UNIVERSITY OF MOHAMED BOUDIAF – M'SILA
FACULTY OF MATHEMATICS AND INFORMATICS
DEPARTMENT OF COMPUTER SCIENCE
2nd year Master (IDO)

Correction of exam of Information Retrieval & Data Mining **University year: 2020/2021**

Exercise 01: Course question (06pt)

1. Describe in short definition exact match retrieval and best match retrieval? (1pt)
   - Exact match gives results that match a query.
   - Best match gives results that are similar to some degree to the query.

2. Data mining can bring new insight of the data being analyzed… (1pt)
3. CRISP-DM defined and validated a data mining process that could be applicable in any domain application sectors. Shortly what is the benefits of using such process? (1pt)
   - Make data mining project faster, cheaper…
   - …

4. What is the result of using Clustering algorithm? (1pt)
   - A number of cluster (groups) contained data, each of those data on the same cluster share the same similarities.

5. What is the benefit of building decision tree? (1pt)
   - To help decision makers take the good decision in short time.

6. What is the role of association rules algorithm? (1pt)
   - Extract the relation between variable in the database.

Exercise 02: Information Retrieval Models (04pt)

1- Calculation of X :
   - X= $log_2\left(\frac{N}{df}\right) = log_2\left(\frac{806791}{25235}\right)$ =4.99

2- NNC.NTC

NNC weights for documents

We should calculate the normalize vector of each doc

Term	**Doc 01**	**Doc 01 Norm**	**Doc 02**	**Doc 02 Norm**	**Doc 03**	**Doc 03 Norm**
Car	12	12/15.74 = 0.76	25	25/51.72 = 0.48	39	39/45.45 = 0.85
Auto	2	2/15.74 = 0.12	33	33/51.72 = 0.63	1	1/45.45 = 0.02
Insurance	0	0/15.74 = 0	31	31/51.72 = 0.59	20	20/45.45 = 0.44
Best	10	10/15.74 = 0.63	0	0/51.72 = 0	12	12/45.45 = 0.26

NTC weights for query

	Query				***Product***		
Term	***Tf r***	***idf***	***Tf-idf***	***Norm V***	***Doc 01***	***Doc 02***	***Doc 03***
Car	2	1.65	3.3	3.3/3.67 = 0.89	0.76*0.89 =0.67	0.48*0.89 = 0.42	0.85*0.89 =0.75
Auto	0	2.08	0	0	0	0	0
Insurance	1	1.62	1.62	1.62/3.67 = 0.44	0	0.59*0.44 = 0.25	0.44*0.44 =0.19
Best	0	4.99	0	0	0	0	0

- Score (q, Doc 01) = 0.67
- Score (q, Doc 02) = 0.42+0.25 = 0.67
- Score (q, Doc 03) = 0.75+0.19 = 0.94

So the result of this schema are: ***Doc 03***, ***Doc 02***. Or ***Doc 03***, ***Doc 01***

Exercise 03: Decision Tree (04pt)

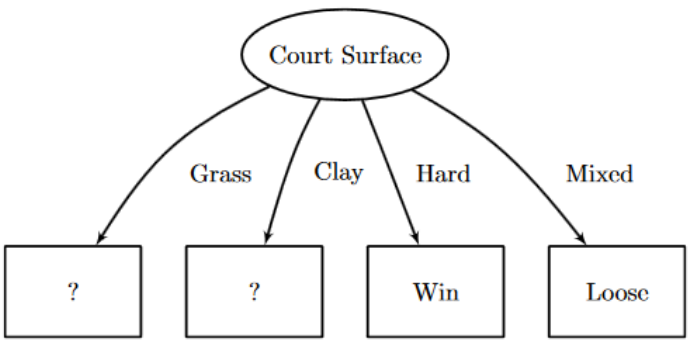
1. We compute the entropy for the whole dataset. 16 entries, 11 *win* and 5 *loose*, it gives an entropy of about 0.99
2. We split the dataset for each attribute, and compute the entropy of each splits. A table shows the details of the calculations

Attribute value	Win	Loose	Entropy
Time			
Morning	2	0	0
Afternoon	7	4	0.94
Night	2	1	0.91
Match Type			
Master	3	3	1
Grand Slam	6	1	0.59
Friendly	2	1	0.91
Court Surface			
Grass	4	1	0.72
Clay	2	2	1
Hard	5	0	0
Mixed	0	2	0
Best effort			
Yes	9	4	0.89
No	2	1	0.91

3. We compute the entropy gain for each split The best entropy gain is

Attribute	Entropy gain
Time	$0.99 - \frac{1}{16}(2 \times 0 + 13 \times 0.94 + 3 \times 0.91) = 0.17$
Match Type	$0.99 - \frac{1}{16}(6 \times 1 + 7 \times 0.59 + 3 \times 0.91) = 0.18$
Court Surface	$0.99 - \frac{1}{16}(5 \times 0.72 + 4 \times 1) = 0.51$
Best Effort	$0.99 - \frac{1}{16}(13 \times 0.89 + 3 \times 0.91) = 0.09$

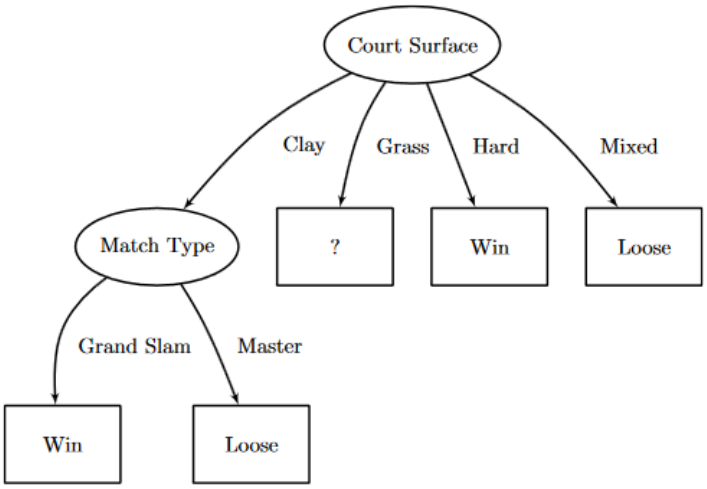
obtained if we use the attribute *court surface*. The tree, at this point, will look like this



