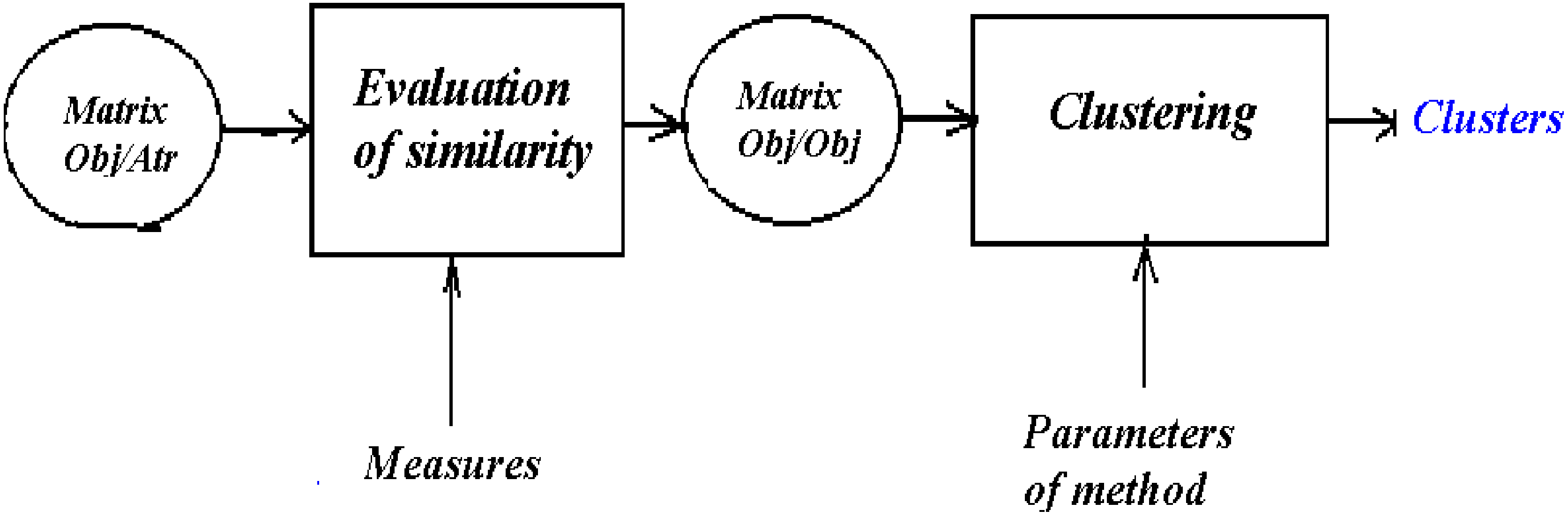
Preprocessing <= Processing

Principal idea :

To transform texts to **numerical** form in order to use **matematical** tools

Remember: our problem is grouping textual documents but not undestanding

General Scheme of Clustering Process - II

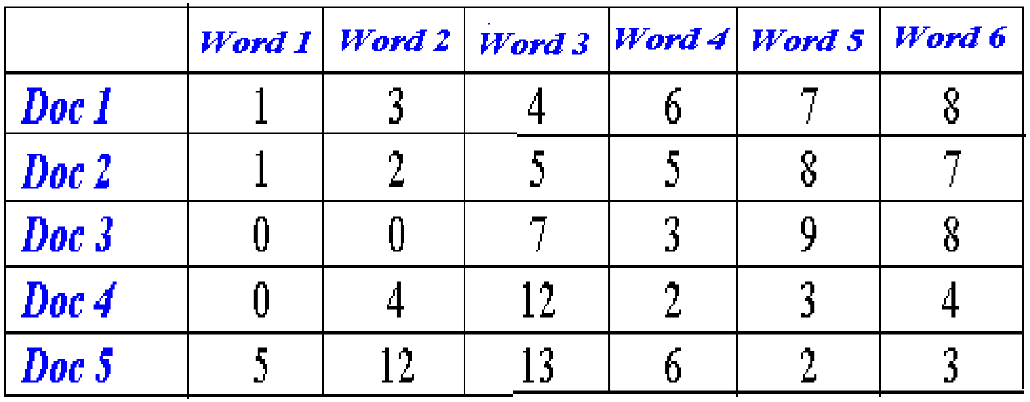


Preprocessing

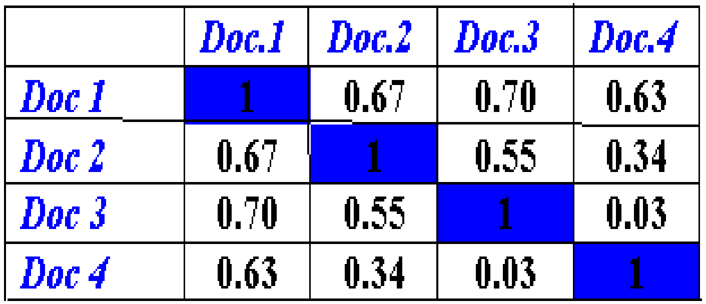
Processing <=

Here:
matrix **Attribute/Attribute**
can be used instead of matrix
Object/Object

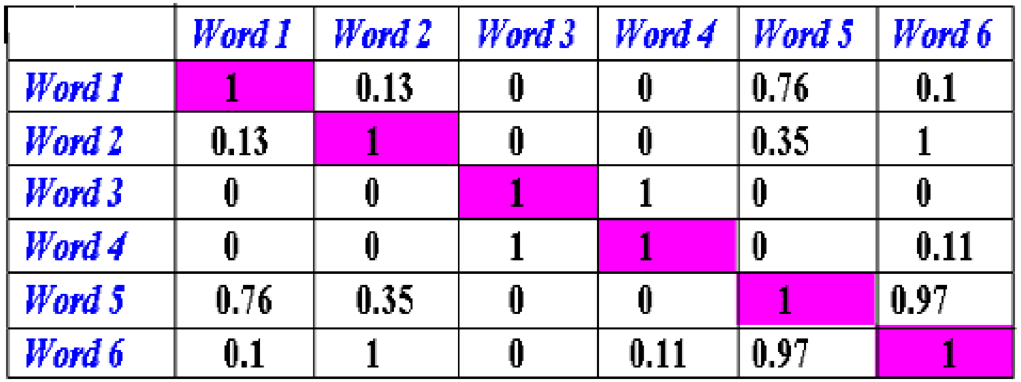
Matrixes to be Considered



	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6
Doc 1	1	3	4	6	7	8
Doc 2	1	2	5	5	8	7
Doc 3	0	0	7	3	9	8
Doc 4	0	4	12	2	3	4
Doc 5	5	12	13	6	2	3



	Doc.1	Doc.2	Doc.3	Doc.4
Doc 1	1	0.67	0.70	0.63
Doc 2	0.67	1	0.55	0.34
Doc 3	0.70	0.55	1	0.03
Doc 4	0.63	0.34	0.03	1



	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6
Word 1	1	0.13	0	0	0.76	0.1
Word 2	0.13	1	0	0	0.35	1
Word 3	0	0	1	1	0	0
Word 4	0	0	1	1	0	0.11
Word 5	0.76	0.35	0	0	1	0.97
Word 6	0.1	1	0	0.11	0.97	1

