

Frequent Pattern Analysis in Web Data

Contents

- Problem Statement
- Dataset Overview
- Data Distribution
- Data Preprocessing
- Implementation
- Results
- Conclusion
- References

Problem Statement

- To predict the next web page that will be visited by the user in a particular website based on the data which has the sequence of page visits of many users to that website previously.

Data Set Overview

Data Type: Discrete sequence

This data describes the page visits of users who visited msnbc.com on a particular day.

% Sequences:

1 1

2

3 2 2 4 2 2 2 3 3

5

1 10 5 10 4 4 4 10 11 11 6 6 8 8

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 10 11 11 12 2 2 2 3

1 1 1 1 1 1 1 1

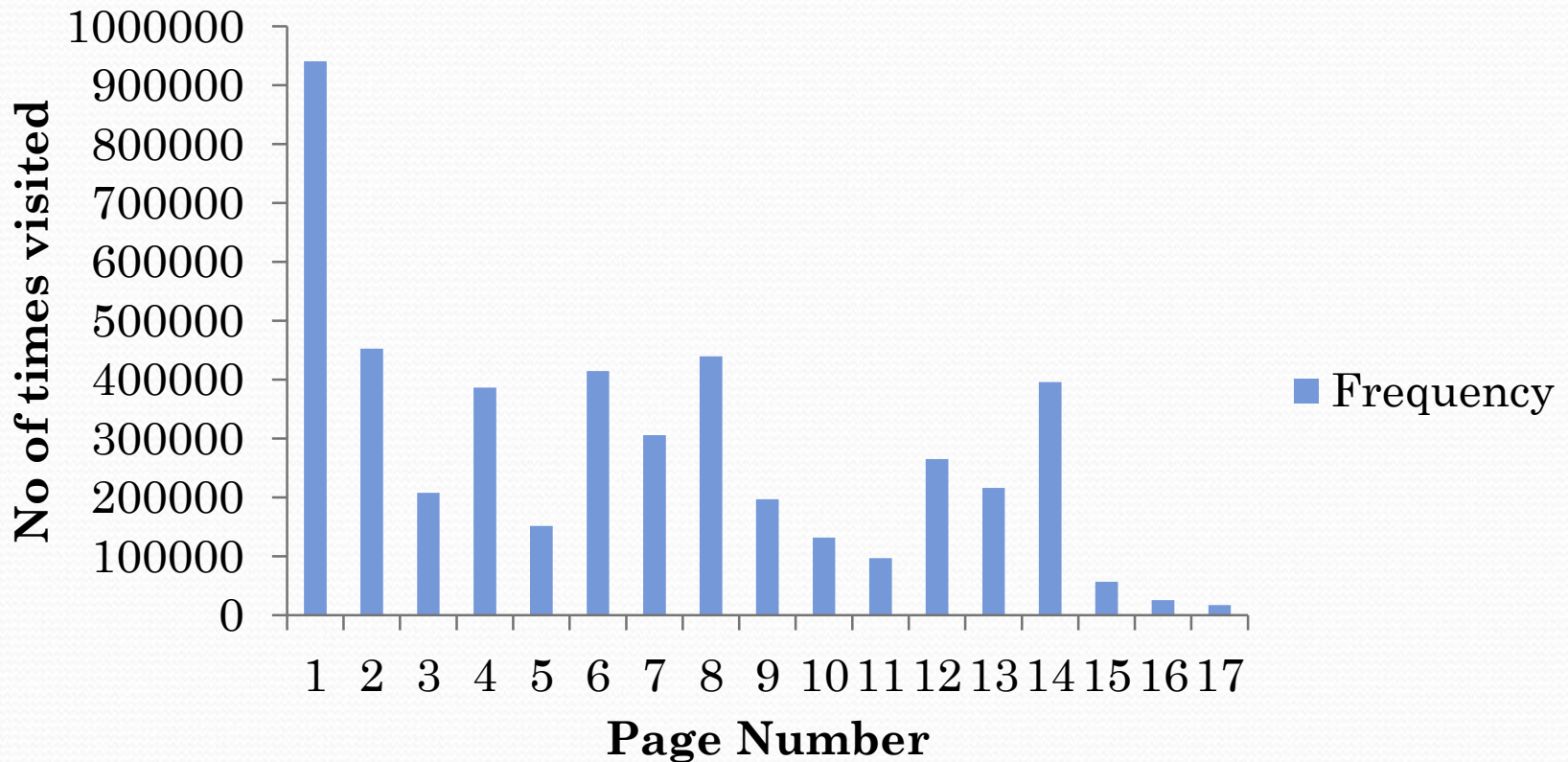
12 12

1 1

Each row below "% Sequences:" describes the hits--in order--of a single user.

