COMP4433 Data Mining and Data Warehousing (2019-2020 Sem. 1)

Individual Project Report (Weblog Mining)

CHENG Yiran 16098521d

Outline

I. Data Loading and Pre-processing

II. Association Rule Mining for Web Pages (Apriori)

III. Sequential ARM for Web Pages (AprioriAll) *(if the original data is sequential)*

IV. User intention clustering

V. Strategy for the Microsoft Website Developers

References

I. Data Loading and Pre-processing

Number of users = 32711, Number of web pages = 294

Code: function *readfile* in individual_project.py

Special pre-processing method will be introduced in III.3.

II. Association Rule Mining for Web Pages (Apriori)

1. **Motivation**
   We may find association rules from the dataset to detect the users' purpose of visiting this website. Also, apply association rule mining can help the developers understand how the webpages are visited and the relationship between two pages.

2. **Code & Parameters**

   min_support = 500 (≈1.5%), min_confidence = 0.5

   The function ***ARM*** in individual_project.py. Use the pymining library [1].

3. **Frequent Itemset**
   a) Top 10 frequent visited with 1 item

** Due to the pages limit, I only show part of the data found here. **

ID	*Name*	*Occurrence*
1025	Web Site Builder's Gallery	2123
1003	Knowledge Base	2968
1026	Internet Site Construction for Developers	3220
1001	Support Desktop	4451
1009	Windows Family of OSs	4628

1017	Products	5108
1018	isapi	5330
1004	Microsoft.com Search	8463
1034	Internet Explorer	9383
1008	Free Downloads	10836

*Table **(ARM - Result 1)***

4. Association Rules

a) Frequent Itemsets

Totally **<u>114</u>** frequent itemsets. ***(ARM - Result 2)***

b) Association Rules (min_support = 500 (≈1.5%), min_confidence = 0.5)

(Remark: due to the page limit, I only show the rules with more than 3 elements here.)

{1035: Windows95 Support, 1018: isapi, 1003: Knowledge Base, } => {1001: Support Desktop, }, Sup=502, Conf=0.7254335260115607

{1035: Windows95 Support, 1003: Knowledge Base, } => {1001: Support Desktop, 1018: isapi, }, Sup=502, Conf=0.6354430379746835

{1035: Windows95 Support, 1001: Support Desktop, 1003: Knowledge Base, } => {1018: isapi, }, Sup=502, Conf=0.8745644599303136

{1001: Support Desktop, 1035: Windows95 Support, } => {1018: isapi, 1003: Knowledge Base, }, Sup=502, Conf=0.518595041322314

{1035: Windows95 Support, 1001: Support Desktop, 1018: isapi, } => {1003: Knowledge Base, }, Sup=502, Conf=0.6330390920554855

{1009: Windows Family of OSs, 1018: isapi, 1035: Windows95 Support, } => {1008: Free Downloads, }, Sup=666, Conf=0.7085106382978723

{1009: Windows Family of OSs, 1035: Windows95 Support, } => {1008: Free Downloads, 1018: isapi, }, Sup=666, Conf=0.6201117318435754

{1008: Free Downloads, 1018: isapi, 1035: Windows95 Support, } => {1009: Windows Family of OSs, }, Sup=666, Conf=0.826302729528536

{1008: Free Downloads, 1035: Windows95 Support, } => {1009: Windows Family of OSs, 1018: isapi, }, Sup=666, Conf=0.748314606741573

{1008: Free Downloads, 1009: Windows Family of OSs, 1035: Windows95 Support, } => {1018: isapi, }, Sup=666, Conf=0.9073569482288828

{1008: Free Downloads, 1009: Windows Family of OSs, 1018: isapi, } => {1035: Windows95 Support, }, Sup=666, Conf=0.6727272727272727

Table: (ARM - Result 3)

5. Patterns Retrieved from Data Mining

a) Most users use Microsoft website for downloading, knowing more about IE and using the search service. *(Ref: ARM - Result 1)*

b) We can summarize some classical usage scenarios from Association Rule Mining such as downloading Windows 95. *(Ref: ARM - Result 3)*

c) During the data mining process, we found that page index 1018: isapi seems unique. If the dataset let us know the real meaning of it, we may remove it based on real situation. *(Ref: ARM - Result 3)*

6. Further Improvement

To improve the current analysis, we may additionally use lift rather than support and confidence only to determine the importance of the rules [2].

