Topic 9. Association Rules

Case 6: Web Viewership Associations

using page sequence data at MSNBC

Reading Assignment

Berry and Linoff (2000)

- Page 307. Market basket analysis. Pages 424-429. Association rules.

Associateion Rules

Association rule mining finds interesting relationships in data.

The leading special case is Market Basket Analysis. Some market baskets at checkout:

1. `bread, peanut_butter, jelly`
2. `vodka, caviar`
3. `milk, cereal`

Some association rules:

1. `bread` $\leftarrow$ `peanut_butter, jelly`
2. `caviar` $\leftarrow$ `vodka`

The itemset to the left of the arrow is the consequent. The itemset to the right is the antecedent.

Nomenclature

- **Itemset:** Sets of items. E.g., `{ peanut_butter, jelly }`
- **Support:** The probability of an itemset as estimated by the percentage occurance of the itemset in the data. E.g., the support of the itemset `{ peanut_butter, jelly }` is

$$P(\{\text{peanut\_butter, jelly}\}) = \frac{\#\{\text{ peanut\_butter, jelly }\}}{\#\text{observations}}$$

- **Confidence:** The conditional probability of one itemset given another as estimated in the data. E.g., the confidence of the rule `bread` $\leftarrow$ `peanut-butter, jelly` is

$$P(\{\text{bread}\}|\{\text{peanut\_butter, jelly}\}) = \frac{\#\{\text{ bread, peanut\_butte}}{\#\{\text{ peanut\_butter, j}}$$

- **Lift:** The confidence of a rule divided by the support of the consequent. E.g., the lift of `bread` $\leftarrow$ `peanut-butter, jelly` is

$$\frac{P(\{\text{bread}\}|\{\text{peanut\_butter, jelly}\})}{P(\{\text{bread}\})} = \frac{\#\{\text{ bread, peanut\_bu}}{\#\{\text{ bread }\}\#\{\text{ peanut\_}}$$

The Goal of Associative Rule Mining

The goal of associative rule data mining is to find all associateve rules that have high confidence in the data set.

The problem is impossible as stated; i.e., it is too computationally intensive.

One has to cut corners in designing algorithms. The most popular algorithm is the apriori algorithm due to Agrawal, Mannila, Srikaut, Toivonen and Verkamo. The most popular implementation of it is the public domain code due to Christian Borgelt at `www.borgelt.net`

The Apriori Algorithm

- Compute the support of all singleton itemsets. Discard those that are smaller than a threshold (default 10%).
- Compute support of all doubleton itemsets made up of survivors from the previous step.
- $\cdots$
- Blend in some advanced data structure techniques.
- $\cdots$

Concern: A rule such as

$$\text{caviar} \leftarrow \text{vodka}$$

will not be discovered because the support of `{ caviar }` is smaller than any reasonable threshold.

MSNBC Web Page Visits

The data list the web page visits of all who visited MSNBC on September 28, 1999.

Visits are recorded in time order for main pages – subpage visits are not recorded.

The main pages are frontpage, news, tech, local, opinion, on-air, misc, weather, health, living, business, sports, bbs, travel.

The first nine records look like this

1 1
2
3 2 2 4 2 2 2 3 3
5
1
6
1 1
6
6 7 7 7 6 6 8 8 8 8

There are 989,818 records. The average number of visits per record is 5.7.

Input to Apriori

The data are in the format expected by apriori. However, replacing the page number by the page names will make apriori output easier to understand. E.g.,

The first seven records will look like this

frontpage frontpage
news
tech news news local news news news tech tech
opinion
frontpage
on-air
frontpage frontpage

Analysis

We shall analyze these data from the perspective of two business purposes

1. Page pushing and caching.

2. Advertisement placement.

What Page Will be Viewed First?

Method

1. Generate a data set that has the first page viewed on each line.

2. Compute the itemset supports for those data.

Table 29. First Page Support

Page	Support
frontpage	28.0
on-air	16.5
news	14.4
sports	11.8
weather	7.3
tech	6.7
business	5.7
local	5.2
living	1.4
health	1.3

Itemsets with support less than 1% are not shown.

What Page Will be Viewed Second?

Method

1. Generate a data set that has the name of the first page viewed followed by the name of the second with a "-2" suffix appended on each line.

2. Compute the association rules for that data set.

Business purpose

1. Page pushing and caching.

2. Advertisement placement.

Table 30. Second Page Association Rules

Association Rule			Support	Confidence	Lift
weather-2	←	weather	5.0	68.4	1158.4
sports-2	←	sports	6.6	55.7	613.8
local-2	←	local	2.1	40.9	951.0
news-2	←	news	5.2	36.5	366.2
frontpage-2	←	frontpage	10.1	36.0	300.8
health-2	←	health	0.4	34.3	2169.6
business-2	←	business	1.9	33.5	830.4
living-2	←	living	0.4	25.1	1206.7
tech-2	←	tech	1.4	21.1	593.9
on-air-2	←	on-air	2.9	17.7	367.3
news-2	←	frontpage	2.6	9.4	94.7
misc-2	←	on-air	1.2	7.1	204.0
news-2	←	business	0.4	6.7	67.1
frontpage-2	←	living	0.1	5.8	48.4
business-2	←	frontpage	1.4	4.9	122.8
news-2	←	health	0.1	4.7	47.7
news-2	←	local	0.2	4.7	47.4
sports-2	←	frontpage	1.3	4.7	51.6
living-2	←	frontpage	1.3	4.6	221.6

A rule such as weather-2 ← weather is a page refresh or return from a subpage. Rules with confidence less than 4.5% are not shown.

Eliminating Refreshes, What Page Will be Viewed Second?

Method

1. Eliminate all page repeats. Generate a data set that has the name of the first page viewed followed by the name of the second with a "-2" suffix appended on each line.

2. Compute the association rules for that data set.

Business purpose

1. Page pushing and caching.

2. Advertisement placement.

Table 31. Second Page Association Rules, No Refresh

Association Rule			Support	Confidence	Lift
news-2	←	frontpage	2.6	9.4	194.5
misc-2	←	on-air	1.2	7.1	210.7
news-2	←	business	0.4	6.7	137.9
frontpage-2	←	living	0.1	5.8	308.1
business-2	←	frontpage	1.4	4.9	234.3
news-2	←	health	0.1	4.7	97.9
sports-2	←	frontpage	1.3	4.7	154.6
news-2	←	local	0.2	4.7	97.4
living-2	←	frontpage	1.3	4.6	268.2
sports-2	←	sports	0.5	4.4	145.2
frontpage-2	←	news	0.6	4.3	226.7
misc-2	←	sports	0.5	4.3	127.8
on-air-2	←	frontpage	1.1	4.1	214.7
news-2	←	tech	0.3	4.0	82.3
news-2	←	living	0.1	3.8	79.2
misc-2	←	news	0.5	3.8	112.0
news-2	←	on-air	0.6	3.8	77.9
tech-2	←	frontpage	1.0	3.5	163.8

Refresh sequences are deleted from each record before computing rules. Rules with confidence less than 3.5% are not shown.

Main Points

- Association rules mining finds interesting relationships in data.

- The apriori algorithm can quickly analyze very large commercial data sets.

- Association rules are among data mining's biggest successes (Hastie, Tibshiranai, and Friedman, 2001).