Data Distribution

Page Visits



- Average number of visits per user: 5.7
- Number of users: 989818

DATA PREPROCESSING

Transforming the MSNBC dataset into a required format

○Steps involved:

1. Data Cleaning:

Removal of unwanted text content in the data.

2. Data Transformation:

Transformation by introducing parameters such as:

SID → SequenceID

EID → EventID

Transformation:

- Let us consider a sample data represented as

1 1 2
2 3
4 5 1 2 1

Transformation

SID	EID	trans action
1	1	1
1	2	1
1	3	2
2	1	2
2	2	3
3	1	4
3	2	5
3	3	1
3	4	2
3	5	1

On applying the pre processing steps, the above horizontal data is converted to vertical format.

IMPLEMENTATION


- We make use of the association rules and probability matrix for predicting the web pages.

These include the following steps:

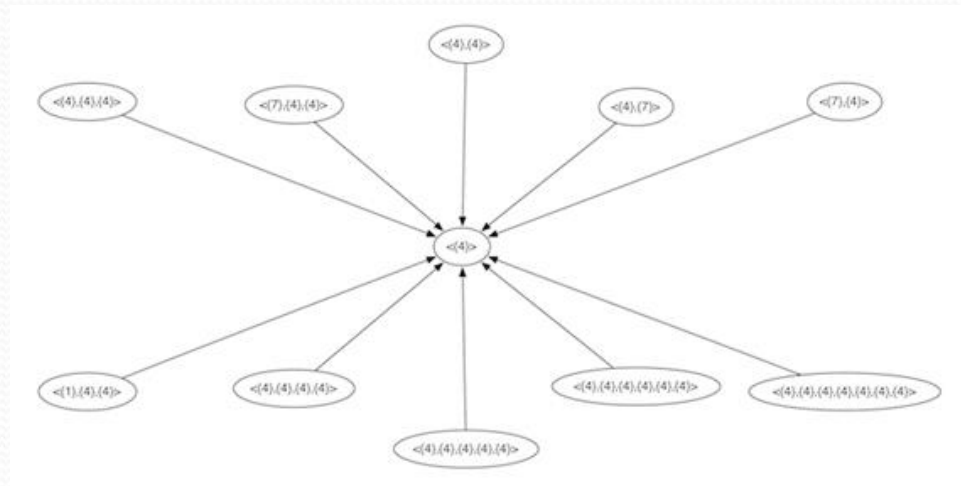
- Mining Frequent Sequences and Association Rules.
- Construct Probability matrix.
- Prediction.

Mining Frequent Sequences and Association Rules:

- In order to mine the frequent sequences we used the SPADE implementation in R called CSPADE.
- First we load the data into a variable and apply the algorithm. The input to the algorithm is the pre processed dataset and the value of minimum support.
- The value of minimum support is chosen rather small to obtain lot of sequences.
- Once the algorithm is applied the output of the algorithm is the frequent sequences and their support values.

- 
- Using the frequent sequences obtained we generated the association rules.
 - We used the traditional apriori based approach for generating rules.
 - For generating the association rules we chose minimum confidence as 0.5.
 - We obtained 269 frequent sequence and 147 association rules.
 - For predicting the result we are not entirely depending on the association rules but we also make use of Probability matrix.

Graphical representation:



Probability matrix:

- For construction of probability matrix we first need to generate a transaction matrix.
- A transaction matrix is square matrix which has a value for all possible combinations in a set.
- Here we count the number of times every page is visited after a particular page. We do this for all the pages.
- Example: For page 1 we calculate the number of times every page is visited after page 1 i.e. 1->1, 1->2, 1->3,1->17
- We calculate the same for every page.