III. Sequential ARM for Web Pages (AprioriAll) *(if the original data is sequential)*

1. Motivation

The dataset description **didn't** mention if the data is sequential for each user. So, for this part, it is just a trial. I tried sequential ARM to show my understanding about it and I know that it is only applicable when the data is sequential.

2. Code & Parameters

Min_support = 0.02

The function *sequentialARM* in individual_project.py.
Reference: *aprioriall.py* from the internet [3].

3. Pre-processing

Remove the users with only one visiting page because it is useless for sequential association rule mining. *(set the PREPROCESSING to True in code)*

After the pre-processing, the running time is less than before and the filtered data scale is: number_of_user = 22717.

4. Frequent Sequences

Remark: Each line represents one sequence. Only shows sequences whose length >= 2. The sequences with item "1018: isapi" are removed manually.

1001: Support Desktop, 1003: Knowledge Base
1001: Support Desktop, 1004: Microsoft.com Search
1001: Support Desktop, 1008: Free Downloads
1001: Support Desktop, 1009: Windows Family of OSs
1001: Support Desktop, 1017: Products
1001: Support Desktop, 1034: Internet Explorer
1001: Support Desktop, 1035: Windows95 Support
1003: Knowledge Base, 1004: Microsoft.com Search
1003: Knowledge Base, 1008: Free Downloads
1003: Knowledge Base, 1035: Windows95 Support
1004: Microsoft.com Search, 1008: Free Downloads
1004: Microsoft.com Search, 1009: Windows Family of OSs
1004: Microsoft.com Search, 1017: Products

1004: Microsoft.com Search, 1034: Internet Explorer
1008: Free Downloads, 1009: Windows Family of OSs
1008: Free Downloads, 1017: Products
1008: Free Downloads, 1026: Internet Site Construction for Developers
1008: Free Downloads, 1034: Internet Explorer
1008: Free Downloads, 1035: Windows95 Support
1009: Windows Family of OSs, 1017: Products
1009: Windows Family of OSs, 1034: Internet Explorer
1009: Windows Family of OSs, 1035: Windows95 Support
1009: Windows Family of OSs, 1037: Windows 95
1017: Products, 1034: Internet Explorer
1025: Web Site Builder's Gallery, 1026: Internet Site Construction for Developers
1026: Internet Site Construction for Developers, 1034: Internet Explorer
1026: Internet Site Construction for Developers, 1038: SiteBuilder Network Membership
1026: Internet Site Construction for Developers, 1041: Developer Workshop

Maximal (Frequent) Sequences ***(SeqARM - Result 1)***

5. Patterns Retrieved from Data Mining

a) Internet Explorer and Windows 95 are two most commonly downloaded software. *(Ref: SeqARM - Result 1)*

b) People usually search for downloads, products, IE and Oss. *(Ref: SeqARM - Result 1)*

c) In product page, people usually want to learn about IE. *(Ref: SeqARM - Result 1)*

IV. User intention clustering

1. Motivation

Based on the users' visiting records, we may cluster the users into different groups. Each group of people may have some same purpose for visiting the Microsoft website.

2. Design

a) Distance measure:

For each user u_i he/she have a visiting list V_i. When two users visit more same websites, they should have smaller distance between them. So, the distance between two user u_i and u_j is calculated as:

$$\text{Distance}(u_i, u_j) = 1 - \text{Size}(Vi \cap Vj) / \text{Size}(Vi \cup Vj)$$

And the distance should be between 0 and 1.

b) Method selection & "bridge" problem:

I first select AGNES with complete linkage clustering and group average clustering. Why select these two clustering methods? It's because these two methods can distinguish two cluster rather than combine them together when they have some intersections.

For example (why not use single linkage clustering because of the "bridge" problem):

The users in group A has such visiting websites ID:

(1, 2, 3, 4, 5)

(only common items are shown, the below is the same)

The users in group B has such visiting websites ID:

(60, 70, 80, 90)

Now we have a user *u* with visiting websites:

(4, 5, 60, 70)

Then, it is easy to cluster the user *u* to either group A or group B. Let's assume it is group A that u is clustered to. If we apply single linkage clustering, then, it is easy to combine group A and group B to one cluster since they have a **"bridge"** – user *u*.

