In [ ]:

```
import pandas as pd
from collections import Counter
import matplotlib.pyplot as plt
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori, association_rules
```

In [ ]:

```
# 1. 数据预处理
# 读取数据
df = pd.read_csv('./anonymous-msweb.data', skiprows=7, header=None)
```

```
/var/folders/zs/_rd25w5j2ksgfxgty_jstgl00000gn/T/ipykernel_82229/2360474538.
py:2: DtypeWarning: Columns (3,4) have mixed types. Specify dtype option on
import or set low_memory=False.
  df = pd.read_csv('./anonymous-msweb.data', skiprows=7, header=None)
```

In [ ]:

```
# 去除引号和逗号
df = df.replace({'"':'', ',': ''}, regex=True)
```

In [ ]:

```
df.head
```

Out[ ]:

```
<bound method NDFrame.head of          0      1      2                       
3            4
0       A   1287      1          International AutoRoute   /autoroute
1       A   1288      1                          library     /library
2       A   1289      1  Master Chef Product Information  /masterchef
3       A   1297      1                  Central America    /centroam
4       A   1215      1         For Developers Only Info   /developer
...    ..    ...    ...                              ...          ...
131654  V   1035      1                              NaN          NaN
131655  V   1001      1                              NaN          NaN
131656  V   1018      1                              NaN          NaN
131657  C  42711  42711                              NaN          NaN
131658  V   1008      1                              NaN          NaN

[131659 rows x 5 columns]>
```

In [ ]:

```
attribute_lines = df[df[0]=='A'] # 属性行
other_lines = df[df[0]!='A'] # 案例行
other_lines.head
```

Out[ ]:

```
<bound method NDFrame.head of          0      1      2    3    4
294     C  10001  10001  NaN  NaN
295     V   1000      1  NaN  NaN
296     V   1001      1  NaN  NaN
297     V   1002      1  NaN  NaN
298     C  10002  10002  NaN  NaN
...    ..    ...    ...  ...  ...
131654  V   1035      1  NaN  NaN
131655  V   1001      1  NaN  NaN
131656  V   1018      1  NaN  NaN
131657  C  42711  42711  NaN  NaN
131658  V   1008      1  NaN  NaN

[131365 rows x 5 columns]>
```

