

An Effective method for Web Log Preprocessing and Page Access Frequency using Web Usage Mining

Jayanti Mehra¹

*Research Scholar, Department of computer Application,
Maulana Azad National Institute of Technology (MANIT), Bhopal,
Madhya Pradesh, India.*

Dr. R S Thakur²

*Assistant Professor, Department of computer Application,
,Maulana Azad National Institute of Technology (MANIT), Bhopal,
Madhya Pradesh, India.*

Abstract

World Wide Web is rising rapidly and enormous amount of information is produced due to user's communications with web sites. To exploit this information, recognizing usage pattern of users is very significant. Web Usage Mining is the application of data mining techniques to find out the useful, hidden information about the users and interesting patterns from data extracted from Web Log files. It supports to know frequently accessed pages, suppose user navigation, progress web site structure etc. In command to relate Web Usage Mining, variety of actions is executed. This paper describes the procedure of Web Usage Mining consisting steps: Data Collection, Pre-processing, Pattern Discovery and Pattern Analysis. It has furthermore existing several approaches for example statistical analysis; clustering, association rules and sequential pattern are individual used to determine patterns in web usage mining. In this paper describes data preprocessing technique of web usage mining, after completion of data preprocessing, any kind of irrelevant information can be sort out. We have also proposed an algorithm and its implementation for web log preprocessing in web usage mining. Every page has been allocated with an individual token. According to this token and frequency, data mining technique (Classification, Association Rules, and Clustering) can be applied. In this article we can simply discover the highest and lowest value according to page access frequency.

Keywords: Web Usage Mining, Web Log Data, Preprocessing, Frequency.

INTRODUCTION

The World Wide Web is a source of web pages that gives the lot of information to the internet users. For internet users the information presented on web has develop into a essential source. for the reason that , there is rising expansion and complication of websites accessible on internet, the amount of

web is huge. A web site is the connection the consumer to company. The companies can revise visitor's performance during web investigation, and discover the patterns. Web mining is generally distinct as find out and study of helpful information commencing the World Wide Web. Web mining split into three parts: Web Contents Mining, Web structure mining and Web Usage Mining. Web Contents Mining represents the taking out of helpful information and web knowledge from web resources or web contents such as text, image, audio, video, and structured records. Web Usage Mining can be as the find and analysis of access patterns of user, throughout the mining of log files. The output of the WUM can be used in web personalization, recovering the system performance, site alteration, usage description etc. Web log file is a server log file which is a fundamental data sources in Web usage mining, in which it include - access logs of the web server. The significant step in the WUM is Data Preprocessing segment. It includes of data cleaning, session identification, user identification, path completion. Data preprocessing is used to clean the inappropriate data from log file so it can be give to the pattern discovery to recognize the user pattern. [1]

Preprocessing is the initial step need after that step we include to execute data mining methods like association rules, classification and clustering for recovering interesting information [4]. The data preprocessing contains data cleaning, user identification, session identification and path completion [3].

Preprocessing is proficient then it can be simply search frequent pattern or interesting rule between web data with limited quantity of time. In this paper we have analyze the web log data and discriminate its attributes. After fetch the data we need to pre-process of that data and provide an algorithm for data preprocessing.

