DeepL
Subscribe to DeepL Pro to translate larger documents.
Visit www.DeepL.com/pro for more information.

Clustering

Advanced Programming Methods Case Study (Course A)

AA 2022-2023

Data Mining

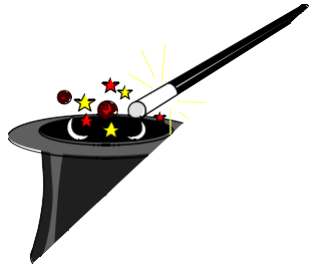
The purpose of data mining is the (semi-)automatic *extraction* of *knowledge* hidden in voluminous databases in order to make it available and directly usable

Clustering

Data

:

- a collection D of transactions where, each transaction is a vector of attribute-value (item) pairs;
- a whole k;



The aim is:

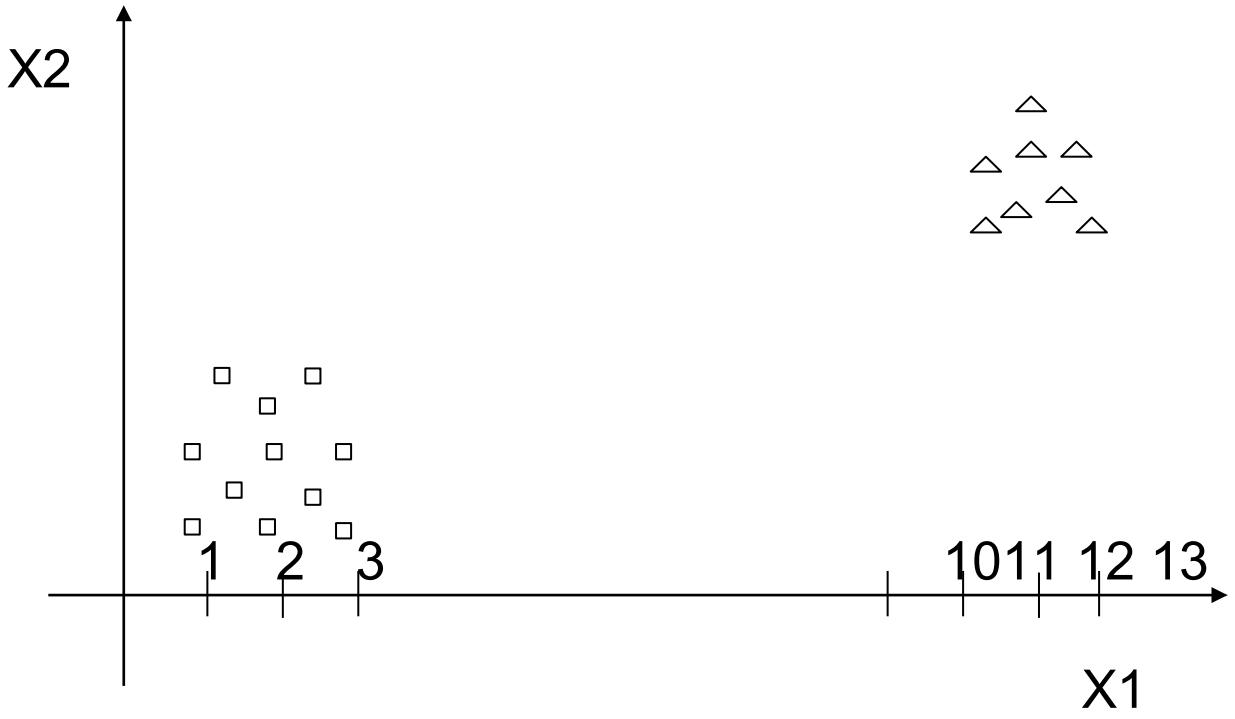
- Partition D into k sets of transactions D_1 ,...D_k, such that:
  - D_i (i=1,...,k) is a homogeneous segment (selection) of D;

k

- $D = \bigcup_{i=1} D_i$ and $D_i \cap D_j = \Phi$.