Clustering for Categorization

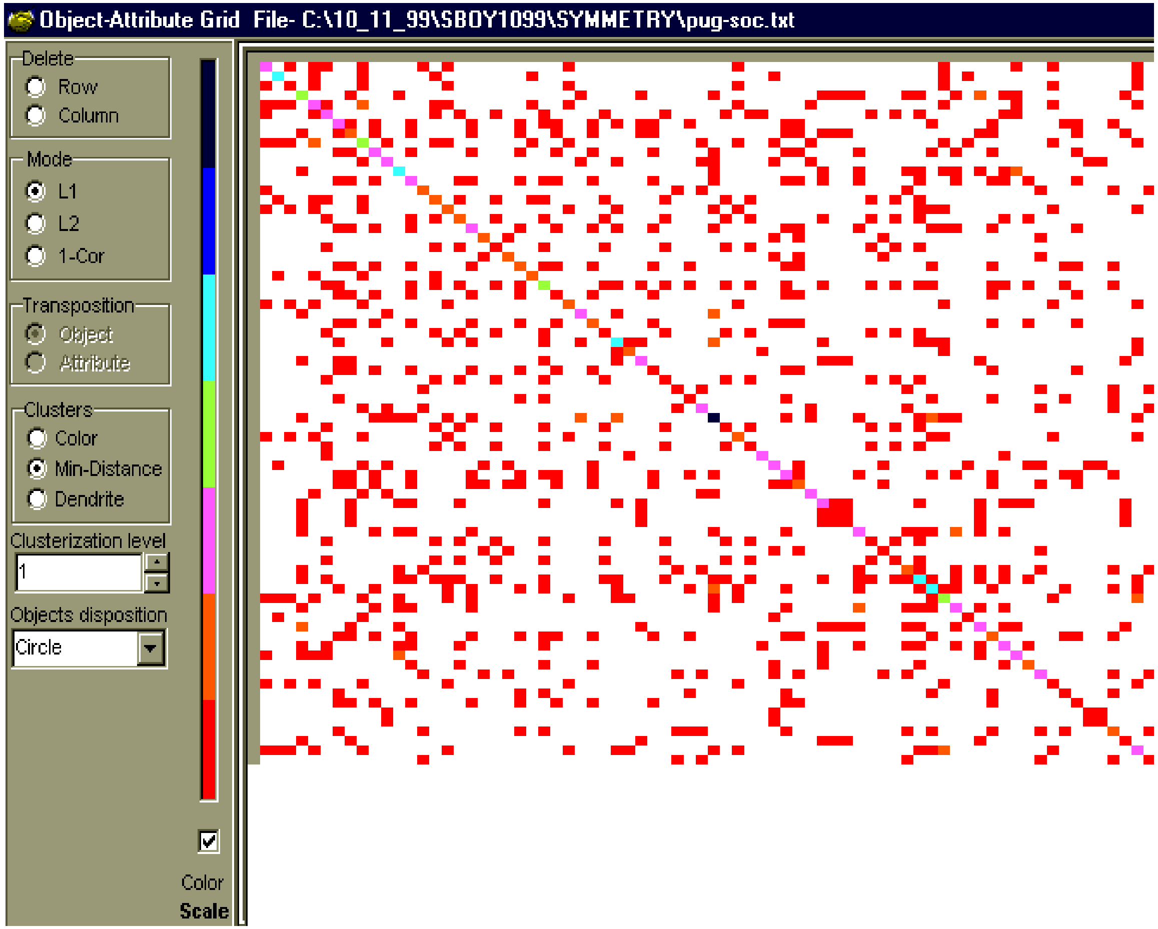
Colour matrix “words-words”

before clustering

Matrix contains the value of word co-occurrences in texts.

Red: if value more than some threshold.

White: if less.



Object-Attribute Grid File- C:\10_11_99\SBOY1099\SYMMETRY\pug-soc.txt

Delete
○ Row
○ Column

Mode
● L1
○ L2
○ 1-Cor

Transposition
● Object
○ Attribute

Clusters
○ Color
● Min-Distance
○ Dendrite

Clusterization level: 1

Objects disposition: Circle

☑ Color Scale

Clustering for Categorizatión

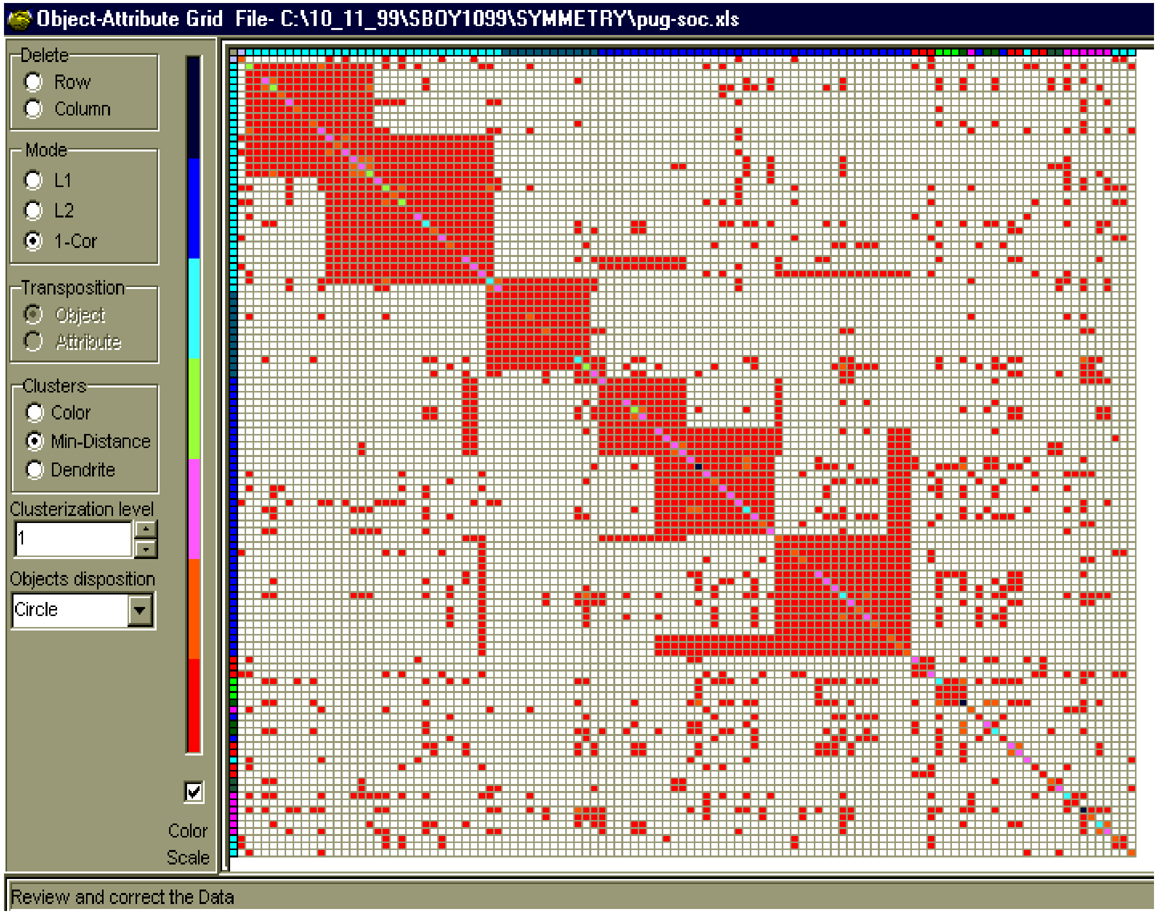
Colour matriz “words-words”

after clustering

Words are groupped.