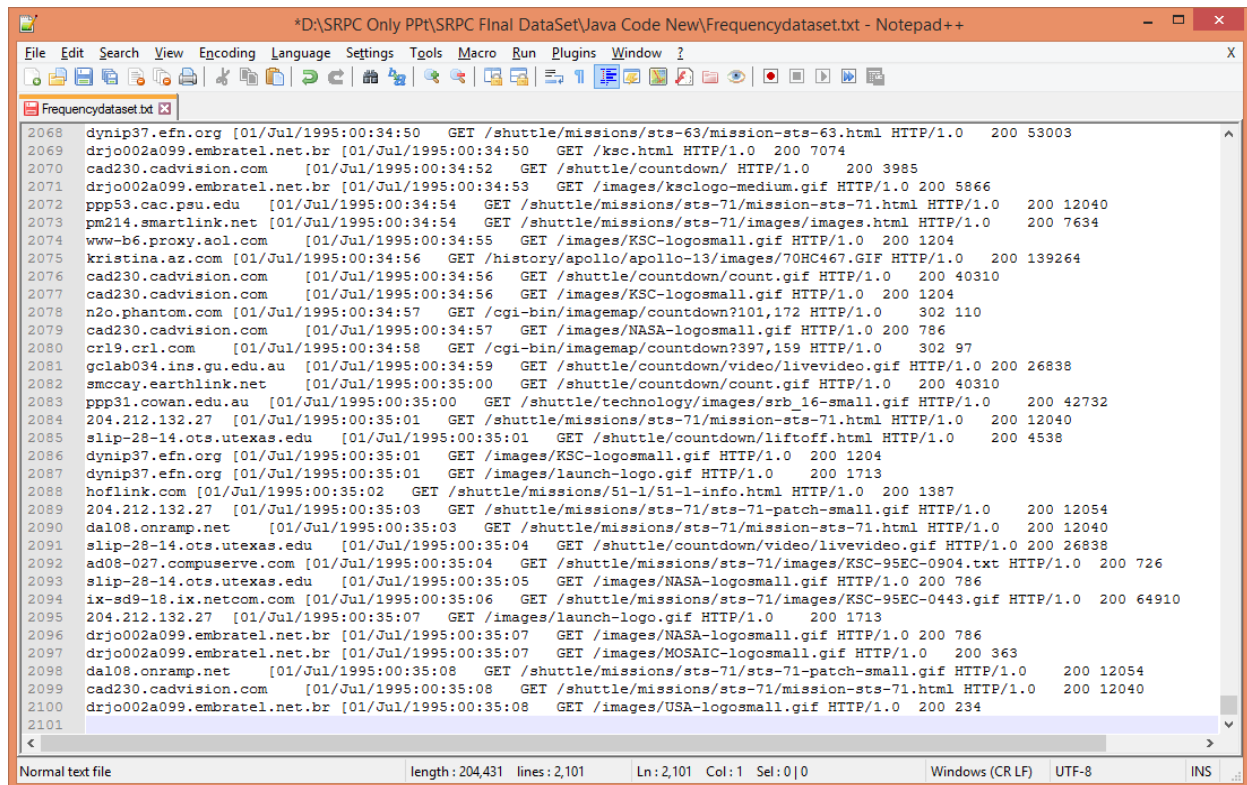


Figure 1. Sample of web log data

A. Data Preprocessing

The data gathered from the web log file is incomplete, noisy and not appropriate for mining initially. Pre-processing is required to exchange the data into relevant form for pattern finding. We start in on Pre-processing stage by data removal then data cleaning and data filtering because the source of web logs data causes are combined with inappropriate information. Data preprocessing acting an main function in Web usage mining. It is used to sift and systematize just suitable information by using web mining algorithms scheduled the web server logs. The innovative server logs are cleaned, formatted, and then grouped addicted to significant sessions earlier than organism used by WUM. This stage holds three sub steps: Data Cleaning, User Identification, and Session Identification [3]

B. Pattern Discovery

Pattern finding illustrates upon methods and algorithms expanded from numerous fields for example statistics, data mining, machine learning and pattern recognition though, it is not the objective of this paper to explain every the obtainable algorithms and techniques received from these fields.[3] This part explains the varieties of mining performance that have been functional to the Web field. Methods build up from other areas ought to obtain into deliberation the diverse types of data generalizations and previous knowledge obtainable for Web Mining. For illustration, in association rule finding, the idea of a deal for market-basket analysis does not get into deliberation the direct in which pieces are chooses. Yet, in

Web Usage Mining, a server session is an efficient series of pages demanded by a user. Also appropriate to the complexity in recognizing unique sessions, extra previous knowledge is compulsory [3][8]

C. Pattern Analysis

Pattern analysis is the final action in the usually Web Usage mining procedure as explained in Figure 1. The reason at the back pattern analysis is to sort out unexciting convention or patterns commencing the set create in the pattern finding stage. The accurate analysis methodology is typically ruled by the function for which Web mining is completed. As a rule common form of pattern analysis consists of a knowledge query method. A further process is to load usage data into a data cube to facilitate perform OLAP operations. Visualization techniques, such as graphing patterns or allocating colours to diverse values, can frequently highlight on the whole patterns or developments in the data. Content and structure information be able of be used to sort out patterns enclosing pages of a usage category. Content type or pages that match a definite hyperlink organization.