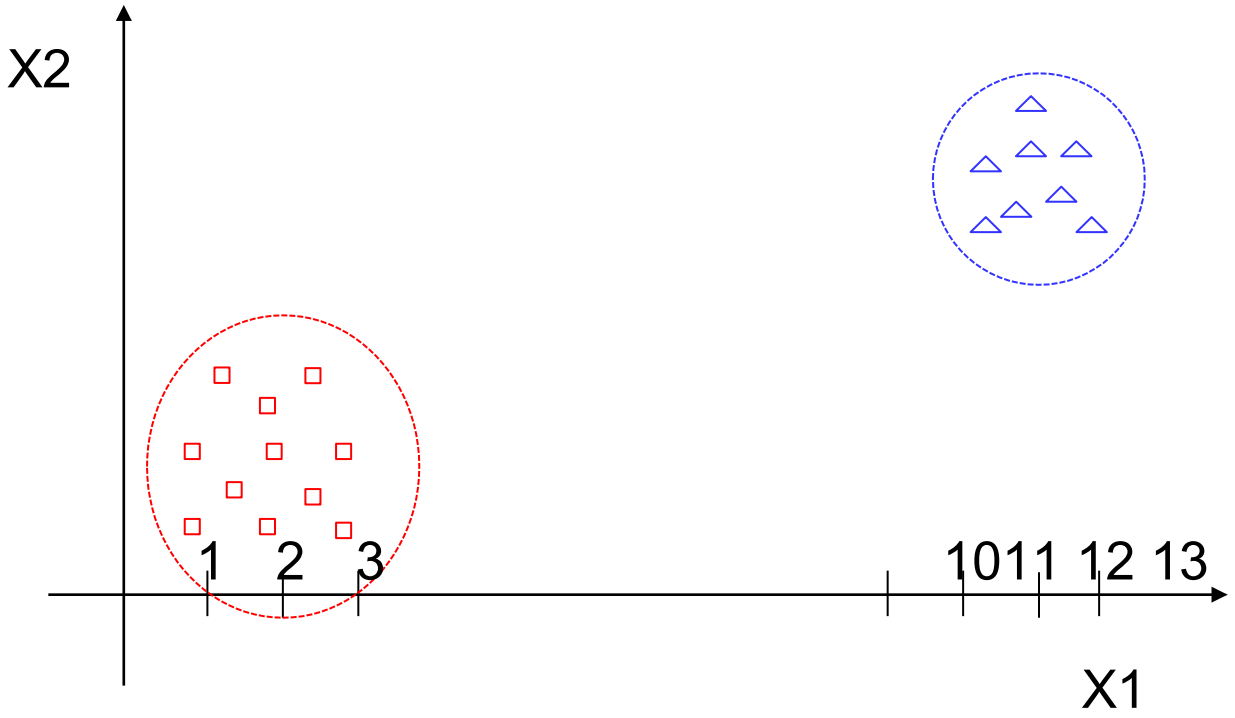
Clustering

X 1	X 2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3



Clustering

X 1	X 2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

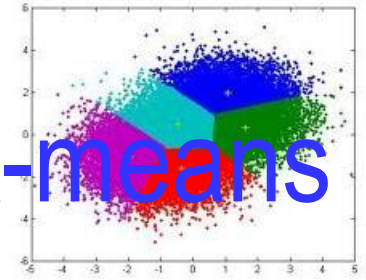


Problems

1. How do I perform clustering?
   - *K -means.*
2. How do I represent clusters?
   - *Calculating and storing cluster centroids.*
3. How do I use clusters in real applications?
   - *Minimising the distance between a new transaction and the cluster representation (centroids) for*

     discover the cluster they belong to.

K-means



http://it.wikipedia.org/wiki/K-means_clustering

Kmeans(D,k)-:clusterSet

clusterSet: set of k segments $_{Di}$: each segment $_{Di}$ is a set of transactions in *D*

begin

1. initialise *clusterSet* with initially empty segments
2. assigns each *clusteSet* segment a transaction randomly chosen by *D*
3. **do**
   for(*transaction*:*D*)
   3.1 $_{Di=cluster}$(clusterSet,transaction)
   3.2 *transaction* in segment $_{By}$
   3.4 Cluster Seeds as the centroids of the clusters
   while (at least one transaction changes cluster)
4. **return** *clusterSet*;

end

X 1	X 2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: how?

STEP 1: initialisation of k seeds (empty sets)

clusterSet={D_1 ,D }$_2$

D1 = {}

D2 = {}

X 1	X 2
0.9	1
0.9	**1.2**
1.3	2
1.2	3.7
1.9	1
2	**2.2**
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: how?

