Knowledge Discovery And Data Mining Course Term Project

Sequential Pattern Mining on Webpage Hits Data

Ashwini Kulkarni, Ningle Lei

Computer Science Department, Indiana Universtiy-Purdue University
Fort Wayne, IN, 46805,

{ kulka02, lein01, }@students.ipfw.edu

Abstract: Sequential pattern mining has been heavily investigated within data mining field. One of the wide range of applications is in webpage hits data. In this paper, we want to answer three questions to the webpage hits data: what are the top 10 most frequently visited web pages? What are the sequential patterns of the visited web pages? What are the association rules of the frequently visited web pages? One of the purposes to answer these question is to provide advertisement solutions. In this paper three approaches are taken respectively to answer those three questions.

Keywords: sequential pattern mining, frequent item set, association rules

---

1 Introduction

Large amount of web data has been generated by World Wide Web page users because of its accessibility. The greatest advantage of having the large amount of web data is to provide services to advertisement companies or create a dynamic user profile, which both of these purposes require us to analyze the data and find the hidden information. Hidden information for web data include identifying frequently visited website pages, in what order a user is visiting the web pages and how likely a user would visit page C given he has already visited page A and B, etc. Once we find the hidden information from web data, we could provide solutions to what advertisements should be posted so a user would likely to click on. However, sometimes, only using data mining is not enough to completely solve the current problem, we will also need to comprehend data mining approach with other techniques such as machine learning, natural language processing, etc.

In this paper, we will mainly focus on using data mining approach, such as frequent pattern mining to achieve our goal. Our goal is to provide answers to the following three questions: what are the top 10 most frequently visited web pages? What are the sequential patterns of the visited web pages? What are the association rules of the frequently visited web pages? Our data set is from MSNBC.com and is called Anonymous Web Data Data Set. It contains 989,818 instances. Each instance is a sequence of a user's page views at a level of page category during twenty-four hour's periods. The focus of this paper is to provide an overview of how to use frequent pattern mining to discover frequent item set (frequently visited web pages), sequence pattern recognition. For each of the problem, an algorithm has been designed and implemented. The frequent item set (frequent page set) is discovered using basic Apriori algorithm approach.

The paper is organized in the following order: in section 2, we will talk about some related work that has been done in the same field. Section 3 data preprocessing and methodology of solutions to our problems. Section 4 is about technologies used in the project. Section 5 is about the approach we took to reach our goal. Section 6 is what kind of result we have at the end.

2 Related Work

- https://www.uni-obuda.hu/journal/Ivancsy_Vajk_5.pdf
  Its previous work on web page log data using algorithm such as SM-Tree algorithm
- https://pdfs.semanticscholar.org/cb3b/fa9ef75431c745c4ab947c6b35a3c5e668c5.pdf
  The sequence tree algorithm is one such web usage mining technique which extracts frequent sequential patterns by formation of a tree
- http://www.datalab.uci.edu/papers/webcanvas.pdf
  They have developed a simple approach for clustering and visualizing user behavior on a web site, and implemented approach method in a visualization tool called WebCANVAS..
- User Behavior analysis
- Scientific Research (Similar Chemical Compound, Biological Study)

- Language Processing
- Targeted Marketing

3 Methodology

- **Frequent item set**
  Given appropriate support, list all the number of times a certain page visited. The ones that are visited more than support threshold will be counted as frequent item set.

- **Frequent sequence mining**
  Given appropriate support, list a sequence of page visited. We went through different combinations of how different pages get visited in different order and recorded the number of times each certain sequence occurred, and lastly find the support for those sequence and list all the sequence that surpass the minimum support as frequent sequence.

- **Association rule mining**
  Given appropriate support and confidence level, we used buit-in functions in python to find potential association rules. Each rule is in a form of list and once we get the list, we can clean up the result by denoting each letter by its name.

4 Technology

Platform: Spyder , Rstudio
Programing language: Python, R

5 Dara Description

This data describes the page visits of users who visited msnbc.com on one day. Visits are recorded at the level of URL category (see description) and are recorded in time order.
Type: Sequential
Instance: 998999
Categories: 17
Input Log File:-

```
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
6 9 4 4 4 10 3 10 5 10 4 4 4
1 1 1 11 1 1 1
12 12
1 1
8 8 8 8 8 8
6
2
9 12
3
```

Integer Value	Webpage Name
1	Front-page
2	News
3	Tech
4	Local
5	Opinion
6	on-air
7	Misc
8	Weather
9	msn-news
10	Health
11	Living
12	Business
13	msn-sports
14	Sports
15	Summary
16	Bbs
17	Travel

6 Approach

Marcov Chain:

We used R Studio library for Marcov chain.