RELATED WORK

DATA PREPROCESSING

The information existing in the web is diverse and unstructured. Consequently, the preprocessing segment is a requirement for find out patterns. The objective of

preprocessing is to change the raw click stream data into a set of user profiles. Data preprocessing presents a number of exceptional challenges which led to a diversity of algorithms and heuristic techniques for preprocessing step such as integration and cleaning, user and session identification etc. A variety of research works are approved in this preprocessing part for combination sessions and transactions, which is used to determine user behavior patterns. [5][9]

A. Data Collection

In this article, the data source which is in IIS file format, for the finding hidden information of visitor is collected by NASA-HTTP. The log files .We use the part of the logs during the period of August 1995. For session identification, set the maximum elapsed time to 30 min, which is used in many commercial applications.

The raw data for mining purpose is collected from NASA website. It contains approximately 1727 records in Common log file format. The sample log file used for the task was in raw log format. Size of the file before cleaning was 164 KB with 1727 entries. We can see this in Fig 1.

Data cleaning Log data is stored in database for supplementary processing of data by way of queries and program .Data file acquired was very enormous and it obtains approximately 80% of total time to mine the data. [7] In data cleaning process, the unnecessary information is removed from the log database. The data cleaning obtains the following steps:

Step1: Elimination of the entries having image files, graphic or multimedia files. The records which are accessing file with extension gif,jpg, jpeg etc. are to be removed.[6]After performing this step around 392 records left.

Step 2: The elimination of entries with failed status code. A variety of status codes for HTTP 1.1 in this step the entries having status code of 200 will be retained, rest are removed.

Step3: Removal of records with bytes transferred field zero. The records having entries zero in the byte transferred field specifies that the requested page is not opened, and is to be removed. After performing the above two steps the number of records left are 380. [1]

C. An algorithm for data cleaning

- 1) Start
- 2) Scan the Log Record in log file
- 3) For every record in log file
- 4) Read all the fields (Referrer, Methods, Status etc)
- 5) If status code = Success
- 6) then
- 7) Take IP Address and URL
- 8) If(Suffix of URL=*.txt, *.mpg,*.gif , *.css, *.jpg)

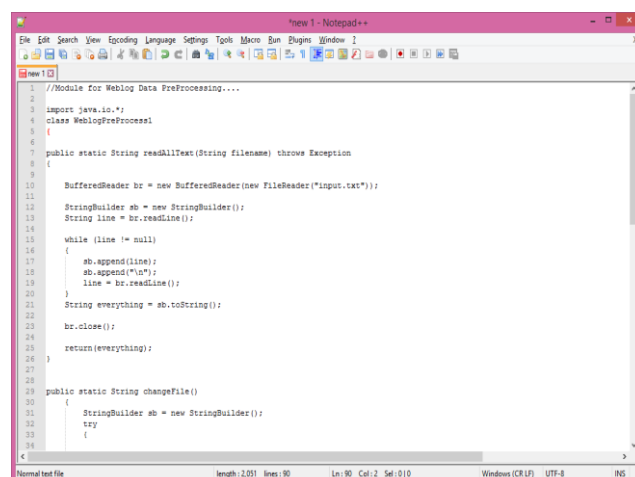
- 9) then
- 10) Remove suffix from URL
- 11) Otherwise
- 12) Save records
- 13) End if
- 14) Fetch the next record
- 15) End if
- 16) Stop

METHODOLOGY

We implemented the algorithm in java programming language. To clean the web log data, read the web log file and calculate all the record. The method is so as to, we read character by character from the file and evaluate the character from ASCII value of space and enter key and count up all the record from web log file. We can see this in the figure 2

