

School of Information Technology and Electrical Engineering
INFS4203/7203 – Data Mining

Tutorial 1 Data Issues

ANSWERS

Question 1.

Concept Lattice

Given a high-dimensional data set represents the context of service offers of an office supplies business.

Office Supply	Furniture	Computers	Copy machines	PDA's	Specialised machines
Consulting	X	X	X	X	X
Planning	X	X			
Assembly and installation	X	X	X	X	X
Instruction		X	X	X	X
Training workshops		X			
Original spare parts and accessories	X	X	X	X	X
Repairs	X	X	X	X	X
Service contracts		X	X	X	

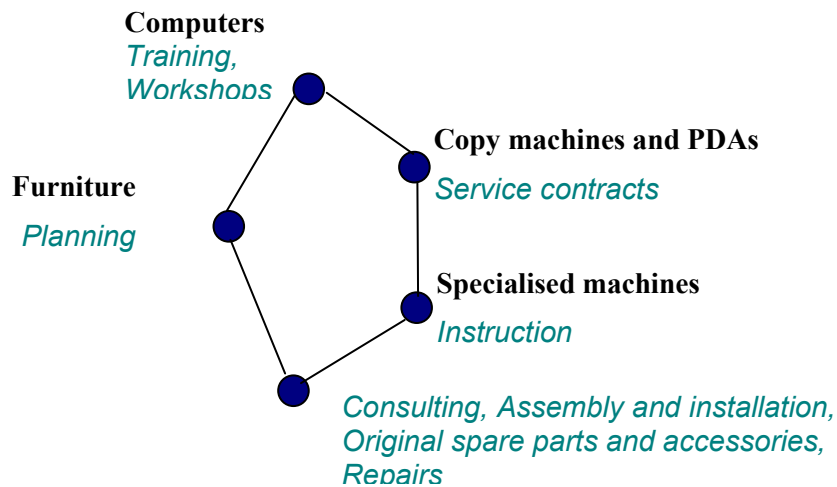
Draw the Concept Lattice for this dataset (refer to Week 1 Lecture Notes).

ANSWER:

Since there are some identical rows and columns in the given table, we need to merge them as the same concept before we draw the Concept Lattice.

So the above table becomes after the merge of the same rows and columns:

Office Supply	Furniture	Computers	Copy machines and PDA's	Specialised machines
Consulting, Assembly and installation, Original spare parts and accessories, Repairs	X	X	X	X
Planning	X	X		
Instruction		X	X	X
Training workshops		X		
Service contracts		X	X	



Question 2.

Simpson’s Paradox

EXAMPLE 1

Ann and Bob are papers reviewers for conferences. They participate in two review cycles: C1 and C2 (e.g., two conferences). Both reviewed the same number of papers in total.

- Ann accepted 55%, Bob accepted 35%

Who is the **stricter reviewer**?

It appears to be Bob, but it’s possible to show that there are cases where **Ann is stricter in both cycles**.

Specifically,

For C1, Ann is stricter since Ann accepted 60% of papers (stricter), Bob accepted 90% of papers;

For C2, Ann is stricter since Ann accepted 10% of papers (stricter), Bob accepted 30% of papers.

TABLE 1. Paper acceptance rates in conferences C1 and C2:

	Conference 1	Conference 2	Total
Ann	60 / 100	1/10	61 / 110
Bob	9 / 10	30 / 100	39 / 110

Discuss the following claims:

1. Bob is a stricter reviewer than Ann.
2. Ann is a stricter reviewer in all cases than Bob.

Which sentence above do you believe (or choose to tell other people)?!

ANSWER :

Both sentences are correct but did not tell the real story and both can be miss-used.

EXAMPLE 2.

Discuss the Simpson’s Paradox in the following example:

TABLE 2. Public-opinion poll of two-Party preferences for Australian Liberal and Labour parties in Brisbane districts:

	Sunny bank	Taringa	Ipswich	St Lucia	Total
Liberal	340/700 (49%)	60/200 (30%)	80/400 (20%)	280/700 (40%)	760/2000 (38%)
Labour	165/300 (55%)	210/600 (35%)	250/1000 (25%)	45/100 (45%)	670/2000 (34%)

According to this table:

1. The Liberal Party lost the poll in every district (implying: the Liberal is losing the Election in these districts).
2. The Liberal Party has a higher overall poll (implying: the Liberal is winning the Election in these districts).

How would this political situation be portrayed? What is your conclusion in order to avoid this kind of problems?

ANSWER :

Again this information could be miss-used by people for different purposes. This problem is not really a paradox, yet non-intuitive.

Generally in If $a/b < A/B$ and $c/d < C/D$, it's possible that $(a+c)/(b+d) > (A+C)/(B+D)$. We are essentially dealing with weighted averages when we combine data segments into an overall evaluation.

To avoid this happening again, the poll should be conducted with the same number of people throughout the experiments for all parties and all districts.

Question 3.

Collection of Data

Read the following paragraph:

Alarm: An increase in cancer among milk drinkers

Cancer, it seems, was becoming increasingly frequent in New England, Minnesota, Wisconsin, and Switzerland, where a lot of milk is produced and consumed. Also it was pointed out, milk-drinking English women get some kinds of cancer eighteen times as frequently as Japanese women who seldom drink milk.

Discuss:

1. The given fact about the increasing of the cancer frequency is true and also the milk-drinking population is true, but the conclusion may not be true. Why?
2. If the following data are also collectable, what would you do in order to give a true story?
 - Data about life spans of women of different nationalities.
 - The average ages of woman of different nationalities who get the cancer.
 - The data about smokers, alcohol consumption, etc of women of different nationalities.
 - The types of cancers in the women populations of different nationalities.
 - Anything else?
3. What should be a general guideline to collect data for the data analysis?
4. Can data be “completely” collected?

ANSWERS

1. This conclusion cannot be made based this correlation between milk-drinking and cancer. More analysis shows that Cancer is predominantly a disease that strikes in middle life or after. Switzerland and the states mentioned first are alike in having populations with relatively long spans of life. English women at the time the study was made were living an average of twelve years longer than Japanese Women. So the process of data collection is wrong and should have collected more relative data for the analysis.
2. All of these data may need to be collected. Some of them may have stronger correlation than others.
3. It is a domain-knowledge guided process to collect all relevant data before a sensible data analysis can be conducted. Statistically the data relevancy and correlations can be evaluated and later on assessed by the domain experts.
4. No, it will never be possible (think about “Butterfly effect” phenomena). Unless we digitize the whole world, there is always a problem to decide what data are relevant or not. In data mining, there is a process namely, feature selection can be applied in dataset preprocessing.

References

Kohavi, R., Mason, L., Parekh, R., and Zheng, Z. 2004. Lessons and Challenges from Mining Retail E-Commerce Data. *Mach. Learn.* 57, 1-2 (Oct. 2004), 83-113.

Huff, Darrell 1993, How to Lie with Statistics, W.W. Norton & Company, 1993.

B. Ganter & R. Wille, 1999, Chapter 1 Concept Lattices of Context, Formal Concept Analysis, Springer 1999.