<Association Rules Problem>

(Khoa Doan)

Executive Summary

D&K is a company that sells various types of newspaper to people. To maximize its profits, the company has organized a marketing campaign. The campaign includes several marketing tasks, each of which is designed for a specific type of newspaper, and aims at a specific group of readers to reduce the marketing cost and increase customers' satisfaction. The association of readers and newspaper sources are based on the company database, which contains information such as the reader's age, marital status, the number of pets he/she has, the music he/she listens to, where he/she lives, the math and language scores, and what newspaper he/she reads. This association can be determined by using the conditional probability, which is a formula measuring how sure we can determine the news source if a characteristic or a combination of characteristics of readers is known.

The conditional probability is calculated for every possible combination of characteristics and the newspaper sources. The relationship between a characteristic and a reader's news source exists only if the probability of this relationship is greater than 50%. The full association is described in the following table:

Reader of	Confidence level	
Telegraaf	100%	Age between 66 and 70
	100%	Age between 56 and 60 and has no pet
	67%	Married and has no pet
	100%	Married and don't like music
	100%	Single, has no pet, and lives in a town
	67%	Has no pet and listen to Pop music
	100%	Has a cat and don't like music
	100%	Don't like music and live in the village
	100%	Don't like music
	100%	Live in Village
Volkskrant	100%	Age between 36 and 45
	60%	Age between 26 and 30
	75%	Age between 26 and 30, and single
	67%	Age between 26 and 30, single, and live in a Town
	100%	Age between 26 and 30, single, and listen to Pop music
	100%	Age between 26 and 30, and listen to Pop music
	57%	Single
	75%	Single and listen to Pop music
	60%	Single and live in a Town
	67%	Married and has a cat
	60%	Has a cat
	100%	Has a cat and listen to Pop music
	67%	Listen to Pop music
	67%	Listen to Pop music and live in a Town
	100%	Listen to Pop music and live in a Country
NRC	100%	Age between 56 and 60, and listen to Classical music
	100%	Age between 66 and 70, and Married
	100%	Has no pet, and listen to Classical music
	67%	Listen to Classical music

Revised: 13 December 2003

Based on this association table, the company can organize its marketing tasks efficiently and profitably. The marketing aims at each newspaper source's reader who has the characteristics as in the rule.

In addition to the findings above, some significant relationships between characteristics other than news source are recognized during the analysis. This information is valuable for many company such as the pet's store, music store, and house seller, thus gives the company some revenue if they sell it. The relationships are below:

Has no pet	67.00%	Age between 10 and 25
	100.00%	Age between 66 and 70
	57.00%	Single
Listen to Pop music	67.00%	Age between 20 and 25
	100.00%	Listen to New Wave music
	75.00%	Live in a country
Listen to Classical music	100.00%	Age between 56 and 60
Live in a Town	60.00%	Age between 26 and 30
	71.00%	Single
	67.00%	Listen to Classical music
Married	60.00%	Has a cat
	75.00%	Live in a Country

Problem Description

D&K newspaper-distribution Company wants to develop a marketing campaign in order to maximize its profits. Since it's currently selling many types of newspaper, including Telegraaf, Volkskrant, NRC and others, the marketing campaign must be developed adaptively to the customers of each newspaper. Therefore, a data mining task is required to find the relationships between the readers and which newspapers they read. The mining is performed on the reader's information stored in the company database. The information includes the reader's age, marital status, the number of pets he/she has, the music he/she listens to, where he/she lives, the math and language scores, and what newspaper he/she reads.

Analysis Technique

The association between readers and newspapers can be expressed in term of a conditional probability. A conditional probability refers to the percentage of an event X which happens in the case an event Y happens. This statement, as well as its formula, can be expressed as below:

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{N(X \cap Y)}{N(Y)}$$

with:

P(X | Y): the percentage of occurrence of X in the occurrence of an event Y

P(X ∩ Y): the percentage of occurrence an event containing both event X and Y in a dataset

P(Y): the percentage of occurrence of an event Y in a dataset

N(X ∩ Y): the number of occurrence an event containing both event X and Y in a dataset

N(Y): the number of occurrence of an event Y in a dataset

With this formula in mind, in order to identify the readers' characteristics that can predict the readers' news source, we just simply find the number of the event containing any combination of the characteristics, and the number of the event containing this combination with the news source. There are 7 attributes, a news source attribute, and a total of 37 categories in this dataset, so there are a total of $2^{37} - 1 = 137438953471$ possible combinations. This number is very high and the analysis on these combinations would take a huge amount of time to complete. Fortunately, there are some attributes that are not significantly related to the news source.