Web log File Cleaning: In this action, the irrelevant log entries are deleted from the log file. This can be completed by examination in the request field of the log file, the suffix of the website URL requested by the user. These suffixes notify us the authentic format or extension of the web files requested by user. Contained by the log file, we will receive only those files which have extensions like .html, .asp, .aspx, .php. So we can also delete every log entries taking extensions like .gif, .jpeg, .flv, .mp3, .mp4, etc. We can also delete log entries with empty URL or having request methods other than GET and POST. We can also delete all those log entries with HTTP.[6]

Status code other than 200 .At the end the cleaned log file is organized for the next steps.



```
//Module for Weblog Data PreProcessing....
2
3 import java.io.*;
4 class WebLogPreProcess
5 {
6
7 public static String readAllText(String filename) throws Exception
8 {
9
10 BufferedReader br = new BufferedReader(new FileReader("input.txt"));
11
12 StringBuilder sb = new StringBuilder();
13 String line = br.readLine();
14
15 while (line != null)
16 {
17 sb.append(line);
18 sb.append("\n");
19 line = br.readLine();
20 }
21 String everything = sb.toString();
22
23 br.close();
24
25 return(everything);
26 }
27
28
29 public static String changeFile()
30 {
31 StringBuilder sb = new StringBuilder();
32 try
33 {
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
2221
2222
2223
2224
2225
2226
2227
2228
2229
2230
2231
2232
2233
2234
2235
2236
2237
2238
2239
2240
2241
2242
2243
2244
2245
2246
2247
2248
2249
2250
2251
2252
2253
2254
2255
2256
2257
2258
2259
2260
2261
2262
2263
2264
2265
2266
2267
2268
2269
2270
2271
2272
2273
2274
2275
2276
2277
2278
2279
2280
2281
2282
2283
2284
2285
2286
2287
2288
2289
2290
2291
2292
2293
2294
2295
2296
2297
2298
2299
2300
2301
2302
2303
2304
2305
2306
2307
2308
2309
2310
2311
2312
2313
2314
2315
2316
2317
2318
2319
2320
2321
2322
2323
2324
2325
2326
2327
2328
2329
2330
2331
2332
2333
2334
2335
2336
2337
2338
2339
2340
2341
2342
2343
2344
2345
2346
2347
2348
2349
2350
2351
2352
2353
2354
2355
2356
2357
```

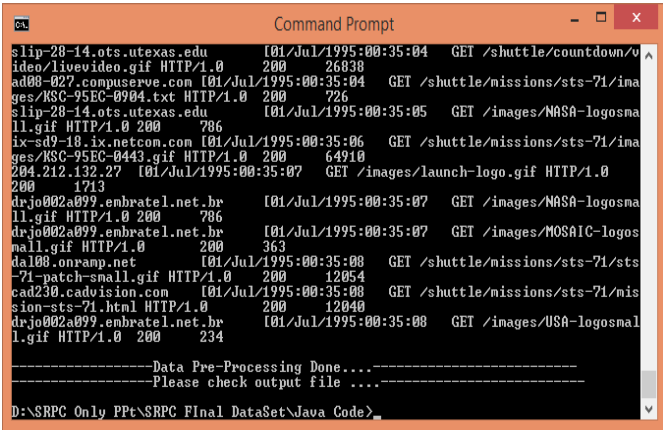


Figure 3. Result after cleaning process

This output already stored in text file we can also store above output in excel sheet below screen shot is output within the excel sheet.