This isn't what we want. We want clear purpose in each group rather (i.e. keep group A and group B as two clusters).

Why not choose density-based method?

Because it can't avoid such above "bridge" problem.

c) Adjustment and sampling:

After my first trial, I found that the programme was keeping running for a very long time. After my calculation, the time complexity is not acceptable for over 30000 users. Therefore, I designed two solutions to solve this problem.

(i) *Reduce the dataset scale (Sampling):*

Only read 300~1000 users as the sample for processing and analysis.

(ii) *Use constant threshold for clustering decision*

First, I calculate all the distances for each user pair. Then, I sorted and print the distribution of the distances and found that the distribution of the distance is:

0%=0.0, 25%=0.875, 50%=1.0, 75%=1.0, 100%=1.0

*25% = 0.875 means the distance 0.875 ranked 25% in all distances.

Then, I choose 0.875 as the threshold for clustering. When any two existing cluster has a distance less than 0.875, combine them.

3. Original Clustering Method and Results (Method 1)

a) Description

This is the original AGNES with complete linkage clustering and group average clustering method without applying the adjustment in 2.c).(ii).

Data: the first 300 users.

b) How to decide the termination of the clustering rather than let it cluster into one set?

I draw a graph and check the min_distance in each merging step:

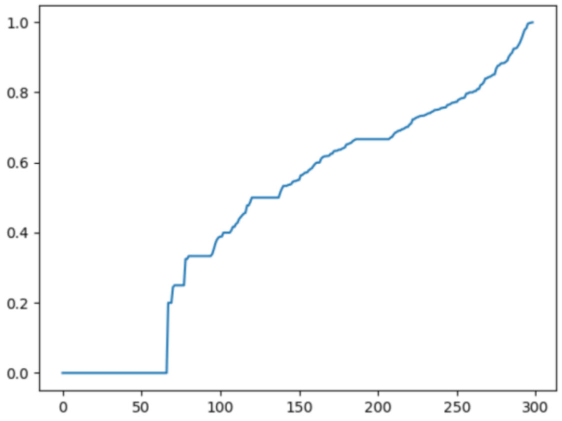

Figure: min_distance growing (x: n-cluster, y: min_distance)

Found that around 150 is a proper number of clustering since the distance later is significantly growing.

c) Parameters

CLUSTER_MIN_PRINT_SIZE = 5 # Only print clusters with more than 5 elements.
DATASCALE_LIMIT = 300 # Sample 1000 data from the dataset.
N_OF_CLUSTERS = 150 # Merge 150 times and end.
DEFAULT_MODE = "complete" # Using complete linkage clustering rather than group avg.
PREPROCESSING = False # Do not pre-process.

Remark: Since using complete linkage clustering and using group average clustering have similar results, in this report, only one of their results are shown below.

d) Clustering Result

Finally, the clusters and their centroid are as below. Since very small cluster is meaningless for pattern retrieve, I only show the cluster with no less than 5 users inside below.

Description: The first line of each box is the cluster centroid number and the cluster size. The second line is the centroid user's visiting record.

Cluster centroid: 10006 Cluster size: 7 1003: Knowledge Base, 1004: Microsoft.com Search,
Cluster centroid: 10009 Cluster size: 13 1008: Free Downloads, 1009: Windows Family of OSs,
Cluster centroid: 10016 Cluster size: 13 1025: Web Site Builder's Gallery, 1026: Internet Site Construction for Developers,
Cluster centroid: 10072 Cluster size: 6 1017: Products, 1004: Microsoft.com Search,
Cluster centroid: 10040 Cluster size: 41 1008: Free Downloads, 1034: Internet Explorer,
Cluster centroid: 10088 Cluster size: 7 1008: Free Downloads, 1007: International IE content,
Cluster centroid: 10063 Cluster size: 7 1034: Internet Explorer, 1018: isapi,

```
Cluster centroid: 10089 Cluster size: 8
1035: Windows95 Support, 1001: Support Desktop, 1003: Knowledge Base, 1018: isapi,
```