Revised: 13 December 2003

Therefore, we can simply takes these attributes out, and they are the *math score* and *language score*. This leaves us to 2^{30} -1 or 1073741823 combinations. This is still very high, but as the analysis takes place, there are some ways to reduce the number of combinations to a much smaller value.

The mining task is first performed by finding the number of each single attribute in the data set as in appendix 1 in the appendix. There's no need to perform this task for the news source attribute, since it's the instance that the associations are built on. After that, the number of news source's category in the event contains each category of other attributes are counted, as in appendix 2. Now, the percentage P(an attribute category | a news source category) can be calculated by dividing the values in appendix 1 by the values in appendix 2 accordingly, just as in appendix 3. The highlighted values are where there are noticeable relationships between an attribute's category and a news source. The word "noticeable" means the following assumption:

- For an event X and an event Y in the studied dataset, if the number of occurrence of Y is less than 2, Y can be ignored from the analysis.
- For an event X and an event Y in the studied dataset, if P(X | Y) equals or less than 50%, it can be said that there's no relationship between X and Y.

As we can see, with this assumption in mind, the number of combination that we take into account can be significantly decreased. For example, when the analysis to identify a combination of age and marital which can predict a news source, only the attribute's categories that have values greater than 1are included. This can be once again reduced when the news source is incorporated into these attributes with the same rules as above in mind. This is illustrated in appendix 4 below.

The process continues for all possible two-attribute combination. The results are shown in appendix 5 of the appendix. Then 3-attribute and 4-attribute combination are taken into account. The 5-attribute combination isn't studied here since there is no significance with our assumption. The results are presented in the "Result section" below.

In the analysis, in addition to our primary goal, we also find that there exists relationship between other attributes. In order words, there are some attribute's values that can predict other attribute's values. This is presented in the result section below.

Assumptions

Here are some assumptions:

- Math score and language score of readers are not related to a news source
- For an event X and an event Y in the studied dataset, if the number of occurrence of Y is less than 2, Y can be ignored from the analysis.
- For an event X and an event Y in the studied dataset, if P(X | Y) equals or less than 50%, it can be said that there's no relationship between X and Y.

Results

In the analysis, there are association rules as following:

One-attribute Rules			**News**	
Age equals to	**2**		**3**	***60.0%***
	4		**3**	***100.0%***
	5		**3**	***100.0%***
	10		**2**	***100.0%***
Marital equals to	**1**		**3**	***57.1%***
Pet equals to	**2**		**3**	***60.0%***
Music equals to	**1**		**4**	***66.7%***
	3		**3**	***66.7%***
	5		**2**	***100.0%***
Lives equals to	**2**		**2**	***100.0%***

Two-attribute Rules					**News**	
Age equals to	**2**	Marital	**1**		**3**	***75.0%***
		Music	**3**		**3**	***100.0%***
	8	Pet	**1**		**2**	***100.0%***
		Music	**8**		**4**	***100.0%***
Marital equals to	**2**	Pet	**1**		**2**	***67.0%***
		Pet	**2**		**3**	***67.0%***
		Music	**5**		**2**	***100.0%***
	1	Music	**3**		**3**	***75.0%***
		Live	**1**		**3**	***60.0%***
Pet equals to	**1**	Music	**1**		**4**	***100.0%***
		Music	**3**		**2**	***67.0%***
	2	Music	**3**		**3**	***100.0%***
		Music	**5**		**2**	***100.0%***
Music equals to	**3**	Live	**1**		**3**	***67.0%***
		Live	**3**		**3**	***100.0%***
	5	Live	**2**		**2**	***100.0%***