A	B	C	D	E
394	seguim1.irc.mcgil.ac.jp	[01/Jul/1995:00:33:47 GET /shuttle/resources/orbiters/endover.html HTTP/1.0	200	6169
395	qpl6-095.pha.primenet.com	[01/Jul/1995:00:33:47 GET /shuttle/missions/ets-71/mission-ets-71.html HTTP/1.0	200	12040
396	netcom3.netcom.com	[01/Jul/1995:00:33:51 GET /history/apollo/apollo-13.html HTTP/1.0	200	18114
397	larryr.cts.com	[01/Jul/1995:00:33:53 GET /shuttle/missions/ets-70/mission-ets-70.html HTTP/1.0	200	13469
398	whelan.cts.com	[01/Jul/1995:00:33:57 GET /shuttle/technology/eta-reserved/eta_asm.html HTTP/1.0	200	71656
399	zco.phantom.com	[01/Jul/1995:00:34:10 GET /shuttle/resources/orbiters/atlasia.html HTTP/1.0	200	7025
400	remotelp1.us.maine.edu	[01/Jul/1995:00:34:12 GET /shuttle/missions/ets-71/images/images.html HTTP/1.0	200	7634
401	hoflink.com	[01/Jul/1995:00:34:16 GET /shuttle/missions/51-1/mission-51-1.html HTTP/1.0	200	6723
402	cad230.cadvision.com	[01/Jul/1995:00:34:38 GET /shuttle/missions/ets-71/mission-ets-71.html HTTP/1.0	200	12040
403	emcay.earthlink.net	[01/Jul/1995:00:34:41 GET /shuttle/missions/ets-71/images/images.html HTTP/1.0	200	7634
404	139.121.119.19	[01/Jul/1995:00:34:42 GET /shuttle/missions/missions.html HTTP/1.0	200	8677
405	www-b5.prcay.asl.com	[01/Jul/1995:00:34:46 GET /shuttle/missions/ets-71/images/images.html HTTP/1.0	200	7634
406	gplab034.iss.gsl.edu.au	[01/Jul/1995:00:34:47 GET /shuttle/countdown/liftoff.html HTTP/1.0	200	4538
407	ppp31.cowan.edu.au	[01/Jul/1995:00:34:48 GET /shuttle/technology/eta-reserved/etb.html HTTP/1.0	200	49555
408	dynap37.efs.org	[01/Jul/1995:00:34:50 GET /shuttle/missions/ets-63/mission-ets-63.html HTTP/1.0	200	53003
409	drjo002a099.embratel.net.br	[01/Jul/1995:00:34:50 GET /kac.html HTTP/1.0	200	7074
410	ppp53.cac.psu.edu	[01/Jul/1995:00:34:54 GET /shuttle/missions/ets-71/mission-ets-71.html HTTP/1.0	200	12040
411	pm14.smartlink.net	[01/Jul/1995:00:34:54 GET /shuttle/missions/ets-71/images/images.html HTTP/1.0	200	7634
412	204.121.132.27	[01/Jul/1995:00:35:01 GET /shuttle/missions/ets-71/mission-ets-71.html HTTP/1.0	200	12040
413	alip-20-14.ots.utexas.edu	[01/Jul/1995:00:35:01 GET /shuttle/countdown/liftoff.html HTTP/1.0	200	4538
414	hoflink.com	[01/Jul/1995:00:35:02 GET /shuttle/missions/51-1/51-1-info.html HTTP/1.0	200	1387
415	onramp.net	[01/Jul/1995:00:35:03 GET /shuttle/missions/ets-71/mission-ets-71.html HTTP/1.0	200	12040
416	cad230.cadvision.com	[01/Jul/1995:00:35:06 GET /shuttle/missions/ets-71/mission-ets-71.html HTTP/1.0	200	12040

Result after data cleaning in excel sheet

Comparison in Size Before and After Cleaning Size (KB)

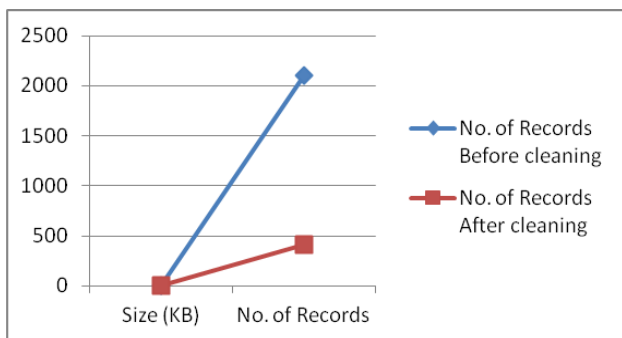


Figure 4. Comparison graph of Data cleaning

Find out the most popular page from web log file

In this section we can find out most popular Page. From this section firstly we can apply data preprocessing technique. Then we counted the frequency of the access page, the term Frequency of page means that the numbers of visibility of that

page in web log file. This method we implement in java language which read the string line by line and checked it with other string. If the string gets matched again and again then we increment its counting by one each time and this counting only shows its frequency. This process is repeated until we reach the end of file.[4]