STEP 2: Initialisation of centroids

One chooses k transactions (centroids) CASUALLY and inserts them into the segments: one centroid per segment.

clusterSet=$\{D_1, D\}_2$

c1= (0.9 , 1.2) : D1 = D1 $\cup$ c1

c2=(2,2.2) : D2 = D1 $\cup$ c2

X 1	X 2
0.9	1
0.9	**1.2**
1.3	2
1.2	3.7
1.9	1
2	**2.2**
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: how?

STEP 3: I assign each transaction to *its* cluster

Whether a transaction belongs to a cluster depends on the distance of the transaction from the cluster centroid.

One chooses to move the transaction to the cluster that minimises this distance.

X 1	X 2
0.9	1
0.9	**1.2**
1.3	2
1.2	3.7
1.9	1
2	**2.2**
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: how?

STEP 3: I assign each transaction to *its* cluster

Whether a transaction belongs to a cluster depends on the distance of the transaction from the cluster centroid.

One chooses to move the transaction to the cluster that minimises this distance.

EuclideanDist((0.9,1),(0.9,1.2))=0.2

EuclideanDist((0.9,1),(2, 2.2))=1.62

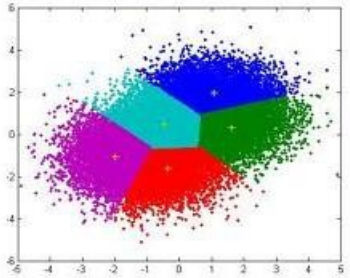
X 1	X 2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: how?

STEP 3: I assign each transaction to *its* cluster

Whether a transaction belongs to a cluster depends on the distance of the transaction from the cluster centroid.

One chooses to move the transaction to the cluster that minimises this distance.



clusterSet={D_1 ,D }$_2$

D_1 ={1,2,5,8}

D_2 ={3,4,6,7,9,10,11,12,1314,15,16,17}

X 1	X 2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7

kMeans: how?

STEP 4: Recalculate cluster centroids

The centroid is a fictitious segment transaction that associates each attribute with the mean value (fashion) calculated on the segment

clusterSet={D_1 ,D }$_2$

c1= (1.65, 1.05) where:

$$\frac{0.9 + 0.9 + 1.9 + 2.9}{4} = 1.65$$

12.2	5.9
12.5	6.2
13	5.3

$$\frac{1+1.2+1+}{\frac{1}{4}} = 1.05$$

X 1	X 2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: how?

STEP 4: Recalculate cluster centroids

The centroid is a fictitious segment transaction that associates each attribute with the mean value (fashion) calculated on the segment

clusterSet={D_1 ,D }$_2$

c1= (1.65, 1.05)

c2=(8.03, 4.66)

X 1	X 2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: how?

STEP 5: Have any transactions changed clusters?

repeat STEP 3 with

c1= (1.65, 1.05)

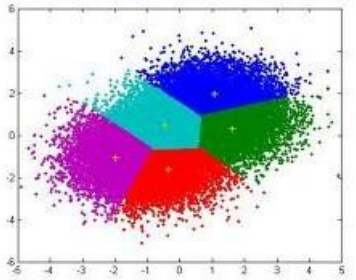
c2=(8.03, 4.66)

X 1	X 2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: how?

STEP 3: I assign each transaction to the nearest cluster

clusterSet={D_1 ,D $\}_2$



D_1 = {1,2,3,4,5,6,7,8,9}

D_2 = {10,11,12,13,14,15,16,17}

X 1	X 2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: how?

STEP 4: Calculate the centroids of new clusters

clusterSet={D1,D2}

c1= (1.76,1.98)

c2=(11.9, 5.875)

X 1	X 2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: how?

STEP 5: Are there any transactions that have changed the cluster? YES

repeat STEP 3 with:

c1= (1.76,1.98)

c2=(11.9, 5.875)

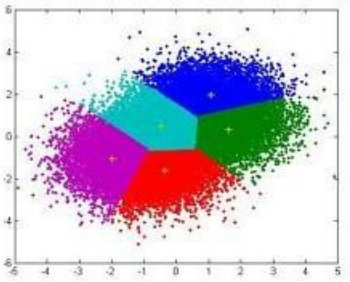
X 1	X 2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

kMeans: how?

STEP 3: I assign each transaction to the nearest cluster.

D_1 = {1,2,3,4,5,6,7,8,9}

D_2 = {10,11,12,13,14,15,16,17}



STEP 4: Calculate the centroids of new clusters

c1= (1.76,1.98) c2=(11.9, 5.875)

STEP 5: There are transactions that have changed the cluster of membership?