Transaction Matrix:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	94494	73820	23795	27130	8604	30579	28117	8095	3257	18870	27260	41211	1628	35992	5739	1049	5038
2	45476	51078	10842	11173	3019	9541	6320	6565	2450	7708	3492	10062	427	8528	5872	182	696
3	16517	7939	14056	3622	1411	2846	741	1256	4817	2505	2988	4192	727	2716	2676	26	655
4	21565	11305	3630	30608	1488	5017	21496	4028	7293	2556	2490	3228	2218	4245	901	61	264
5	5787	2739	851	800	8255	2228	327	892	1346	535	771	640	164	631	1763	247	75
6	18177	10718	5930	7075	2776	33364	36783	4434	7672	6445	2814	4064	2813	3430	11516	371	202
7	25214	3269	812	33018	303	29210	53179	1021	7399	2074	454	955	7984	2719	930	28	126
8	6241	8173	1477	3351	944	3225	2583	20105	3100	1005	987	1320	1857	2372	302	29	81
9	10065	4204	5304	9981	2807	3786	9541	4718	20120	1090	2046	4049	5071	672	44	87	14
10	11944	7191	2830	1826	835	4608	2030	729	1172	13302	1540	2209	154	1141	2249	34	239
11	19014	3992	1816	2368	1505	2977	643	677	1231	1619	9225	1445	318	1762	1471	21	2933
12	27158	8577	5681	3362	1023	2937	1566	1391	5536	2414	1765	21483	947	5329	1252	68	279
13	2898	191	601	1801	99	1501	8438	2443	3748	106	312	623	18218	18090	16	12	102
14	26361	4829	2557	3353	567	2180	3342	1497	1470	955	1300	2865	9731	29187	828	511	245
15	3514	4953	2408	1043	2487	6575	2037	429	201	2850	2154	1130	65	1355	2936	29	156
16	691	225	32	46	141	287	27	16	19	49	18	50	92	405	45	1149	122
17	3707	784	371	337	187	372	550	144	66	281	1611	278	18	262	145	82	1617

- Once, the transaction matrix is constructed we can use that for constructing probability matrix.
- We calculate the probability of each transaction in transaction matrix.
- The probability of transition from page A to B can be calculated as follows:
- $P(A \rightarrow B) = \text{Count}(A \rightarrow B) / \sum_{i=1 \text{ to } 17} \text{Count}(A \rightarrow i);$
where $A, B \in \{1, 2, \dots, 17\}$

Page No	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0.05548177	0.04333431	0.013371	0.015323	0.00505	0.017354	0.01651	0.004753	0.001312	0.011073	0.016006	0.024137	3.56E-04	0.021133	0.00337	6.16E-04	0.002358
2	0.02670105	0.0239302	0.006366	0.00656	0.00177	0.005602	0.00371	0.003855	0.001433	0.004526	0.00205	0.005308	2.51E-04	0.005007	0.003448	1.07E-04	4.03E-04
3	0.00363783	0.0046614	0.008253	0.002127	8.28E-04	0.001671	4.35E-04	7.37E-04	0.002828	0.001471	0.001754	0.002461	4.27E-04	0.001535	0.001571	1.53E-05	3.85E-04
4	0.0126618	0.0066377	0.002131	0.017371	8.74E-04	0.002346	0.01262	0.002365	0.004282	0.001501	0.001462	0.001835	0.001302	0.002432	5.23E-04	3.58E-05	1.55E-04
5	0.00333781	0.0016082	5.00E-04	4.70E-04	0.00485	0.001308	1.32E-04	5.24E-04	7.30E-04	3.14E-04	4.53E-04	3.76E-04	3.63E-05	3.70E-04	0.001035	1.45E-04	4.40E-05
6	0.01067255	0.006293	0.003482	0.004154	0.00163	0.01953	0.0216	0.002603	0.004505	0.003784	0.001652	0.002386	0.001652	0.002014	0.006762	2.18E-04	1.13E-04
7	0.0148043	0.0013134	4.77E-04	0.013386	1.78E-04	0.017151	0.03122	5.33E-04	0.004344	0.001218	2.67E-04	5.61E-04	0.004688	0.001536	5.46E-04	1.64E-05	7.40E-05
9	0.00530362	0.0024684	0.003114	0.00586	0.00165	0.002223	0.0056	0.00277	0.011813	6.40E-04	0.001201	0.002377	0.002377	3.95E-04	2.58E-05	4.70E-06	8.22E-06
10	0.00701287	0.0042222	0.001662	0.001072	4.30E-04	0.002706	0.00119	4.28E-04	6.88E-04	0.00781	3.04E-04	0.001237	3.04E-05	6.70E-04	0.00132	2.00E-05	1.40E-04
11	0.01116333	0.0023433	0.001066	0.00133	8.84E-04	0.001748	3.78E-04	3.37E-04	7.23E-04	3.51E-04	0.005416	8.48E-04	1.87E-04	0.001035	8.64E-04	1.23E-05	0.001722
12	0.01534571	0.005036	0.003336	0.001374	6.01E-04	0.001724	3.13E-04	8.17E-04	0.00325	0.001417	0.001036	0.012614	5.56E-04	0.003123	7.35E-04	3.33E-05	1.64E-04
13	0.00170155	1.12E-04	3.53E-04	0.001057	5.81E-05	8.81E-04	0.00435	0.001434	0.002201	6.22E-05	1.83E-04	3.66E-04	0.010637	0.010621	3.33E-06	7.05E-06	1.17E-06
14	0.01547775	0.0028353	0.001501	0.001363	3.33E-04	0.00128	0.00136	8.73E-04	8.63E-04	5.61E-04	7.63E-04	0.001682	0.005714	0.017137	4.86E-04	3.00E-04	1.44E-04
15	0.00206323	0.0023081	0.001414	6.12E-04	0.00146	0.00386	0.0012	2.52E-04	1.18E-04	0.001673	0.001265	6.63E-04	3.82E-05	7.36E-04	0.001724	1.70E-05	3.16E-05
16	4.06E-04	1.32E-04	1.88E-05	2.70E-05	8.28E-05	1.63E-04	1.53E-05	3.33E-06	1.12E-05	2.88E-05	1.06E-05	2.34E-05	5.28E-06	2.38E-04	2.64E-05	6.75E-04	1.17E-06
17	0.00217655	4.60E-04	2.18E-04	1.38E-04	1.10E-04	2.18E-04	3.23E-04	8.45E-05	3.88E-05	1.65E-04	3.46E-04	1.63E-04	1.06E-05	1.54E-04	8.51E-05	1.17E-06	3.43E-04