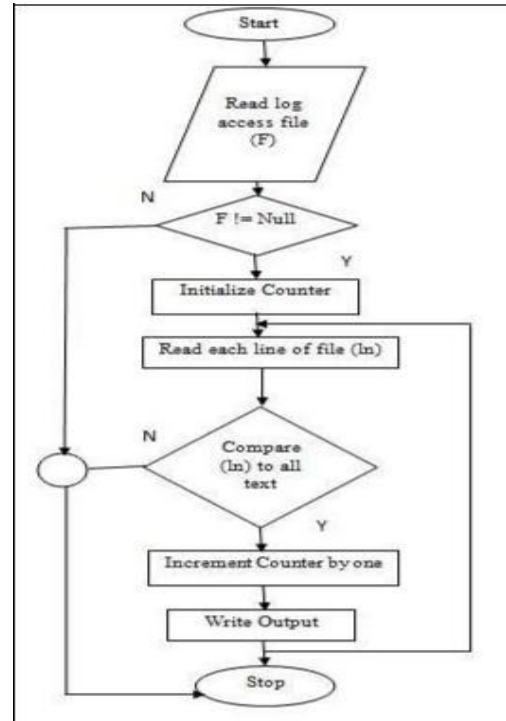


Figure 5. Flow diagram for frequency of page

	Size (KB)	No. of Records
No. of Records Before cleaning	200 KB	2100
No. of Records After cleaning	41 KB	416

```

11 public class Frequency
12 {
13     public static void main(String[] args)
14     {
15         //Creating wordCountMap which holds words as keys and their occurrences as values
16         HashMap<String, Integer> wordCountMap = new HashMap<String, Integer>();
17         BufferedReader reader = null;
18         try
19         {
20             //Creating BufferedReader object
21             reader = new BufferedReader(new FileReader("E:\Data.txt"));
22             //Reading the first line into currentline
23             String currentline = reader.readLine();
24             while (currentline != null)
25             {
26                 //splitting the currentline into words
27                 String[] words = currentline.toLowerCase().split(" ");
28                 //Iterating each word
29                 for (String word : words)
30                 {
31                     //If word is already present in wordCountMap, updating its count
32                     if(wordCountMap.containsKey(word))
33                     {
34                         //Incrementing the count
35                         Integer count = wordCountMap.get(word);
36                         count++;
37                         wordCountMap.put(word, count);
38                     }
39                     else
40                     {
41                         //Adding new word to the map
42                         wordCountMap.put(word, 1);
43                     }
44                 }
45             }
46         }
47         catch (IOException e)
48         {
49             e.printStackTrace();
50         }
51         finally
52         {
53             if (reader != null)
54             {
55                 reader.close();
56             }
57         }
58     }
59 }
    
```

Figure 6. Java code for page access Frequency

In this we can find out most popular page after that

Figure 7. Result for page access frequency

	A	B
1	Page	Frequency
2	/shuttle/missions/sts-71/images/images.html HTTP/1.0	54
3	/shuttle/missions/sts-71/mission-sts-71.html HTTP/1.0	52
4	/shuttle/missions/missions.html HTTP/1.0	25
5	/ksc.html HTTP/1.0	24
6	/shuttle/countdown/liftoff.html HTTP/1.0	22
7	/shuttle/countdown/countdown.html HTTP/1.0	14
8	/shuttle/missions/sts-71/movies/movies.html HTTP/1.0	14
9	/story/apollo/apollo-13/apollo-13.html HTTP/1.0	11
10	/software/winwn/winwn.html HTTP/1.0	11
11	/shuttle/resources/orbiters/atlantis.html HTTP/1.0	9
12	/shuttle/countdown/tour.html HTTP/1.0	8
13	/shuttle/missions/sts-78/mission-sts-78.html HTTP/1.0	8
14	/story/apollo/apollo.html HTTP/1.0	7
15	/history/apollo/apollo-13/apollo-13-info.html HTTP/1.0	6
16	/shuttle/missions/sts-70/mission-sts-70.html HTTP/1.0	6
17	/shuttle/missions/sts-71/sts-71-day-04-highlights.html HTTP/1.0	6
18	/shuttle/technology/sts-newsref/stsref-toc.html HTTP/1.0	6
19	/history/history.html HTTP/1.0	5
20	/shuttle/missions/sts-67/mission-sts-67.html HTTP/1.0	5
21	/facilities/tour.html HTTP/1.0	4
22	/shuttle/countdown/lps/fr.html HTTP/1.0	4
23	/facilities/lc39a.html HTTP/1.0	3