Three-attribute Rules								
Age equals to	**2**	Marital	**1**	Live	**1**		**3**	***67.00%***
Marital equals to	**1**	Pet	**1**	Live	**1**		**2**	***100.00%***

Here is the formal result:

Reader of	Confidence level	
Telegraaf	100%	Age between 66 and 70
	100%	Age between 56 and 60 and has no pet
	67%	Married and has no pet
	100%	Married and don't like music
	100%	Single, has no pet, and lives in a town
	67%	Has no pet and listen to Pop music
	100%	Has a cat and don't like music
	100%	Don't like music and live in the village
	100%	Don't like music
	100%	Live in Village
Volkskrant	100%	Age between 36 and 45
	60%	Age between 26 and 30
	75%	Age between 26 and 30, and single
	67%	Age between 26 and 30, single, and live in a Town
	100%	Age between 26 and 30, single, and listen to Pop music
	100%	Age between 26 and 30, and listen to Pop music
	57%	Single
	75%	Single and listen to Pop music
	60%	Single and live in a Town

Revised: 13 December 2003

	67%	Married and has a cat
	60%	Has a cat
	100%	Has a cat and listen to Pop music
	67%	Listen to Pop music
	67%	Listen to Pop music and live in a Town
	100%	Listen to Pop music and live in a Country
NRC	100%	Age between 56 and 60, and listen to Classical music
	100%	Age between 66 and 70, and Married
	100%	Has no pet, and listen to Classical music
	67%	Listen to Classical music

With this rules, the marketing campaign can more easily to aim at customers who have characteristics similar to the rules with a confidence level as above.

Has no pet	67.00%	Age between 10 and 25
	100.00%	Age between 66 and 70
	57.00%	Single
Listen to Pop music	67.00%	Age between 20 and 25
	100.00%	Listen to New Wave music
	75.00%	Live in a country
Listen to Classical music	100.00%	Age between 56 and 60
Live in a Town	60.00%	Age between 26 and 30
	71.00%	Single
	67.00%	Listen to Classical music
Married	60.00%	Has a cat
	75.00%	Live in a Country

The extra findings can also help the company to generate more profits. The company can sell this information to the Pet store, Music store, and House seller to help them to identify the right target customers.

Issues

One issue is that the number of rows in the database is very small. This creates difficulties in arriving with assumptions to make the result more useful.

Appendices

Appendix 1

Age	1	2	3	4	5	6	7	8	9	10
	3	5	0	1	1	0	0	2	1	2
Marital	1	2	3							
	7	6	2							
Pet	1	2	3	4	5					
	8	5	1	0	1					
Music	1	2	3	4	5					
	3	3	6	0	3					
Live	1	2	3							
	8	3	4							

Appendix 2

News/Age	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	0	0	0	0
2	1	2	0	0	0	0	0	0	1	2
3	1	3	0	1	1	0	0	0	0	0
4	1	0	0	0	0	0	0	2	0	0
5	0	0	0	0	0	0	0	0	0	0

News/Marital	1	2	3
2	2	3	1
3	4	2	0
4	1	1	1

News/Pet	1	2	3	4	5
2	3	2	1	0	0
3	2	3	0	0	1
4	3	0	0	0	0

News/Music	1	2	3	4	5
2	0	1	2	0	3
3	1	1	4	0	0
4	2	1	0	0	0

News/Live	1	2	3
2	2	3	1
3	4	0	2
4	2	0	1

Appendix 3

News/Age	1	2	3	4	5	6	7	8	9	10
1	0.00	0.00	#####	0.00	0.00	#####	#####	0.00	0.00	0.00
2	0.33	0.40	#####	0.00	0.00	#####	#####	0.00	1.00	1.00
3	0.33	0.60	#####	1.00	1.00	#####	#####	0.00	0.00	0.00
4	0.33	0.00	#####	0.00	0.00	#####	#####	1.00	0.00	0.00
5	0.00	0.00	#####	0.00	0.00	#####	#####	0.00	0.00	0.00

News/Marital	1	2	3
2	0.286	0.5	0.5
3	0.571	0.333	0
4	0.143	0.167	0.5

Revised: 13 December 2003

News/Pet	1	2	3	4	5
2	0.375	0.4	1	#####	0
3	0.25	0.6	0	#####	1
4	0.375	0	0	#####	0