Cluster => **Subdictionary**

Absence of blocks means absence of **Subthemes**



Object-Attribute Grid File- C:\10_11_99\SBOY1099\SYMMETRY\pug-soc.xls

Delete
○ Row
○ Column

Mode
○ L1
○ L2
● 1-Cor

Transposition
● Object
○ Attribute

Clusters
○ Color
● Min-Distance
○ Dendrite

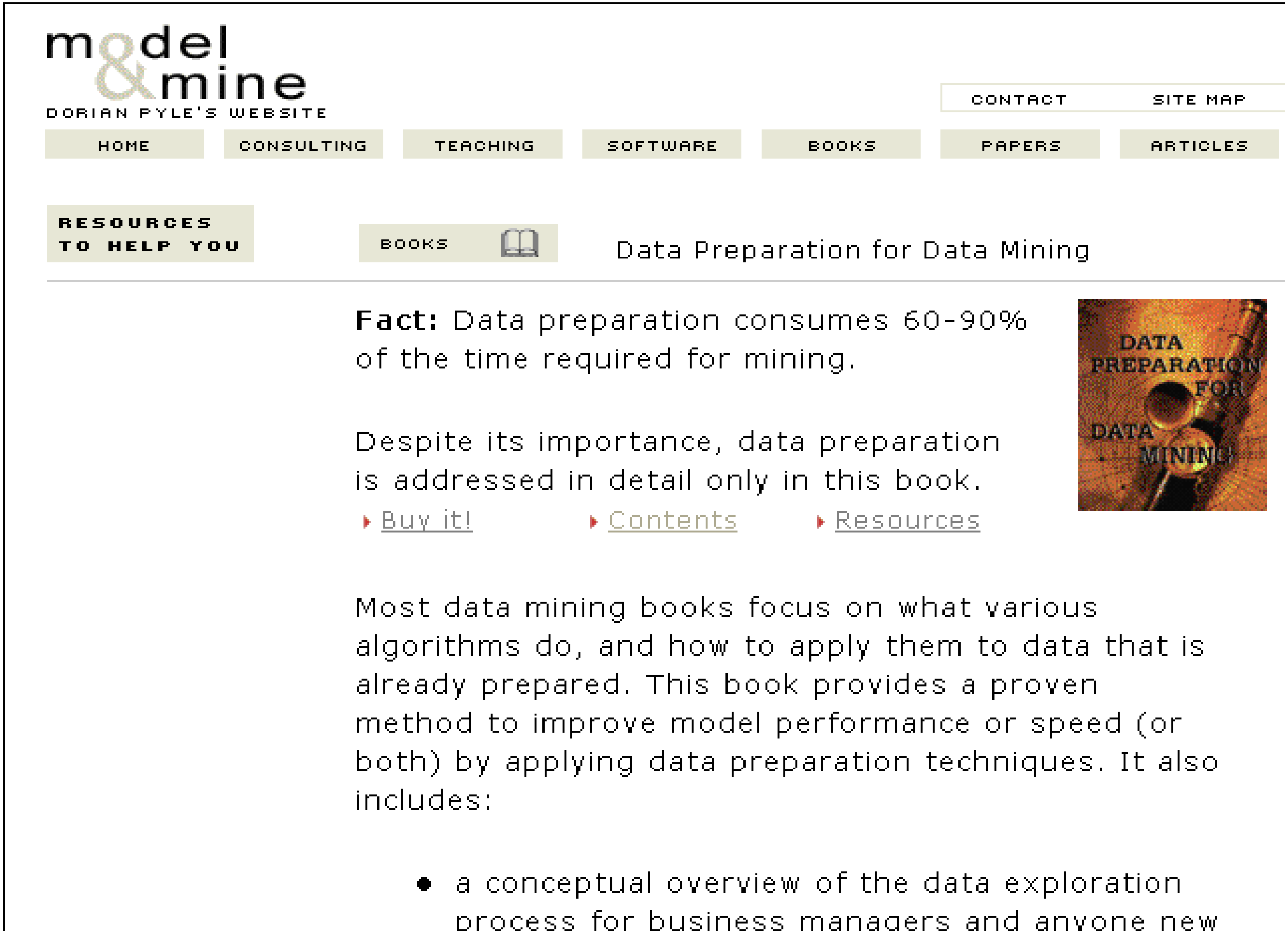
Clusterization level
1

Objects disposition
Circle

☑ Color Scale

Review and correct the Data

Importance of Preprocessing (it takes 60%-90% of efforts)



Data Preparation for Data Mining

Fact: Data preparation consumes 60-90% of the time required for mining.

Despite its importance, data preparation is addressed in detail only in this book.

Buy it! Contents Resources

Most data mining books focus on what various algorithms do, and how to apply them to data that is already prepared. This book provides a proven method to improve model performance or speed (or both) by applying data preparation techniques. It also includes:

- a conceptual overview of the data exploration process for business managers and anyone new

CONTENTS

Introduction

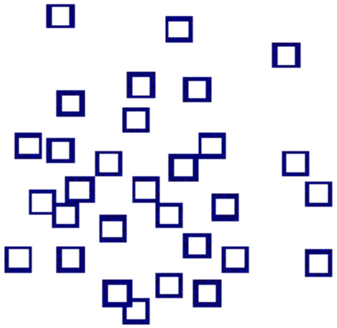
Definitions

Clustering ←

Discussion

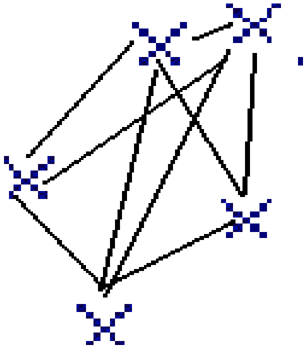
Open Problems

Definitions



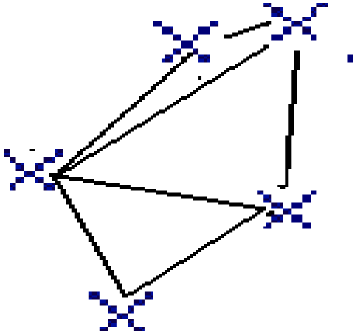
Set

Def. 1 “Let us V be the set of objects. Clustering $C = \{ Ci \mid Ci \in V \}$ of V is division of V on subsets, for which we have : $\bigcup_i Ci = V$ and $Ci \cap Cj = 0 \quad i \neq j$“



Clique

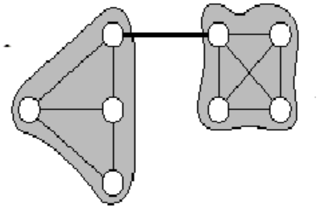
Def. 2 “Let us V be the set of nodes, E be arcs, φ is weight function that reflects the distance between objects, so we have weighted graph **$G = \{ V,E, \varphi \}$.** In this case C is named as clustering of G.”



Graph

In the framework of the second definition every Ci produced subgraph G(Ci). Both subsets **Ci** and subgraphs **G(Ci)** are named **clusters.**

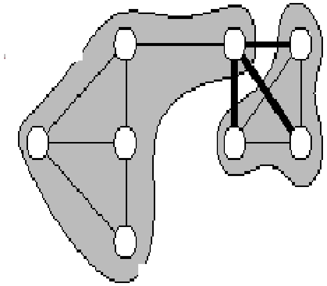
Definitions



Principal note

Both definitions **SAYS NOTHING**:

- about quality of clusters
- about **numbers of** clusters

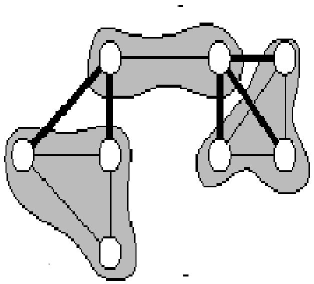


Reason of difficulties

Nowadays **there is no any general agreement** about any universal defintion of the term **‘cluster’**

What means that clustering is good ?

1. Closeness between objets **inside clusters** is essentially more than the closeness **between clusters** themselves
2. Constructed clusters correspond to **intuitive presentations** of users (they are **natural** clusters)