Clustering result (cluster size>5) ***(Cluster – Result 1)***

4. Simplified Clustering Method (Method 2)

a) Description

It is a simplified method for clustering. It may have some problem because it based on the distance value. It combines two cluster once their distance is less than a specific value.

b) Parameters

CLUSTER_MIN_PRINT_SIZE = 10 # Only print clusters with more than 10 elements.
DATASCALE_LIMIT = 1000 # Sample 1000 data from the dataset.
CLUSTER_THRESHOLD = 0.4 # the max distance available for merging two clusters is 0.4.
DEFAULT_MODE = "groupavg" # Using group Avg. clustering rather than complete linkage
PREPROCESSING = False # Do not pre-process.

Remark: Since using complete linkage clustering and using group average clustering have similar results, in this report, only one of their results are shown below.

c) Clustering Result

Description: The first line of is the cluster centroid number and the cluster size. The second line is the centroid user's visiting record. The third line is another cluster's centroid ...

```
Cluster centroid: 10008 Cluster size: 64
1004: Microsoft.com Search,
Cluster centroid: 10009 Cluster size: 22
1008: Free Downloads, 1009: Windows Family of OSs,
Cluster centroid: 10016 Cluster size: 21
1025: Web Site Builder's Gallery, 1026: Internet Site Construction for Developers,
Cluster centroid: 10023 Cluster size: 24
1008: Free Downloads,
Cluster centroid: 10026 Cluster size: 54
1034: Internet Explorer,
Cluster centroid: 10040 Cluster size: 72
1008: Free Downloads, 1034: Internet Explorer,
Cluster centroid: 10033 Cluster size: 13
1032: Games,
Cluster centroid: 10050 Cluster size: 17
1025: Web Site Builder's Gallery,
Cluster centroid: 10097 Cluster size: 17
1004: Microsoft.com Search, 1034: Internet Explorer,
Cluster centroid: 10163 Cluster size: 13
1009: Windows Family of OSs
```

Clustering result (cluster size>10) ***(Cluster – Result 2)***

5. Patterns Retrieved from Data Mining

a) We can find some common usage scenarios from the clustering results such as visiting "Internet Site Construction for Developers" to view "Web Site Builder's Gallery". *(Ref: Cluster – Result 1 & 2)*

b) There are some different types of users with different purpose such as visiting developer website to check the reference, search for some information, and download Internet Explorer. *(Ref: Cluster – Result 1 & 2)*

c) The result here can be used to build personas to analyse different types of users' behaviour. *(Ref: Cluster – Result 1 & 2)*

d) Some most common scenarios are showed as the result. *(Ref: Cluster – Result 1 & 2)*

V. Strategy for the Microsoft Website Developers

Strategy	**Refer to technique/data**
Improve and allocate more hardware and network resource for the web pages which are frequently visited.	***ARM - Result 1*** e.g. Free Downloads
In a web page, use larger/obvious hyperlink for the pages which are more likely to be visited next.	***ARM - Result 3 & SeqARM - Result 1***
Use direct link from A to C rather than A to B to C.	***ARM - Result 3 & SeqARM - Result 1***
Consider combine the content of two pages which are usually visited together.	***ARM - Result 3 & SeqARM - Result 1***
Use some direct link for users to download Internet Explorer in the download page.	***SeqARM - Result 1***
Consider combine the content of page 1025 and 1026 since they are usually visited together.	***SeqARM - Result 1,*** ***Cluster – Result 1 & 2***
Based on the purpose of users, design special functions directly to meet their needs.	***Cluster – Result 1 & 2***

References:

[1] Andy_shenzl. "关联规则 (Association Rules) 原理分析及实例 python 实现". Available: https://blog.csdn.net/Andy_shenzl/article/details/83084572

[2] U. Malik. "Association Rule Mining via Apriori Algorithm in Python". Available: https://stackabuse.com/association-rule-mining-via-apriori-algorithm-in-python/

[3] L. Tang. "序列模式挖掘-AprioriAll 算法详解". Available: http://hexo.tanglei.name/blog/aprioriall-algorithm-in-python.html