News/Music	1	2	3	4	5
2	0	0.333	0.333	#####	1
3	0.333	0.333	0.667	#####	0
4	0.667	0.333	0	#####	0

News/Live	1	2	3
2	0.25	1	0.25
3	0.5	0	0.5
4	0.25	0	0.25

(*Note*: #### stands for non-existence relationship)

Appendix 4

Age	1	2	3	4	5	6	7	8	9	10
	3	5	0	1	1	0	0	2	1	2

Marital	1	2	3
	7	6	2

Full counting table:

Marital/Age	1	2	3	4	5	6	7	8	9	10
1	2	4	0	1	0	0	0	0	0	0
2	1	1	0	0	1	0	0	1	0	2
3	0	0	0	0	0	0	0	1	1	0

Reduced counting table:

Marital/Age	1	2	8	10
1	2	4	0	0
2	1	1	1	2
3	0	0	1	0

Reduced counting table with news source:

Mar-A/News	2	3	4
1-1	1	0	1
1-2	1	3	0
2-10	2	0	0

Appendix 5

Pet/Age	1	2	3	4	5	6	7	8	9	10
1	2	1	0	1	0	0	0	2	0	2
2	1	2	0	0	1	0	0	0	1	0
3	0	1	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	0	1	0	0	0	0	0	0	0	0

P-A/News	2	3	4
1-1	1	0	1
1-8	0	0	2
1-10	2	0	0
2-2	1	1	0

0.50	0.00	0.50
0.00	0.00	1.00
1.00	0.00	0.00
0.50	0.50	0.00

Music/Age	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	1	0	0	2	0	0
2	1	2	0	0	0	0	0	0	0	0
3	2	2	0	1	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0	0
5	0	1	0	0	0	0	0	0	1	1

Mu-A/News	2	3	4
1-8	0	0	2
2-2	1	1	0
3-1	1	1	0
3-2	0	2	0

0.00	0.00	1.00
0.50	0.50	0.00
0.50	0.50	0.00
0.00	1.00	0.00

Live/Age	1	2	3	4	5	6	7	8	9	10
1	1	3	0	1	1	0	0	1	0	1
2	1	0	0	0	0	0	0	0	1	1
3	1	2	0	0	0	0	0	1	0	0