In [ ]:

```
dic = {x[1]:x[3] for _,x in attribute_lines.iterrows()}
print(dic)
```

```
{1287: 'International AutoRoute', 1288: 'library', 1289: 'Master Chef Produc
t Information', 1297: 'Central America', 1215: 'For Developers Only Info', 1
279: 'Multimedia Golf', 1239: 'Microsoft Consulting', 1282: 'home', 1251: 'R
eference Support', 1121: 'Microsoft Magazine', 1083: 'MS Access Support', 11
45: 'Visual Fox Pro Support', 1276: 'Visual Test Support', 1200: 'Benelux Re
gion', 1259: 'controls', 1155: 'Sidewalk', 1092: 'Visual FoxPro', 1004: 'Mic
rosoft.com Search', 1057: 'MS PowerPoint News', 1140: 'Netherlands (Hollan
d)', 1198: 'Picture It', 1147: 'Microsoft Financial Forum', 1005: 'Norway',
1026: 'Internet Site Construction for Developers', 1119: 'Corporation Inform
ation', 1216: 'Virtual Reality Markup Language', 1218: 'MS Publisher Suppor
t', 1205: 'Hardware Supprt', 1269: 'Customer Guides', 1031: 'MS Office', 100
3: 'Knowledge Base', 1238: 'Excel Development', 1118: 'SQL Server', 1242: 'M
S Garden', 1171: 'MS Merchant', 1175: 'MS Project Support', 1021: 'Visual
C', 1222: 'MS Office News', 1284: 'partner', 1294: 'Bookshelf', 1053: 'Jakar
ta', 1293: 'Encarta', 1167: 'Windows Hardware Testing', 1202: 'Advanced Tech
nology', 1234: 'Office Free Stuff News', 1054: 'Exchange', 1262: 'Chile', 10
74: 'Windows NT Workstation', 1027: 'Internet Development', 1061: 'promo', 1
236: 'Developing for Global Markets', 1212: 'World Wide Offices', 1204: 'MS
Schedule+', 1196: 'ie40', 1188: 'Korea', 1228: 'Visual Test', 1078: 'NT Serv
er Support', 1008: 'Free Downloads', 1052: 'MS Word News', 1091: 'Windows Ha
rdware Development', 1280: 'MS Interactive Music Control', 1247: 'Wine Guid
e', 1064: 'MS Site Builder Workshop', 1065: 'Java Strategy and Info', 1133:
'FrontPage Support', 1102: 'Microsoft Home Essentials', 1132: 'MS Money Supp
ort', 1240: 'Thailand', 1225: 'Anti Piracy Information', 1130: 'IT Technical
Information', 1157: 'Windows 32 bit developer', 1058: 'SP Referral (ART)', 1
076: 'NT Workstation Support', 1163: 'Open Type', 1187: 'ODBC Development',
1152: 'Russia', 1139: 'MS in K-12 Education', 1223: 'Finland', 1001: 'Suppor
t Desktop', 1043: 'Connecting Small Business', 1165: 'Poland', 1194: 'Chin
a', 1138: 'Developer Magazine', 1158: 'Interactive Media Technologies', 109
4: 'Microsoft Home', 1055: 'MSHome Kids Stuff', 1277: 'NetShow for PowerPoin
t', 1143: 'Site Builder Workshop', 1068: 'VBScript Development', 1229: 'Urug
uay', 1177: 'Master Marketing Calendar', 1014: 'Office Free Stuff', 1019: 'M
S PowerPoint', 1122: 'Microsoft User Group Program', 1041: 'Developer Worksh
op', 1033: 'MS Store Logo Merchandise', 1233: 'vbscripts', 1211: 'SMSMGT Sup
port', 1199: 'feedback', 1024: 'Internet Information Server', 1179: 'Colombi
a', 1067: 'FrontPage', 1181: 'Kids Support', 1174: 'New Zealand', 1162: 'IIS
Support', 1046: 'IE Support', 1197: 'SQL Support', 1231: 'Windows NT Develop
er Support', 1141: 'Europe', 1120: 'Switching from Competitive Products', 11
12: 'Canada', 1142: 'South Africa', 1250: 'Middle East', 1214: 'MS Financial
Services', 1190: 'Repository', 1098: 'For Developers Only', 1263: 'Education
al Services & Programs', 1049: 'Support Network Program Information', 1073:
'Taiwan', 1166: 'Mexico', 1226: 'MS Schedule+ Support', 1184: 'MS Excel Supp
ort', 1025: "Web Site Builder's Gallery", 1160: 'Visual C Support', 1156: 'P
owered by BackOffice', 1268: 'javascript', 1220: 'Mac Office Support', 1060:
'MS Word', 1203: 'Denmark', 1176: 'Java Script Development', 1168: 'Sales In
formation (infobase)', 1066: 'Music Producer', 1128: 'MS Solutions Framewor
k', 1275: 'security.', 1136: 'WorldWide Offices - US Districts', 1146: 'Micr