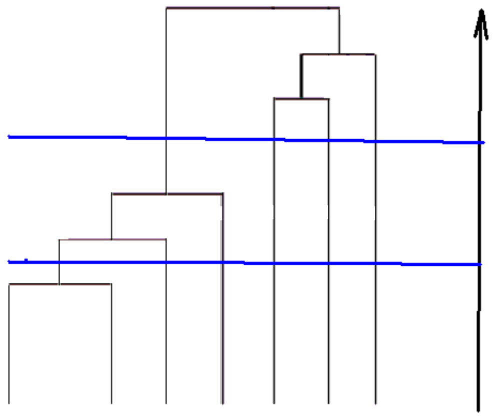
Classification of methods

Based on the way of grouping

1. Hierarchy based methods

Any neighbors

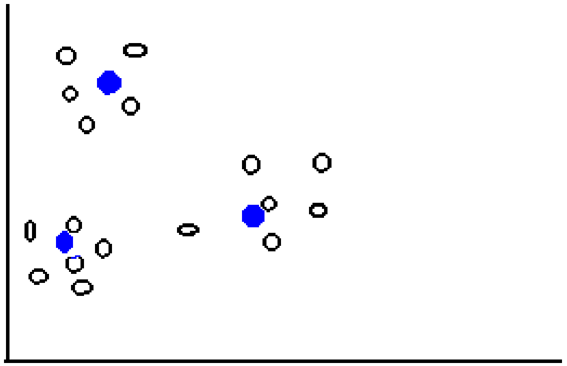
N =? N is not given



2. Exemplar based methods

K-means

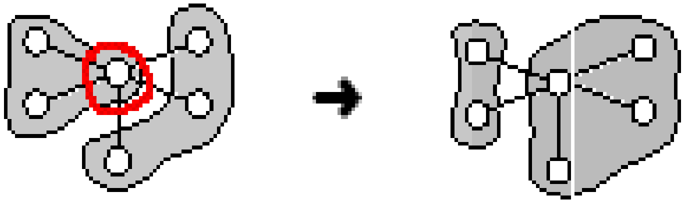
N = ? N is given



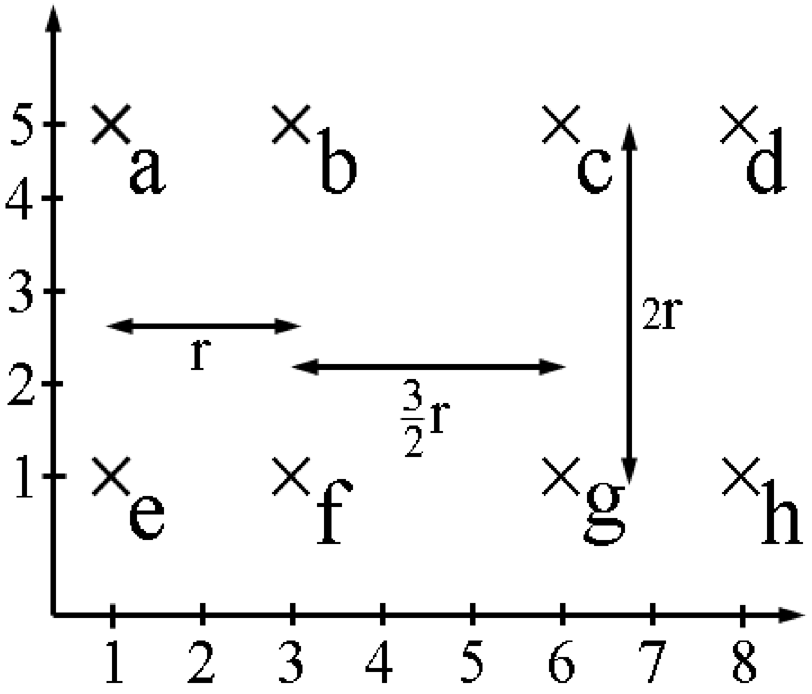
3. Density based methods

MajorClust

N = ? N is calculated automatically



Hierarchy based methods



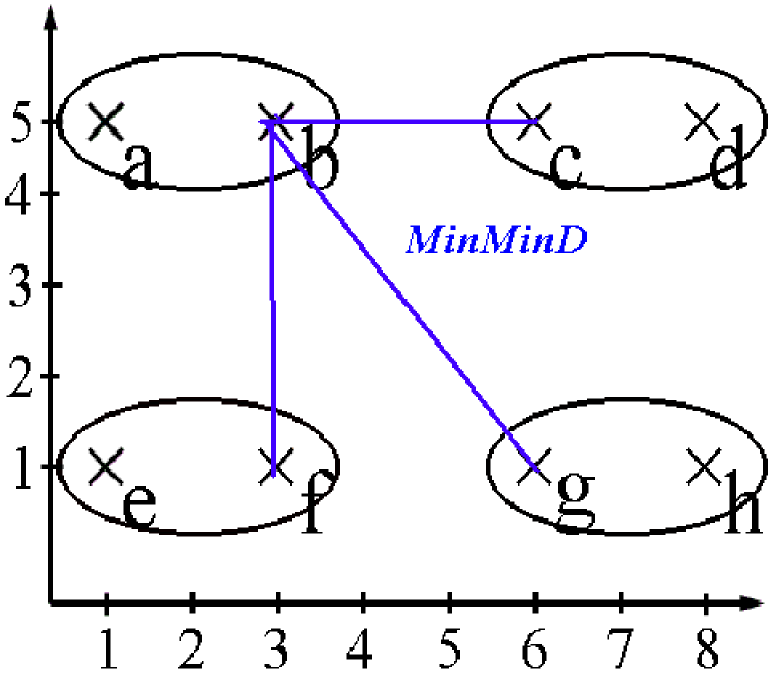
General algorithm

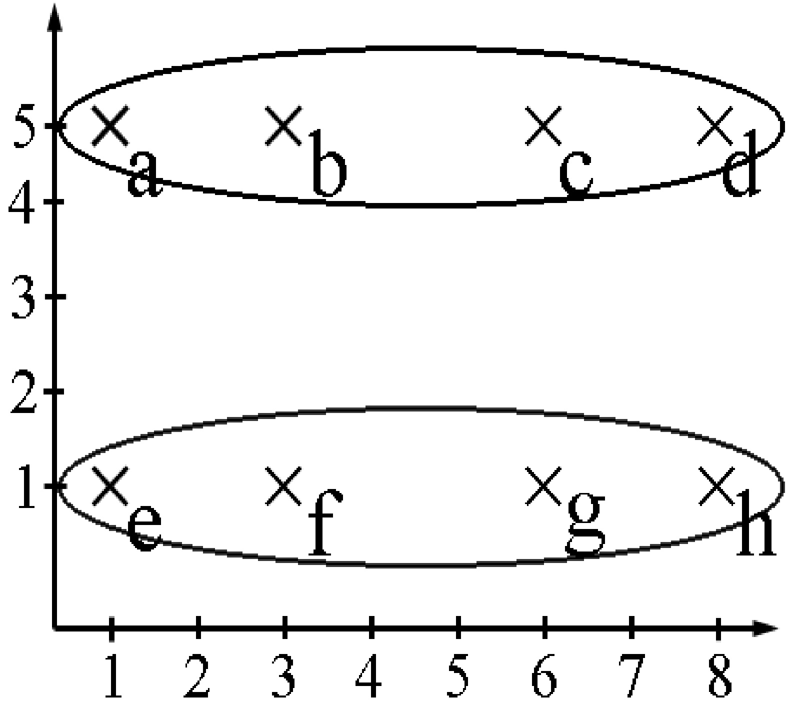
3. *Initially every object is one cluster*
5. *The series of steps are performed. On every step the pair of cluster being the closest ones are merged.*
6. *At the end we have one cluster.*

Neighbors.

Every object is cluster

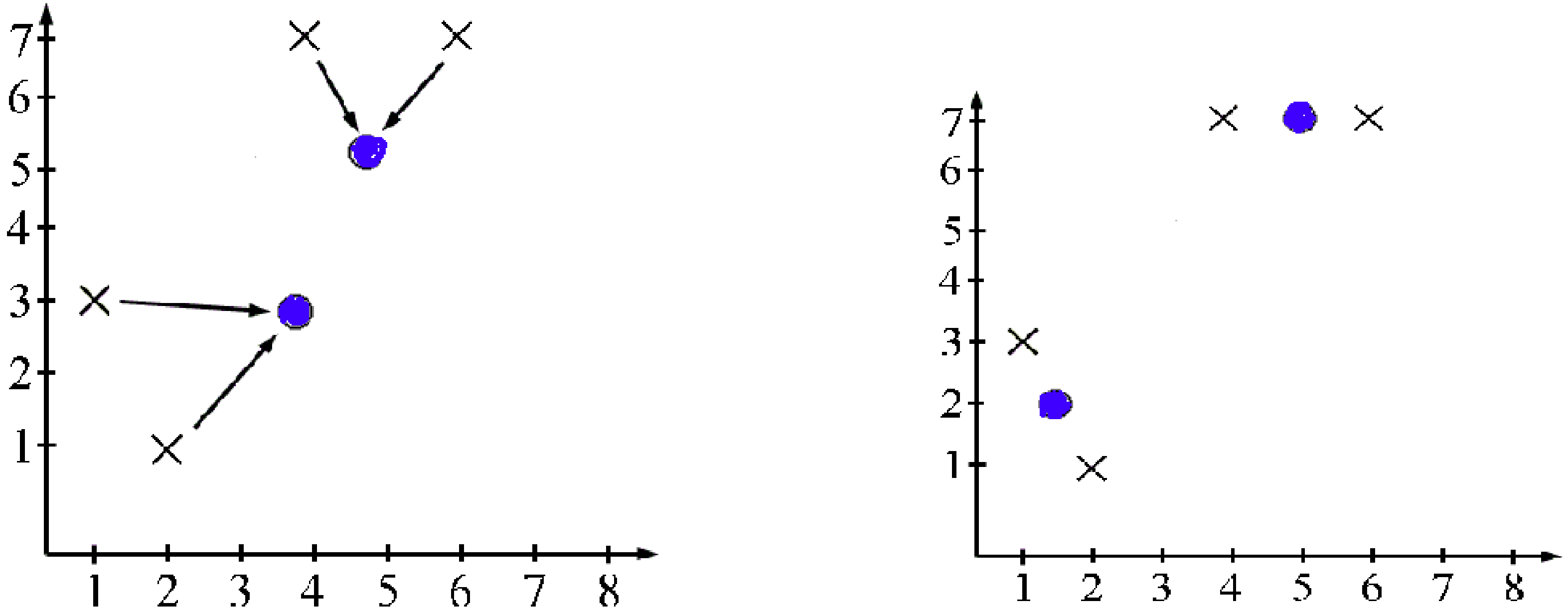
Hierarchy based methods





Nearest neighbor method (NN)