L-A/News	2	3	4
1-2	1	2	0
3-2	1	1	0

0.33	0.67	0.00
0.50	0.50	0.00

Marital/Pet	1	2	3	4	5
1	4	1	1	0	1
2	3	3	0	0	0
3	1	1	0	0	0

Mar-P/News	2	3	4
1-1	1	2	1
2-1	2	0	1
2-2	1	2	0

0.25	0.50	0.25
0.67	0.00	0.33
0.33	0.67	0.00

Marital/Music	1	2	3	4	5
1	0	3	4	0	0
2	2	0	2	0	2
3	1.00	0	0	0	1

Revised: 13 December 2003

Ma-Mu/News	2	3	4
1-2	1	1	1
1-3	1	3	0
2-1	0	1	1
2-3	1	1	0
2-5	2	0	0

0.33	0.33	0.33
0.25	0.75	0.00
0.00	0.50	0.50
0.50	0.50	0.00
1.00	0.00	0.00

Marital/Live	1	2	3
1	5	1	1
2	2	1	3
3	1	1	0

Mar-L/News	2	3	4
1-1	1	3	1
2-1	1	1	0
2-3	1	1	1

0.20	0.60	0.20
0.50	0.50	0.00
0.33	0.33	0.33

Pet/Music	1	2	3	4	5
1	2	2	3	0	1
2	1	0	2	0	2
3	0	1	0	0	0
4	0	0	0	0	0
5	0	0	1	0	0

P-Mu/News	2	3	4
1-1	0	0	2
1-2	0	1	1
1-3	2	1	0
2-3	0	2	0
2-5	2	0	0

0.00	0.00	1.00
0.00	0.50	0.50
0.67	0.33	0.00
0.00	1.00	0.00
1.00	0.00	0.00

Pet/Live	1	2	3
1	5	2	1
2	2	1	2
3	1	0	0
4	0	0	0
5	0	0	1

Revised: 13 December 2003

P-L/News	2	3	4
1-1	1	2	2
1-2	2	0	0
2-1	0	2	0
2-3	1	1	0

0.20	0.40	0.40
1.00	0.00	0.00
0.00	1.00	0.00
0.50	0.50	0.00

Music/Live	1	2	3
1	2	0	1
2	3	0	0
3	3	1	2
4	0	0	0
5	0	2	1

Mu-L/News	2	3	4
1-1	0	1	1
2-1	1	1	1
3-1	1	2	0
3-3	0	2	0
5-2	2	0	0

0.00	0.50	0.50
0.33	0.33	0.33
0.33	0.67	0.00
0.00	1.00	0.00
1.00	0.00	0.00

Mar-A/Pet	1	2	3	4	5
1-1	2	0	0	0	0
1-2	1	1	1	0	1
2-10	2	0	0	0	0

Mar-A-Pet/News	2	3	4
1-1-1	1	0	1

0.50	0.00	0.50

Mar-L/Age	1	2	3	4	5	6	7	8	9	10
1-1	1	3	0	1	0	0	0	0	0	0
2-1	0	0	0	0	1	0	0	0	0	1
2-3	1	1	0	0	0	0	0	1	0	0

Mar-L-A/News	2	3	4
1-1-2	1	2	0

0.33	0.67	0.00

Ma-Mu/Age	1	2	3	4	5	6	7	8	9	10
1-2	1	2	0	0	0	0	0	0	0	0
1-3	1	2	0	1	0	0	0	0	0	0
2-1	0	0	0	0	1	0	0	1	0	0
2-3	1	0	0	0	0	0	0	0	0	1
2-5	0	1	0	0	0	0	0	0	0	1

Ma-Mu-A/News	2	3	4
1-2-2	1	1	0
1-3-2	0	2	0

0.50	0.50	0.00
0.00	1.00	0.00

Revised: 13 December 2003

Mu-A/Pet	1	2	3	4	5
1-8	2	0	0	0	0
2-2	1	0	1	0	0
3-1	1	1	0	0	0
3-2	0	1	0	0	1

Mu-A-Pet/News	2	3	4
1-8-1	0	0	2

0	0	1

L-A/Pet	1	2	3	4	5
1-2	1	1	1	0	0
3-2	0	1	0	0	1

No further analysis

P-Mu/Marital	1	2	3
1-1	0	1	1
1-2	2	0	0
1-3	2	1	0
2-3	1	1	0
2-5	0	1	1

P-Mu-Ma/News	2	3	4
1-2-1	0	1	1
1-3-1	1	1	0

0	0.5	0.5
0.5	0.5	0

P-L/Marital	1	2	3
1-1	3	1	1
1-2	1	1	0
2-1	1	1	0
2-3	0	2	0

P-L-Mar/News	2	3	4
1-1-1	0	2	1
2-3-2	1	1	0

0	0.667	0.333
0.5	0.5	0

Mu-L/Pet	1	2	3	4	5
1-1	1	1	0	0	0
2-1	2	0	1	0	0
3-1	1	1	0	0	0
3-3	0	1	0	0	1
5-2	1	1	0	0	0

Mu-L-Pet/News	2	3	4
2-1-1	0	1	1

0	0.5	0.5

A-Mar-Pet/Music	1	2	3	4	5
1-1-1	0	1	1	0	0

No further analysis

Revised: 13 December 2003

A-Mar-Pet/Live	1	2	3
1-1-1	1	1	0

No further analysis

A-Mar-Mu/Live	1	2	3
2-1-2	2	0	0
2-1-3	1	0	1

A-Mar-Mu-Live/News	2	3	4
2-1-2-1	1	1	0

A-P-Mu/Live	1	2	3
8-1-1	1	0	1

No further analysis

Mar-P-Mu/Live	1	2	3
1-1-2	2	0	0
1-1-3	1	1	0

Mar-P-Mu-L/News	2	3	4
1-1-2-1	0	1	1