Using fitMarkovChain() function we can generate transition probability Graph which shown in result section of this report

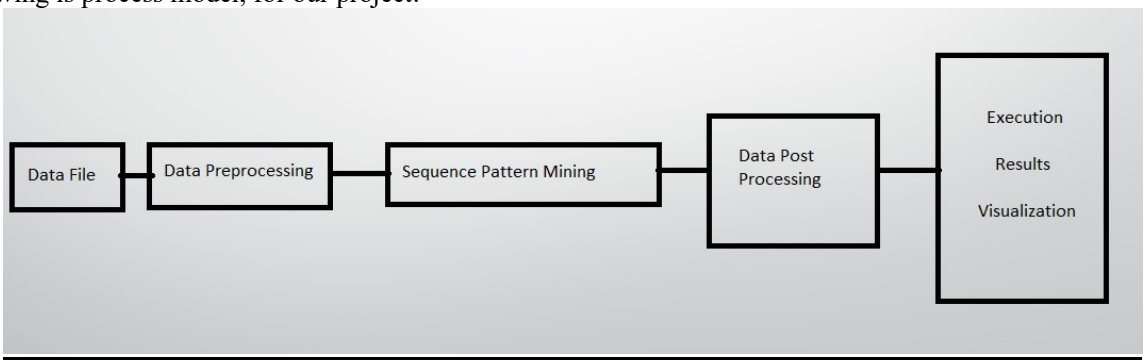
```
library(clickstream)
library(arulesSequences)
library(ggplot2)

clickstreams1 <- readClickstreams(file = "C:/Python/DataMining/Proccessed.txt", se
print(clickstreams1)
print(clickstreams1[4])
print(clickstreams1[5])


mc <- fitMarkovChain(clickstreams1)
```

Sequence Pattern Mining Approch:

Following is process model, for our project:



Using basic Apriori concept we developed algorithm as follows

Minimum Support is: 0.1 of no of records
1.....K=1
2.....Item set size K with minimum Support
3.....Add frequent set in temp array
4.....for each itemset t in temp array
prefix= t
add prefix to final
5..... find frequent item after visiting t with minimum support
if not null (now item set size is k =k+1) gotostep 3
else continue

After getting Frequent sequence pattern, we need to do post processing of data for better understanding and analysis.

7 Data Preprocessing

Removing Extra spaces, lines from raw data(log file).

Here 17 categories are represented as numbers, converting it into single character '11' as 'k', which is easy for computation.

```
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
6 9 4 4 4 10 3 10 5 10 4 4 4
1 1 1 11 1 1 1
12 12
1 1
```

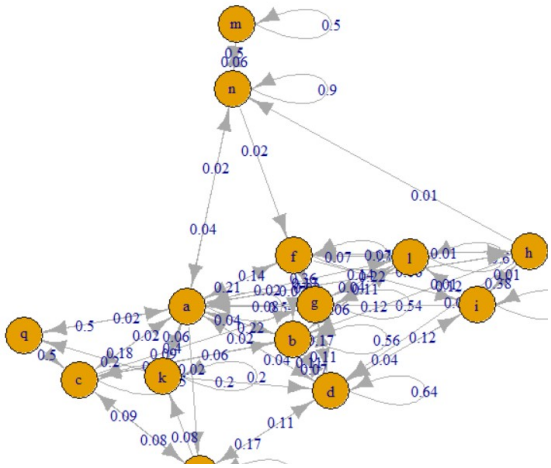
```
a a
b
c b b d b b b c c
e
a
f
a a
f
f g g g f f h h h h
f i d d d j c j e j d d d
a a a k a a a
l l
a a
```

8 Data Post-processing:

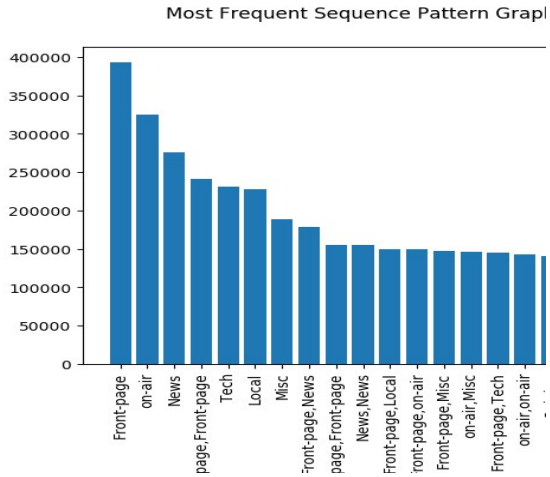
After getting Frequent Sets from Frequent Pattern Mining function, we need to process data which will be in format of alphabets a to q. So need to convert it in actual page_name. Also Finding Size of each item set, and frequencies so we can plot graph/report which can be used for decision making.