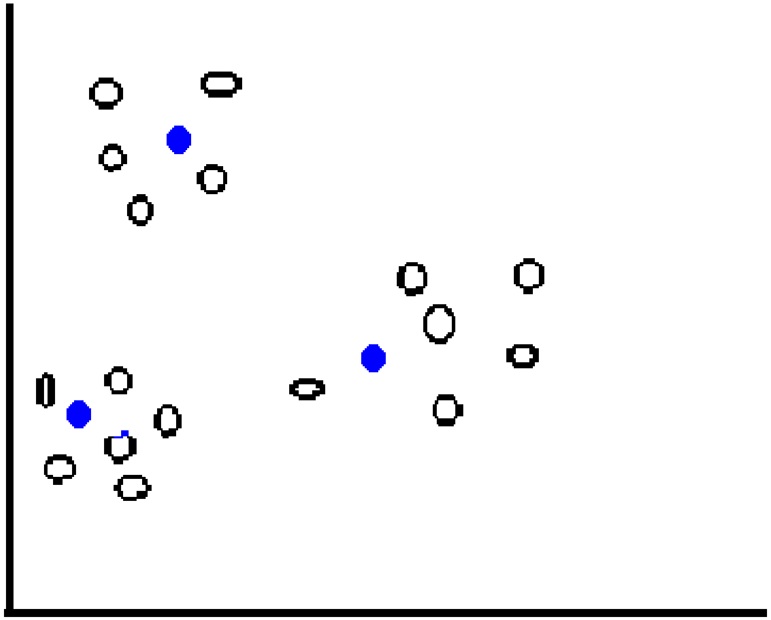
Exemplar based methods



K - means, centroid

Exemplar based methods

Method K-means



General algorithm

3. *Initially K centers are selected by any random way*

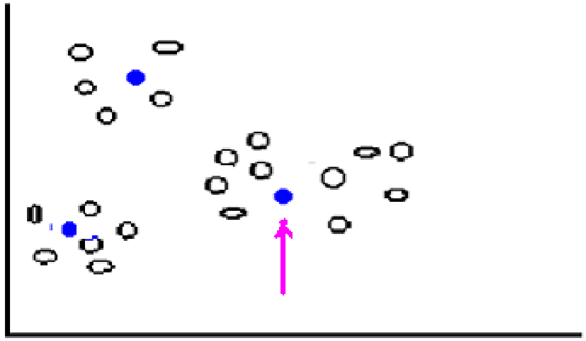
5. *Series of steps are performed. On every step the objects are distributed between centers according the criterion of the nearest center. Then all centers are recalculated.*

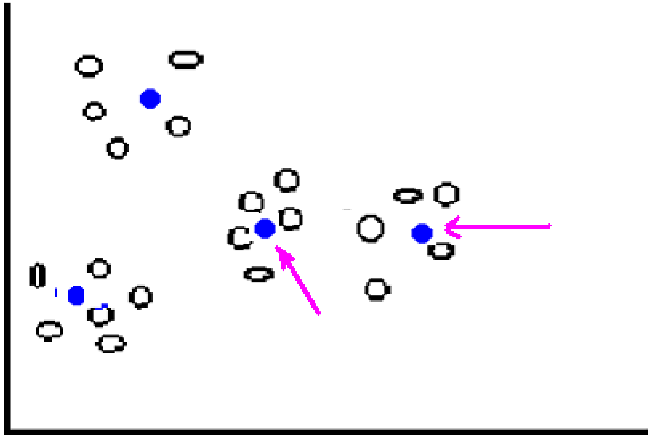
7. *The end is fixed when the centers are not changed.*

Exemplar based methods

Method X-means

(Dan Pelleg, Andrew Moor)





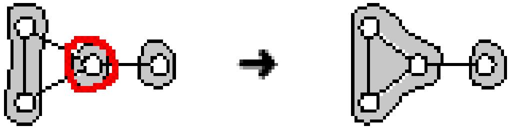
Approach

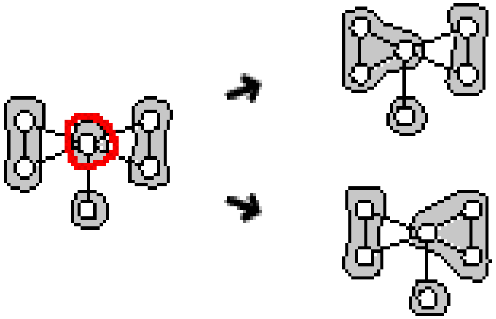
Using evaluation of object distribution
Selection of the most **likely points**

Advantage

- More rapid
- Number of cluster **is not fixed**
  (in all cases it tends to be less)

Density based methods

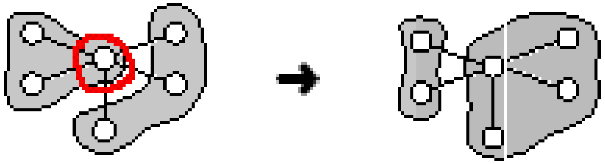




MajorClust method

Principal idea

Total closeness to the objects of his own cluster exceeds the closeness to any other cluster

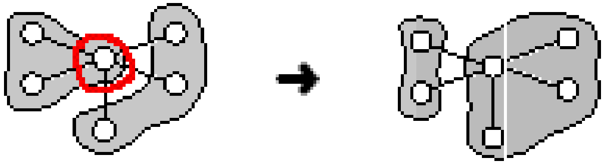


Suboptimal solution

Only part of neighbors are considered on every step (to save time, to avoid mergence)

Density based methods

MajorClust method

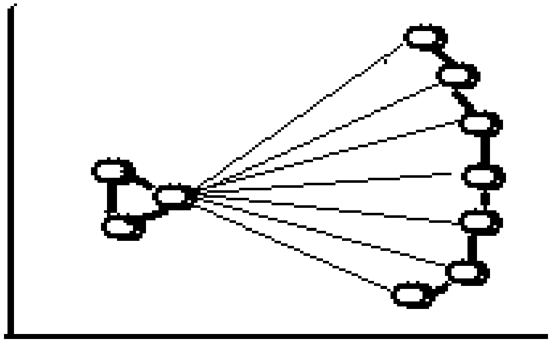


General algorithm

5. *Initially every object is one cluster and it joins to the nearest neighbor*

7. *Every object evaluates the total closeness to his own cluster and separately to all other clusters. After such evaluation the objects change its belonging and go off to the closest one*

9. *The end of searching is fixed when clusters do not change.*



Preprocessing for MajorClust

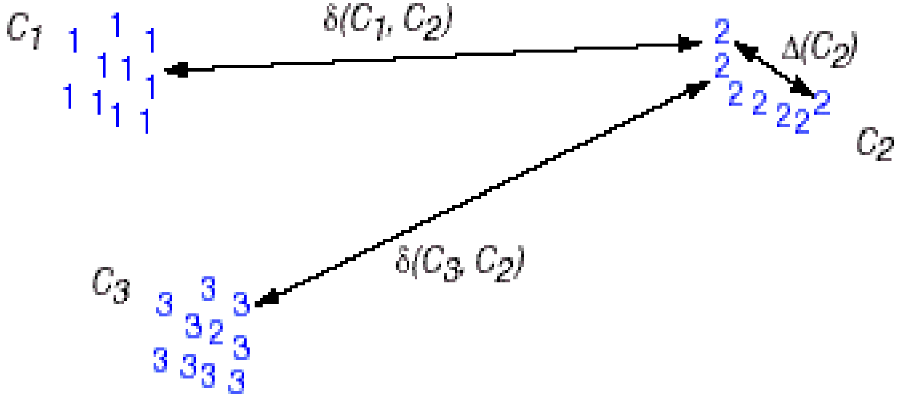
Many weak links can be stronger than the several strongest ones that disfigures results.
So: weak links should be **eliminated** before clustering

Cluster Validity

Definition

It reflects cluster **separability** and formally depends on :

- Scatters inside clusters
- Separation between clusters



Indexes

It is formal characteristics of structure

- **Dunn** index
- **Davies Bouldin** index
- Hypervolume criterion (**Andre Hardy**)
- Density expected measure DEM (**Benno Stein**)