Figure 8. Frequency of an individual page

CONCLUSION AND FUTURE SCOPE

Web data preprocessing is a significant research way of in the field of Web Mining. Web log files are the greatest source to predict user’s behavior. Web log file has useful information and it is also contains entries for unnecessary details like image access, failed entries etc. which are not needed to our mining process. Therefore, it becomes necessary to get divest of this irrelevant information. In this paper, the different phases of data pre-processing have been described. Algorithms for performing the data cleaning technique on server log have also been discussed. The proposed algorithm was successfully tested on the log files for data cleaning. The results which were found after the analysis were acceptable and included important information concerning the log files. The data cleaning approach demonstrated a quite salient reduction in the number of records and in the log files size and therefore enlarges the quality of the available data. Here we also counted the page access frequency and distinct different pages. So that most popular page and least popular page can be find out.

ACKNOWLEDGEMENT

We take opportunity to express our obligation and very thankful to all those who have helped us directly or indirectly to successful completion of this review paper.

REFERENCES

- [1] Surbhi Anand and Rinkle Rani Aggarwal, "An Efficient Algorithm for Data Cleaning of Log File using File Extensions" *International Journal of Computer Applications*, 2012.
- [2] Sheetal A.Raiyani,Shailendra, " Efficient Preprocessing technique using Web log mining," *International Journal of Advancements in Research & Technology*, Volume 1, Issue6,November-2012.
- [3] V.Chitraa and Dr.Antony Selvadoss Thanamani,“A Novel Technique for Sessions Identification in Web Usage Mining Preprocessing”, *International Journal of Computer Applications (0975 – 8887) Volume 34– No.9, November 2011.*
- [4] S. Umamaheswari and S. K. Srivatsa,” Algorithm for Tracing Visitors’ On-Line Behaviors for Effective Web Usage Mining,International Journal of Computer Applications (0975 – 8887) Volume 87 – No.3, February 2014
- [5] Sujith Jayaprakash and Balamurugan E.,” Comprehensive Survey on Data Preprocessing Methods in Web Usage Mining”, *International Journal of Computer Science and Information Technologies*, Vol. 6 (3) , 2015, 3170-3174
- [6] Ketan D. Patel and Satyen M. Parikh,” Preprocessing on Web Server Log Data for Web Usage Pattern Discovery “*International Journal of Computer Applications (0975 – 8887) Volume 165 – No.10, May 2017.*
- [7] Arshi Shamsi, et. All,” Web Usage Mining by Data Preprocessing”, *IJCST Vol. 3, Iss ue 1, Jan. - March 2012.*
- [8] Zhuang Like, Kou Zhongbao and Zhang Changshui, "Session identification based on time intervals in Web log mining," *Journal of Tsinghua University (Science and Technology)*, 2005.
- [9] Brijesh Bakariya and G.S.Thakur, "Preprocessing on Web Log Data in Web Usage Mining," *International Conference on Intelligent Computing and Information System ICICIS*, 2012
- [10] Tasawar Hussain, Dr. Sohail Asghar, Dr. Nayyer Masood, " Web Usage Mining: A Survey on Preprocessing of Web Log File," *IEEE*, 2010.
- [11] T. Murata and K. Saito, "Extracting Users' Interests from Web Log Data," *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings*, 2006.

- [12] R. Cooley, B. Mobasher and J. Srivastava, "Data preparation for mining world wide web browsing patterns," Knowledge and Information System, 1999.
- [13] Thi Thanh Sang Nguyen, Hai Yan Lu and Jie Lu, "Web-page Recommendation based on Web Usage and Domain Knowledge," IEEE, 2013.