4. We now consider the subset of data for which the *court surface* attribute is *clay*. Without even computing the entropy gain for each attribute,

Time	Match type	Best effort	Outcome
Afternoon	Grand Slam	Yes	*Win*
Afternoon	Master	Yes	*Loose*
Afternoon	Master	Yes	*Loose*
Afternoon	Grand Slam	Yes	*Win*

we can see that *time* and *best effort* are the same for all the subset. So they can't be used to explain the *outcome* attribute, giving an entropy gain of 0. But the *match type* is either *grand slam*, outcome is then always *win*, or *master*, outcome is then always *loose*. The entropy gain for the *match type* attribute is 0.99. Thus, we split this subset with the attribute *match type*. The tree, at this point, will look like this



5. We now consider the subset of data for which the *court surface* attribute is *grass*. The entropy of the subset is 0.72. We split the subset for each

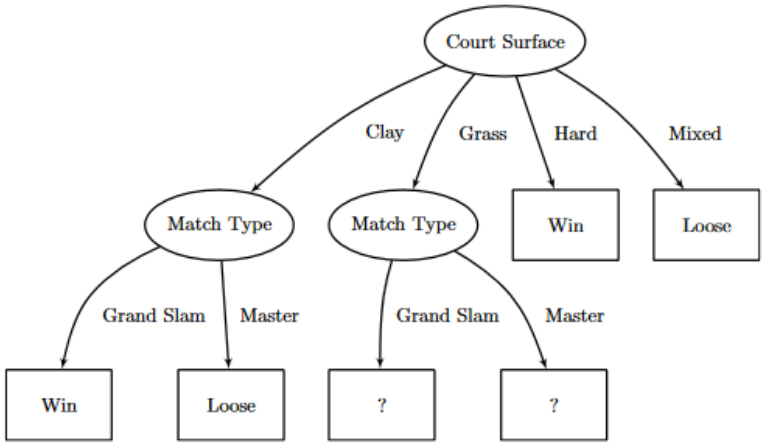
Time	Match type	Best effort	Outcome
Morning	Master	Yes	*Win*
Afternoon	Grand Slam	Yes	*Win*
Afternoon	Master	Yes	*Win*
Morning	Master	Yes	*Win*
Afternoon	Grand Slam	Yes	*Loose*

attribute, and compute the entropy of each splits. A table shows the details of the calculations. We don't need to consider *Best effort*, as it is always equal to *yes*.

Attribute value	Win	Loose	Entropy
Time			
Morning	2	0	0
Afternoon	2	1	0.91
Match Type			
Master	3	0	0
Grand Slam	1	1	1

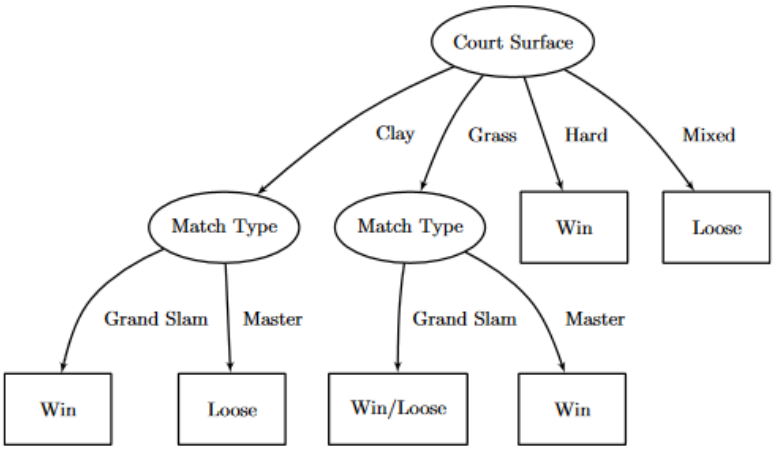
6. We compute the entropy gain for each split The best entropy gain is

Attribute	Entropy gain
Time	$0.72 - \frac{1}{5}(2 \times 0 + 3 \times 0.91) = 0.17$
Match Type	$0.72 - \frac{1}{5}(3 \times 0 + 2 \times 1) = 0.32$



7. For *court surface* = *grass* and *match type* = *master*, the entropy of the subset is 0, we do not need to subdivide it.

8. For *court surface* = *grass* and *match type* = *grand slam*, none of the attribute can help to tell about the outcome, we do not need to subdivide it.



Exercise 04: Association rules (06pt)

A 4 B 4 ~~C 2~~ D 3 ~~E 2~~ ~~K 1~~	A B 4 A D 3 B D 3	A B D 3

Rules:

1- A -> B P(B|A) = |B∩A| / |A| = 4/4, c: 100%
2- B -> A c: 100%
~~3- A -> D c: 75%~~
4- D -> A c: 100%
~~5- B -> D c: 75%~~
6- D -> B c: 100%
~~7- AB -> D c: 75%~~
8- D -> AB c: 100%
9- AD -> B c: 100%
~~10- B -> AD c: 75%~~
11- BD -> A c: 100%
~~12- A -> BD c: 75%~~

Only rules accepted are: 01, 02, 04, 06, 08, 09 and 11.