Dunn index (to be **max**)

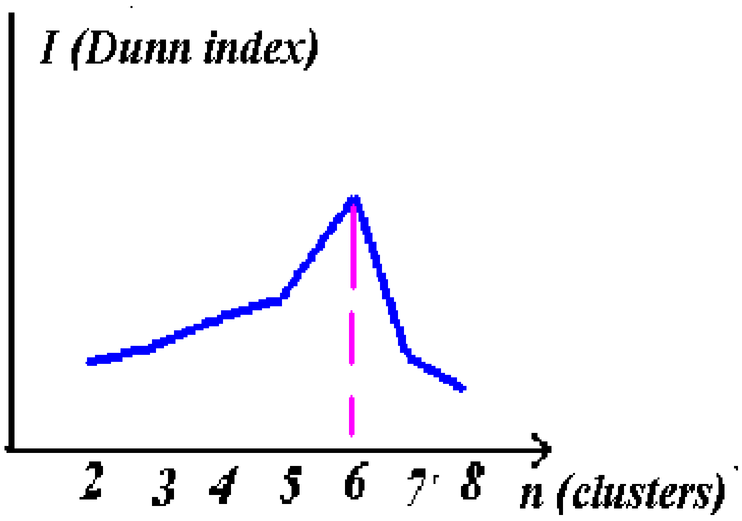
$$I(C) = \frac{\min_{i \neq j}\{\delta(C_i, C_j)\}}{\max_{1 \leq l \leq k}\{\Delta(C_l)\}}$$

Cluster Validity

Number of clusters

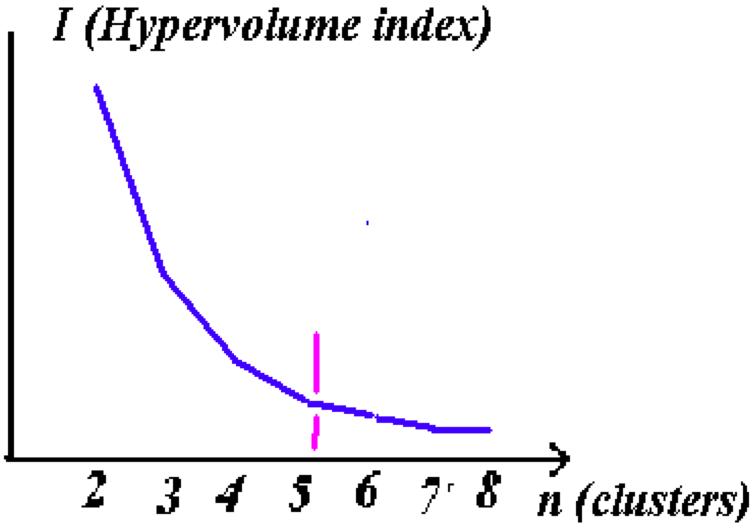
Geometrical approach, two variants:

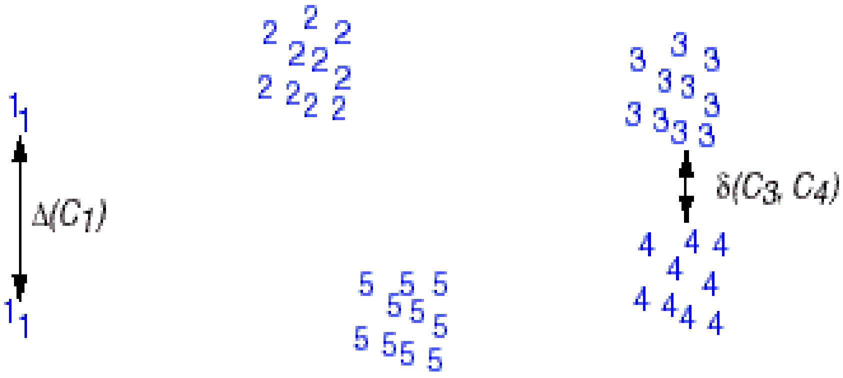
- **Optimum** (min, max) of curve
- **Jump** of curve



Dunn index (to be **max**) is too sensible to extremal cases

$$I(C) = \frac{\min_{i \neq j}\{\delta(C_i, C_j)\}}{\max_{1 \leq l \leq k}\{\Delta(C_l)\}}$$



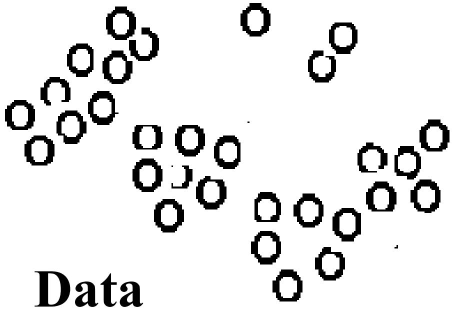


Cluster Usability

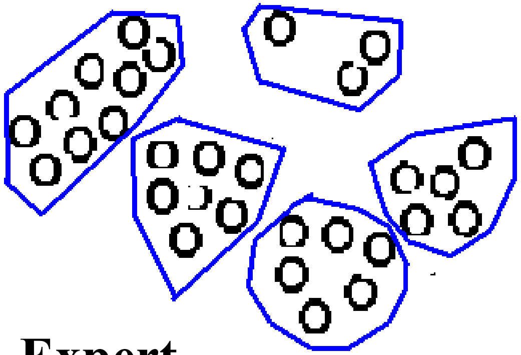
Definition

It reflects user's opinion and formally expresses the difference between :

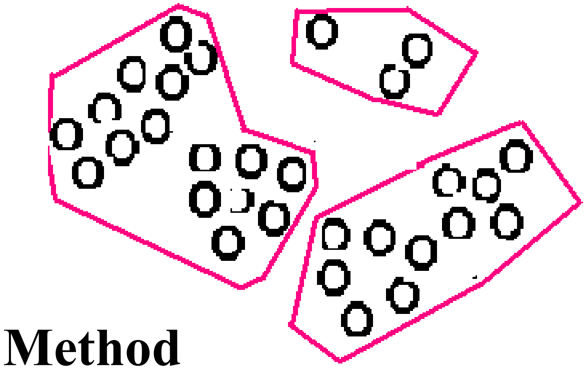
- Classes selected manually by a user
- Clusters constructed by a given method



Data



Expert



Method

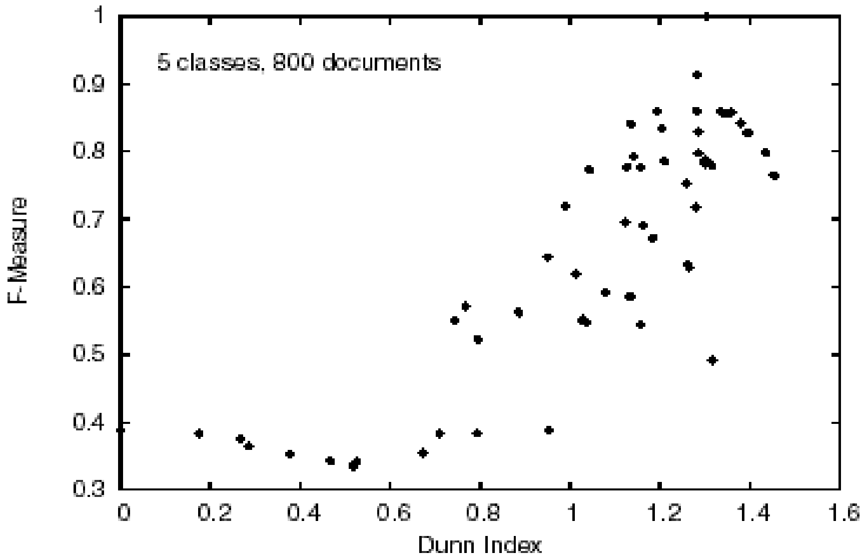
Cluster *F*-measure (*Benno Stein*)