2.Representation of a cluster

1) Extensional description (list of transactions in the cluster).

Cluster 1

X1	X2
0.9	1
0.9	1.2
1.3	2
1.2	3.7
1.9	1
2	2.2
1.9	3.1
2.9	1

Cluster 2

X1	X2
2.9	2.7
11	5
11	6
11.5	5.4
12	6.2
12	7
12.2	5.9
12.5	6.2
13	5.3

Representation of a cluster

2) Intensional description (via cluster centroids).

$$x_{centroid} = \begin{cases} \dfrac{\sum_{(\ldots,x_i,\ldots)\in cluster} x_i}{|\,clusters\,|} & \text{if } X \text{ is nume} \\ \square\text{argma} \end{cases}$$

Cluster 1	Cluster 2
(1.76, 1.98)	(11.9, 5.875)

Calculating a centroid: how?

Genre	Nationality	Age
F	Italian	25
F	Italian	27
F	Italian	34
F	English	23
M	Americana	29

centroid

F	Italian	27.6

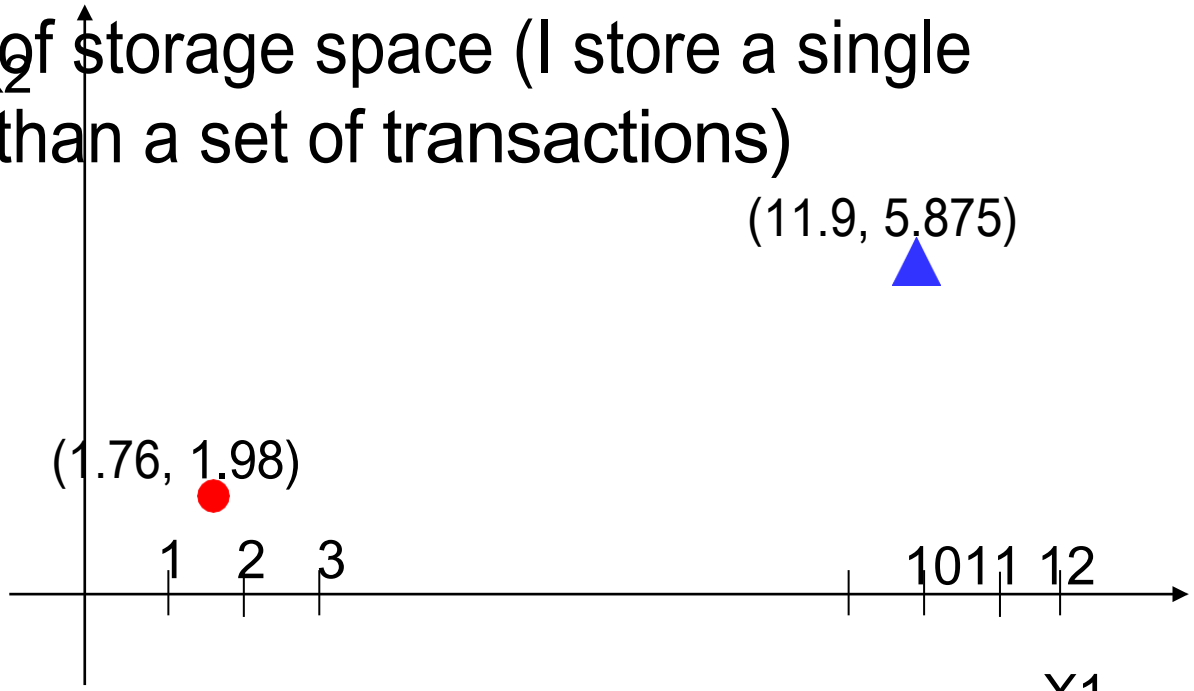
3. Clusters and/or centroids: applications

real

Advantag

es:

1. Compact in terms of storage space (I store a single transaction rather than a set of transactions)



Clusters and/or centroids: applications

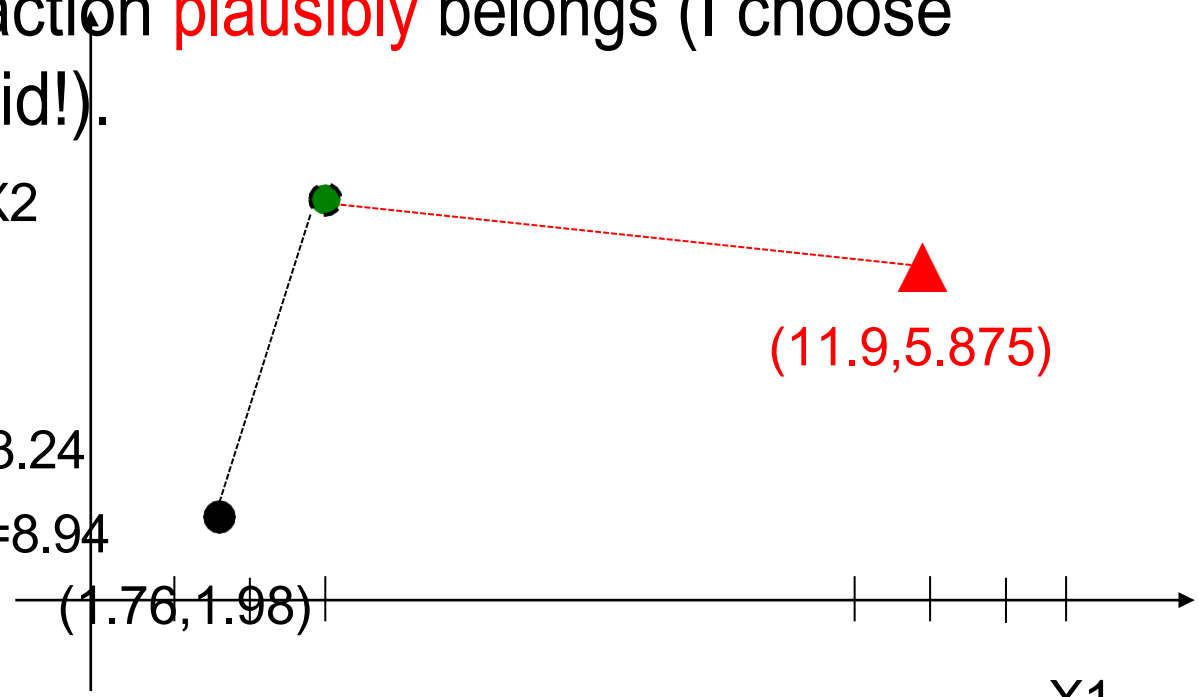
real

Advantag

es:

2. I can use the cluster centroids to find the segment to which a new transaction plausibly belongs (I choose the nearest centroid!).

(3,5) ?



EuclidDist((3,5),(1.76,1.98))=3.24

EuclidDist((3,5),(11.9,5.875))=8.94

1 2 3 1011 12 13

X1

4.Cluster quality: Silhouette

https://en.wikipedia.org/wiki/Silhouette_(clustering)

Validates the consistency of the cluster model with the data

measures how consistent the data is with the cluster to

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases} \qquad -1 \leq s(i) \leq 1$$

Con:

- b(i) the smallest average distance of i from the examples of any other

cluster other than the one in which the

- a(i) the average distance of i from the points grouped in the same clusters of

Cluster quality: how to choose k?

k= 2 avg Silhouette calculation

k=3 avg Silhouette calculation

...

I choose k which represents a local maximum.

Case Study

Designing and implementing a client-server system called 'K-MEANS'.

The server includes data mining capabilities for the discovery of data clusters.

The client is a Java application that enables the remote discovery service and visualises the knowledge (clusters) discovered

Instructions

1. The A.A. 2022 -23 project, called K-MEANS, is only valid for those who pass the written test or in itinere tests within the current A.A.
2. Each project can be carried out by groups of at most THREE (3) students.
3. Those who pass the written test must hand in their project NO LATER than the date set for the corresponding oral test (from the degree programme website). The oral examination will take place on a date following the hand-in (the date will be communicated on esse3 after the hand-in of the project).
4. The discussion of the project will take place upon its delivery, *ad personam* for each member of the group. The maximum mark for the written test is 33. A mark above 30 is equivalent to 30 cum laude.
5. The final grade will be determined on the basis of the grade awarded in the written paper and the project.

Instructions

A project which has not been developed in all its parts (client -server, client interface, db access, serialisation,...) will not be considered sufficient, and as such will not be correct.

Evaluation

Class diagram (2 points) JavaDoc (3 points)
Installation Guide (with Jar+ Bat+ SQL Script) (2 points)
User guide with test examples (2 points) System source (14 points)
Extensions of the project carried out in the laboratory (10 points)