prediction

- We use the obtained association rules and probability matrix for prediction. Implemented this prediction method in Java.
- First, we match the input sequence with antecedent of the association rules to check if there is a match.
- If there is a match then the value of the consequent i.e. the page number is considered as next page.
- There are cases when there is no match. To deal with the no match cases we used probability matrix.
- If there is no match, we find the page with highest probability to be visited after the last page in the input sequence using the calculated probability matrix and give the page with highest probability as next page.

Results

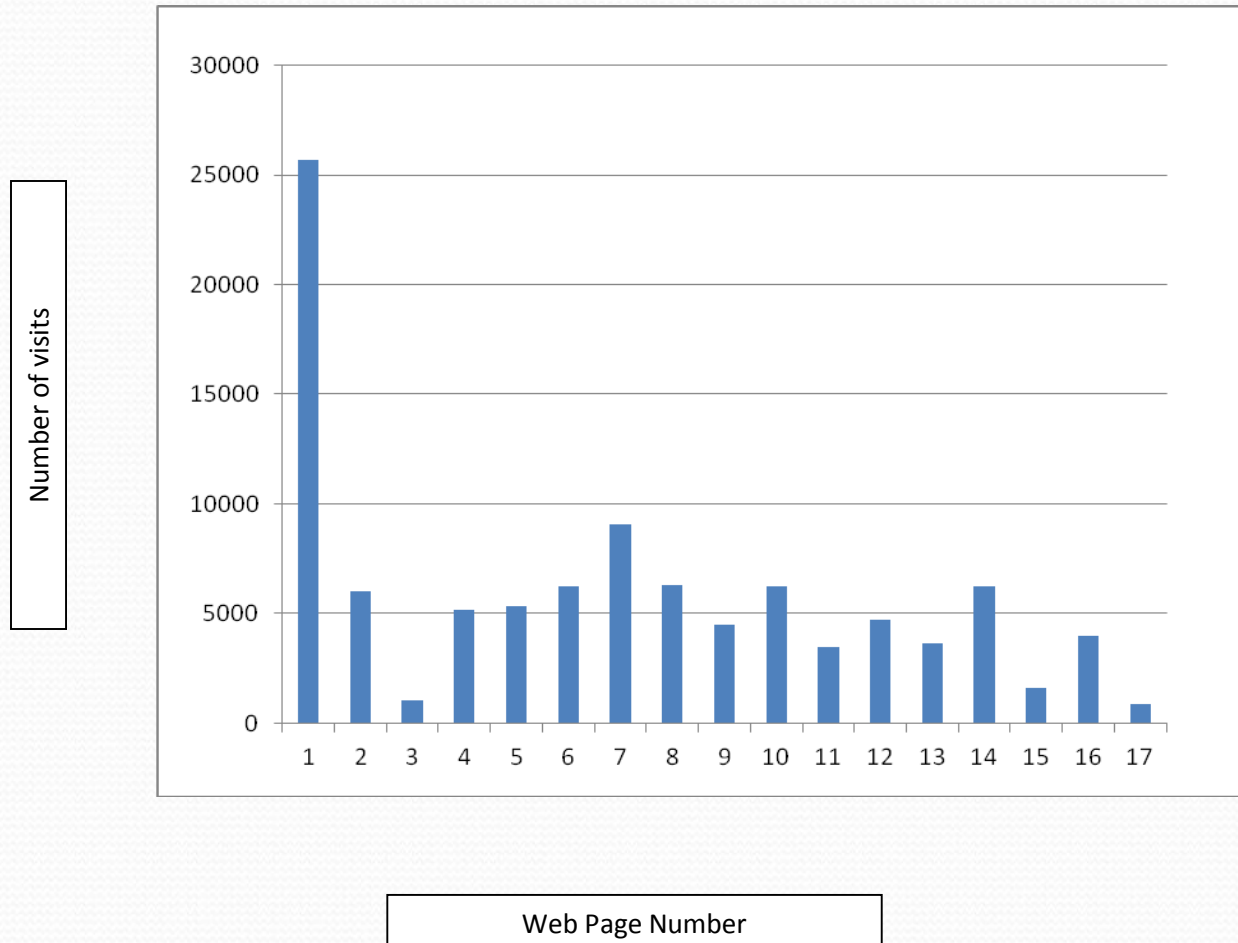
After applying CSPADE on input dataset we obtained frequent sequences as following:

	Sequence	Support
1	$\langle\{1\}\rangle$	0.31640261
2	$\langle\{10\}\rangle$	0.05112657
3	$\langle\{11\}\rangle$	0.05818949
4	$\langle\{12\}\rangle$	0.11333700
5	$\langle\{13\}\rangle$	0.07773954
6	$\langle\{14\}\rangle$	0.12036354
7	$\langle\{15\}\rangle$	0.02950037
8	$\langle\{2\}\rangle$	0.17708912
9	$\langle\{3\}\rangle$	0.12320245
10	$\langle\{4\}\rangle$	0.12297109
11	$\langle\{1\},\{6\}\rangle$	0.03536711
12	$\langle\{7\},\{7\},\{1\}\rangle$	0.02028454
13	$\langle\{7\},\{1\},\{1\}\rangle$	0.01420968
14	$\langle\{1\},\{12\},\{12\}\rangle$	0.01948035

- The association rules obtained from the frequent sequences are as follows:

	rule	support	confidence
1	$\langle\{9\},\{9\}\rangle \Rightarrow \langle\{9\}\rangle$	0.02321639	0.5263881
2	$\langle\{9\},\{9\},\{9\}\rangle \Rightarrow \langle\{9\}\rangle$	0.01316606	0.5671018
3	$\langle\{8\}\rangle \Rightarrow \langle\{8\}\rangle$	0.06783368	0.7022225
4	$\langle\{8\},\{8\}\rangle \Rightarrow \langle\{8\}\rangle$	0.05253996	0.7745409
5	$\langle\{6\},\{6\},\{7\}\rangle \Rightarrow \langle\{7\}\rangle$	0.01258110	0.6922949
6	$\langle\{1\},\{1\},\{7\}\rangle \Rightarrow \langle\{7\}\rangle$	0.01414705	0.7360315
7	$\langle\{6\},\{7\},\{6\}\rangle \Rightarrow \langle\{7\}\rangle$	0.01098081	0.5254786
8	$\langle\{7\},\{6\}\rangle \Rightarrow \langle\{6\}\rangle$	0.01535939	0.6096808
9	$\langle\{6\},\{7\}\rangle \Rightarrow \langle\{6\}\rangle$	0.02089677	0.6885486
10	$\langle\{1\},\{1\},\{6\}\rangle \Rightarrow \langle\{6\}\rangle$	0.01023521	0.5520981
11	$\langle\{5\}\rangle \Rightarrow \langle\{5\}\rangle$	0.01536949	0.6088366
12	$\langle\{5\},\{5\}\rangle \Rightarrow \langle\{5\}\rangle$	0.01137785	0.7402879

- We have tested this implementation of prediction for up to 100000 random input sequences.



Conclusion

- In this project, we have gone through a sequence of steps to predict the next page.
- We have predicted the next page by using the frequent patterns, association rules mined from the input data set and also the probability matrix
- We overcame the problem of no match using the probability matrix.
- The results obtained for large set of test input have a similar pattern to that of input data set.