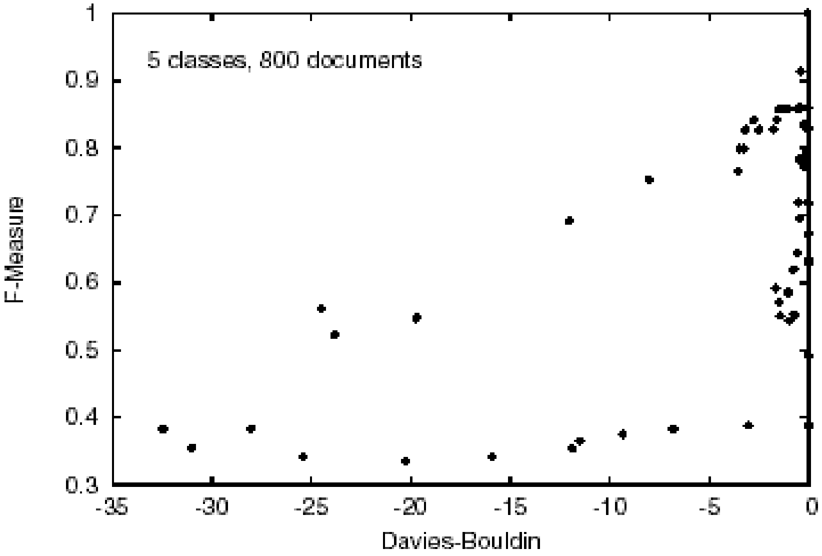
$$F_{i,j} = \frac{2}{\frac{1}{prec(i,j)} + \frac{1}{rec(i,j)}}$$

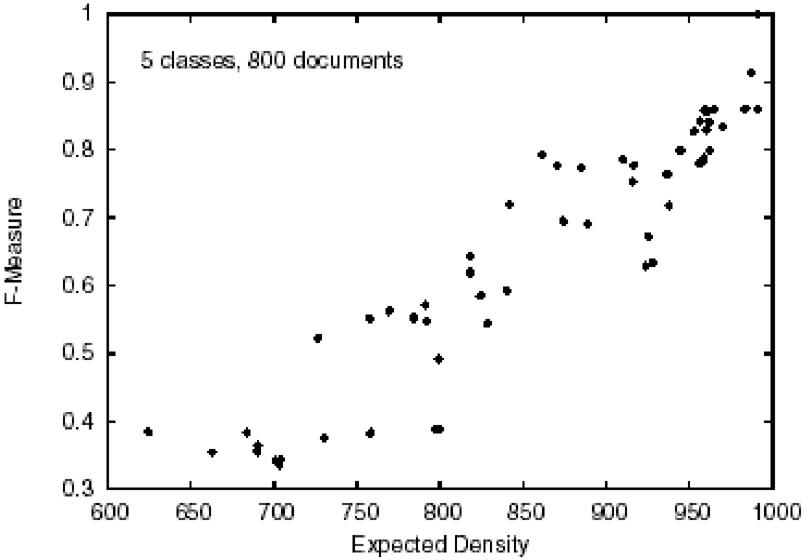
$$F = \sum_{i=1}^{l} \frac{|C_i^*|}{|D|} \cdot \max_{j=1,\dots,k} \{F_{i,j}\}$$

Here: *i, j* are indexes of clusses and clusters
C^*_i, C_j are classes and clusters
prec(i,j), rec(i,j) are precision and recall

Validity and Usability







Conclusion

Density expected measure corresponds to ***F*-measure** reflecting expert's opinion.

So, **DEM** can be an indicator of expert **opinion**

Tecnologies of Clustering

Meta methods

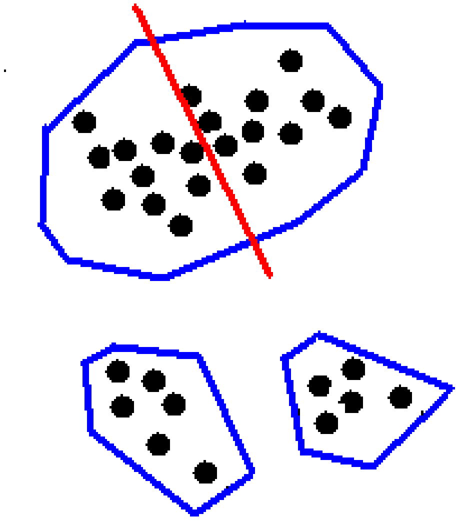
They construct separated data sets using criteria of optimization and **limitations**:
- Neither much nor small **number** of clusters
- Neither large nor small **size** of clusters
etc.

Visual methods

They present visual images to a user in order to select **manually** the clusters
- Using **different** methods
- **Comparing** results

Meta Methods

Algorithm (example)



Notations:

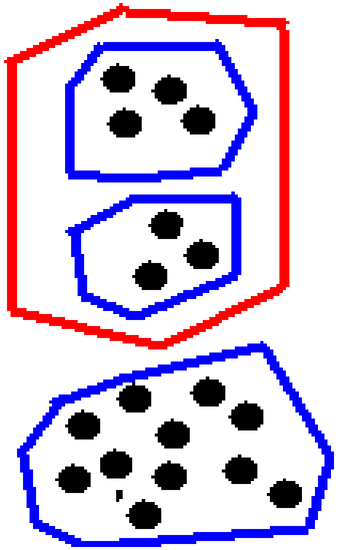
N is the number of objects in a given cluster

D is the diagonal of a given cluster

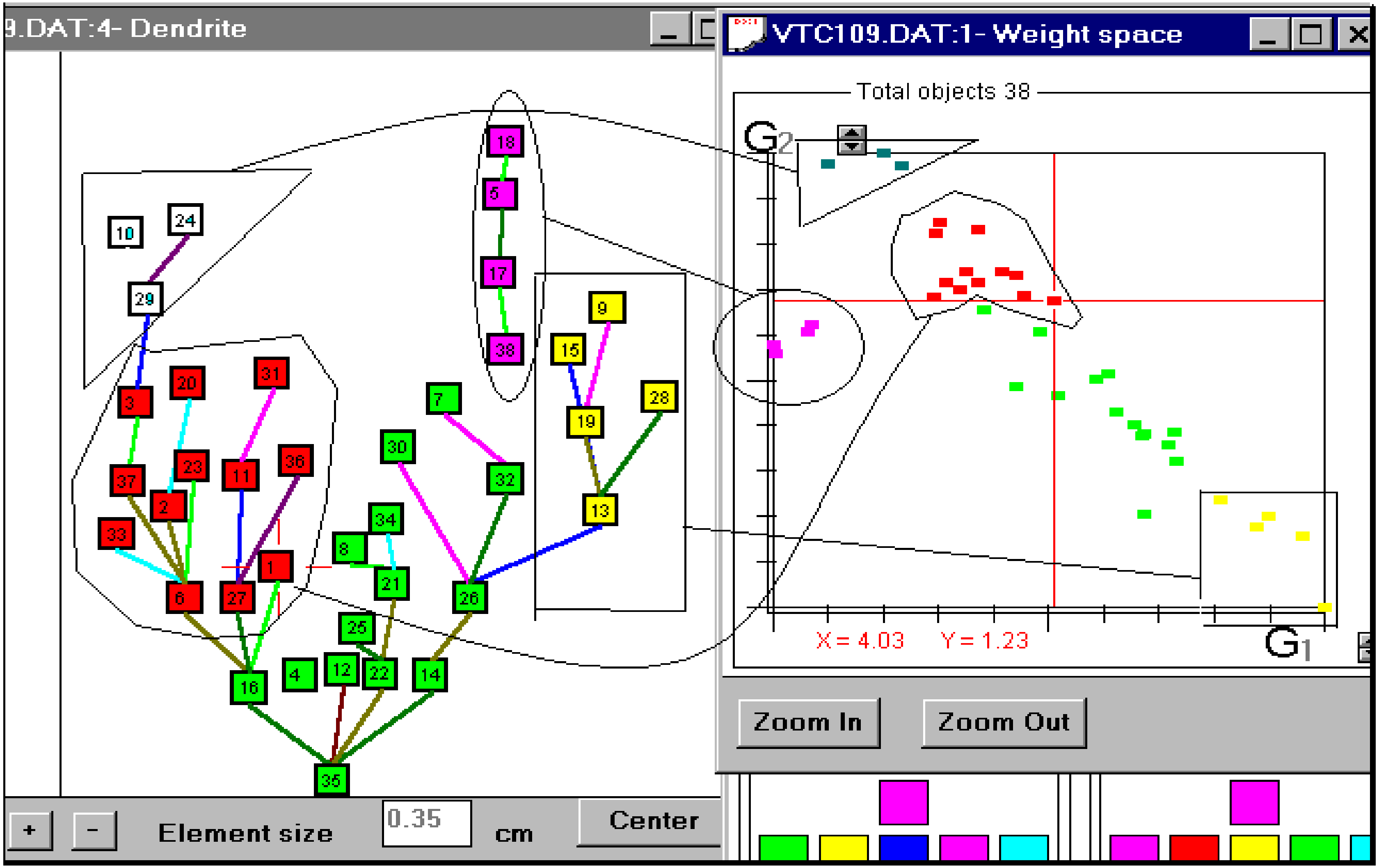
Initially N_0 and their centers Ci are given

Steps

1. Method K-medoid (or any other one) is performed
2. If $N > N\mathrm{max}$ or $D > D\mathrm{max}$ (in any cluster), then this cluster is divided on 2 parts. Go to p.1
3. If $N < N\mathrm{min}$ or $D < D\mathrm{min}$ (in any cluster), then this and the closest clusters are joined. Go to p.1
4. When the number of iteration $I > I\mathrm{max}$, Stop Otherwise go to p.1



Visual Clustering



9.DAT:4- Dendrite

Element size 0.35 cm

VTC109.DAT:1- Weight space

Total objects 38

X = 4.03 Y = 1.23

Clustering on dendrite **Clustering in space of factors**

Authorship

Problem

Authorship of Molier dramatic works (comedies, dramas,...).
Corneille and/or Molier ?