9 Result

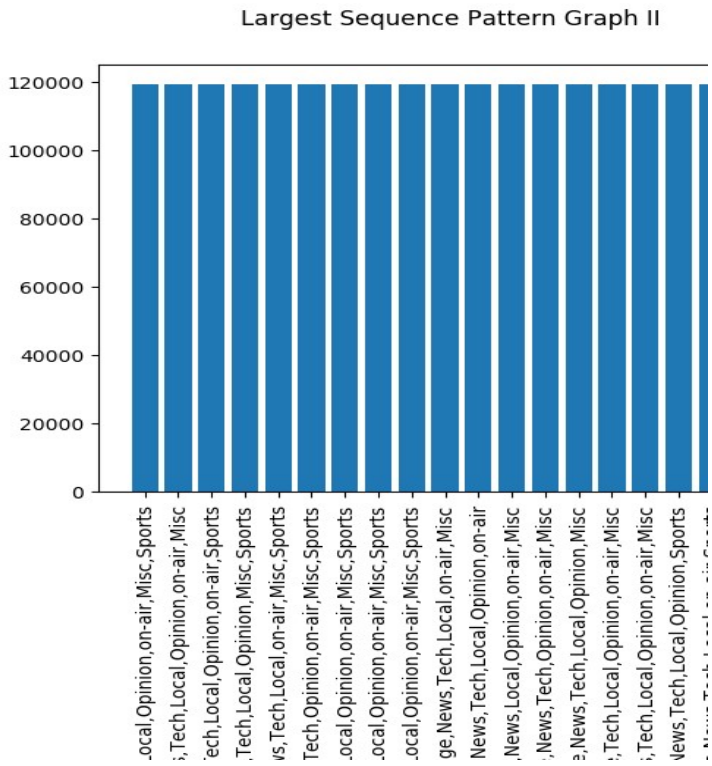
From given input fine, finding following patterns:
Marcov chain behavior of user page visit:
This graph show what is probability that user visit page ‘a’ after page ‘b’.



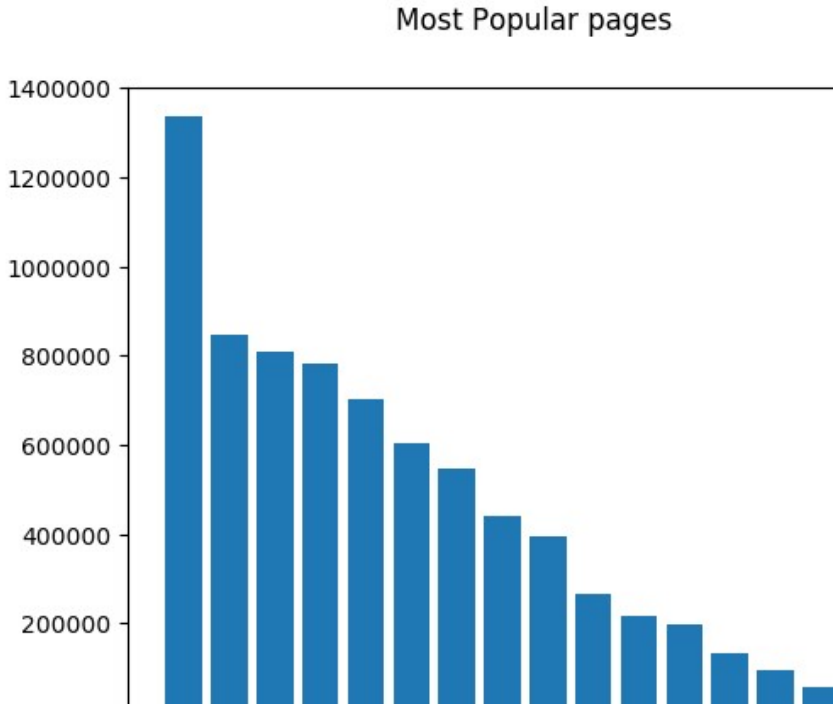
Most Frequent Pattern:



Largest Sequence pattern:



Most Popular Pages



10 Conclusion

Summary
Project Statement: Recognizing Frequent patterns in sequence data.
As we find out most frequent sequence pattern, which shows User's behavior on visit a site and its different sections.
From this analysis, we can find interested site, information flow, also user habbits which can be used for other BI applications

Limitations and future work
Currently, implementation is based on sequence data available for msn.com news site.

- Implementing generic model for any kind sequence pattern recognizing
- Implementation using Real Time data.
- Expand applications for other attributes.
- Minimizing large file processing time.

References

- **http://archive.ics.uci.edu/ml/datasets/MSNBC.com+Anonymous+Web+Data**
- **http://archive.ics.uci.edu/ml/datasets/MSNBC.com+Anonymous+Web+Data**
- **https://docs.python.org/2/library/sets.html**
- **https://docs.python.org/3/tutorial/datastructures.html**

***