Approach

Style based indexing (**NooJ** can be used)
Clustering all dramatic works
Well-known dramatic works should be marked

Style

- Formal style estimations
- Informal style estimations

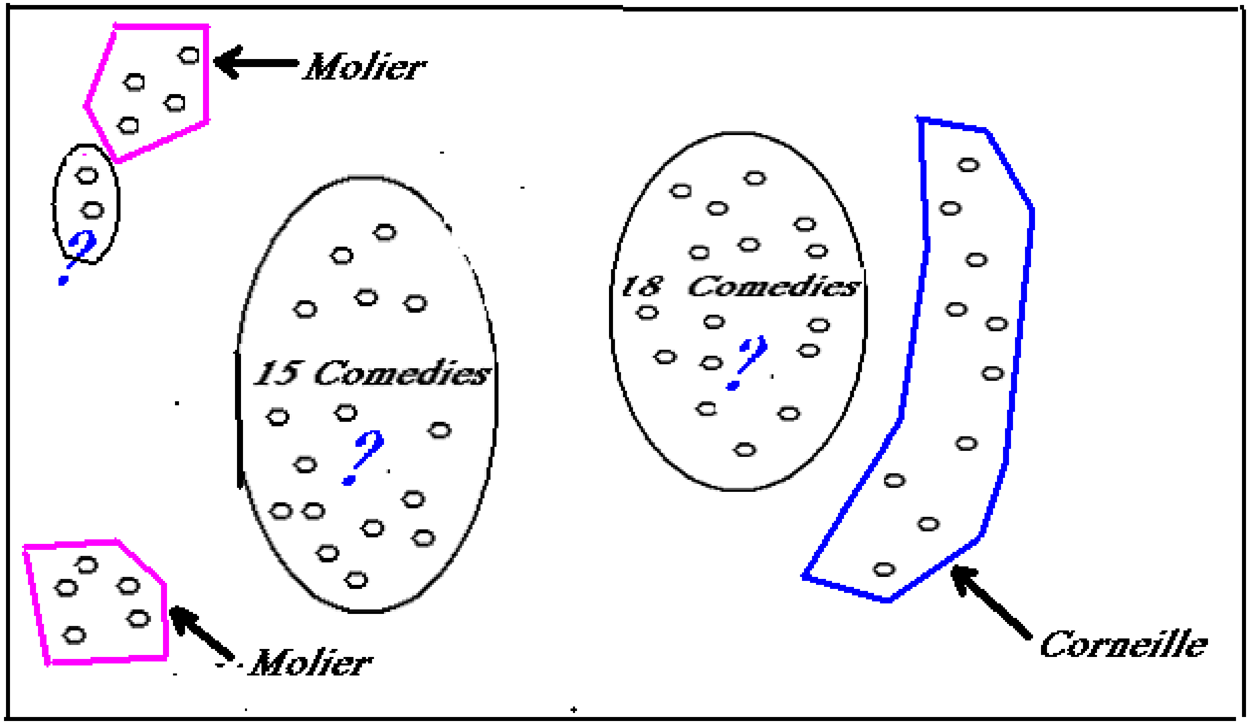
Formal style indicators

- Text Complexity
- Text Harmonicity

References:
Labbe C., Labbe D.
Inter-textual distance and authorship attribution Corneille and Molier.
Journ. of Quantitative Linguistics.
2001. Vol.8, N_3, pp.213-331

Authorship

Clustering



Note:
During a certain time Molier and Corneille were friends

Results

1) 18 comedies of Molier should be belonged to Corneille

2) 15 comedies of Mollier are weak connected with all his other works.
So, they can be written by two authors

3) 2 comedies of Corneille now are considered as works of Molier.
etc.

Learning

Journals and Congresses about Clustering

1. **Journal “Journal of Classification”,** Springer
2. **IFCS** **-** International Federation of Classification Societies, Conferences
3. **CSNA** - Classification Society of North America, Seminars, Workshops

Special and Universal packages with algorithms of Clustering

1. **ClustAn (**Scotland**)** www. clustan.com ***Clustan Graphics-7 (2006)***
2. **MatLab** Descriptions are in Internet
3. **Statistica** Descriptions are in Internet

CONTENTS

Introduction

Definitions

Clustering

Discussion

Open Problems

←

Certain Observations

The numbers of methods for grouping data is a little bit more than the numbers of researchers working in this area.

Problem does not consist in searching the **best method** for all cases.

Problem consists in searching the **method being relevant** for your data.

Only you know what methods are the best for you own data**.**

Principal problems consist in choice of indexes (parameters) and measure of closeness to be adecuate to a given problem and given data

Frecuently the results are bad because of the **bad indexes** and **bad measure** but not the **bad method** !

Certain Observations

Antipodal methods

To be sure that results are really good and **do not depend on the method** used one should test these results using any **antipodal** methods

Solomon G, 1977: “The most antipodes are: **NN-method** and **K-means**”

Sensibility

To be sure that results **do not depend** essentially on the **method’s parameters** one should perform the analysis of sensibility by changing parameters of adjustment.

CONTENTS

Introduction

Definitions

Clustering

Conclusions

Open Problems

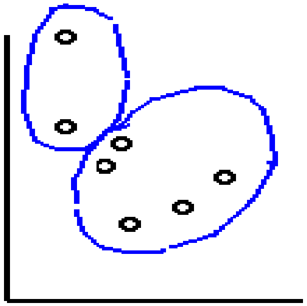
Some Problems

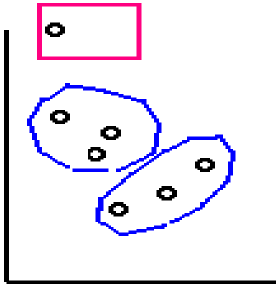
Question 1

How to reveal **alien** objects?

Solution (idea)

Revealing **a stable** structure on different sets of objects. They are subsets of a given set.





Object distribution reflects:
real structure (**nature**)
+
noise (**alien objects**)

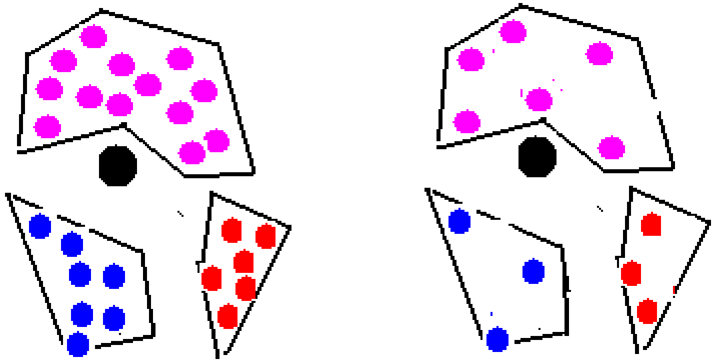
Some Problems

Question 2

How to accelerate classification?

Solution (idea)

Filtering objects, which give a minimum contribution to decisive function



Representative objects of each cluster

CONTACT INFORMATION

Mikhail Alexandrov[1,2], Pavel Makagonov[3]

[1] Autonomous University of Barcelona, Spain
[2] Social Network Research Center with UCE, Slovakia
[3] Mixtec Technological University, Mexico
dyner1950@mail.ru, mpp@mixteco.utm.mx

Petersburg 2008