osoft Solution Providers', 1237: 'Developer Days', 1081: 'Access Developmen
t', 1016: 'MS Excel', 1069: 'Windows CE', 1148: 'Channel Resources', 1161:
'Works Support', 1013: 'Visual Basic Support', 1116: 'Switzerland', 1093: 'V
BA Development', 1249: 'Fortran Support', 1095: 'Product Catalog', 1023: 'Sp
ain', 1192: 'Visual J++ Support', 1080: 'Brazil', 1050: 'Macintosh Office',
1255: 'Message Queue Server', 1273: 'mdn', 1206: 'Volume Purchasing Option
s', 1230: 'Mail Support', 1172: 'Belgium', 1011: 'MS Office Development', 10
09: 'Windows Family of OSs', 1096: 'Microsoft Press', 1235: 'MS Training Eva
luation', 1070: 'ActiveX Technology Development', 1154: 'MS Project', 1099:
'Executive Computing', 1186: 'Job Listings for Pre-Grads', 1291: 'news', 125
6: 'Solutions in Action', 1270: 'developr', 1232: 'SiteBuilder Network Specs
& Standards', 1159: 'Transaction Server', 1035: 'Windows95 Support', 1164:
'Systems Management Server', 1077: 'MS Office Support', 1295: 'Training', 10
56: 'sports', 1006: 'misc', 1272: 'softlib', 1123: 'Germany', 1151: 'MS Powe
rPoint Support', 1103: 'MS Works', 1243: 'MS Usability Group', 1244: 'Develo
per Newswire', 1260: 'Exchange Trial', 1258: 'Peru', 1208: 'Israel', 1106:
'Czech Republic', 1124: 'Industry Marketing Information (Vertical)', 1114:
```

```
'Service Advantage', 1012: 'Outlook Development', 1045: 'NetMeeting', 1082:
'MS Access', 1261: "MS's Complete Do It Yourself Guide", 1137: 'About Micros
oft ', 1059: 'Sweden', 1037: 'Windows 95', 1227: 'Argentina', 1281: 'Intelli
Mouse', 1134: 'BackOffice', 1044: 'Developer Media Development', 1028: 'OLE
Development', 1248: 'Softimage ', 1085: 'Exchange Support', 1131: 'MS Money
Information', 1079: 'Australia', 1048: 'MS Publisher', 1042: 'Visual Studi
o', 1075: 'Job Openings', 1201: 'MS Hardware', 1105: 'France', 1153: 'Venezu
ela', 1292: 'MS North Africa', 1015: 'Excel', 1290: 'Activate the Internet C
onference', 1017: 'Products ', 1010: 'Visual Basic', 1126: 'Media Manager',
1144: 'For Developers Only News', 1191: 'Management', 1002: 'End User Produc
ed View', 1213: 'Corporate Customers', 1084: 'UK', 1178: 'msdownload.', 103
6: 'Corporate Desktop Evaluation', 1257: 'Professional Developers Series', 1
180: 'Slovenija', 1246: 'Developer Media Games', 1088: 'OutLook', 1117: 'Sid
ewinder', 1097: 'Latin America Region', 1266: 'Licenses and Piracy', 1072:
'Visual InterDev', 1169: 'MS Project', 1107: 'Slovakia', 1089: 'Office Refer
ence', 1038: 'SiteBuilder Network Membership', 1224: 'Authorized Technical E
ducation Center Program', 1086: 'OEM', 1108: 'MS TeamManager', 1007: 'Intern
ational IE content', 1252: 'Community Affairs', 1283: 'Cinemainia', 1127: 'N
etShow', 1189: 'Internet News', 1110: 'Mastering Series', 1090: 'Games Suppo
rt', 1109: 'TechNet (World Wide Web Edition)', 1040: 'MS Office Info', 1150:
'Internet Information Server News', 1195: 'Portugal', 1111: 'Visual Source S
afe', 1274: 'Professional Developer Conference', 1267: 'Caribbean', 1113: 'I
nternet Security Framework', 1245: 'Open Financial Connectivity', 1253: 'MS
Word Development', 1087: 'MS Proxy Server', 1185: 'SNA Server', 1209: 'Turke
y', 1063: 'Intranet Strategy', 1101: 'Microsoft OLE DB', 1264: 'MS Partner W
eb', 1032: 'Games', 1173: 'Microsoft OnLine Institute', 1051: 'MS Schedule+
News', 1278: 'MS in Higer Education', 1062: 'MS Access News', 1020: 'Develop
er Network', 1104: 'Hong Kong', 1071: 'N. American Automap', 1000: 'regwiz',
1135: 'MS Word Support', 1207: 'Internet Control Pack ', 1217: 'Ireland', 12
54: 'ie3', 1022: 'Typography Site', 1183: 'Italy', 1170: 'Microsoft Mail', 1
241: 'India', 1149: 'Advanced Data Connector', 1029: 'Clip Gallery Live', 12
21: 'Microsoft TV Program Information', 1115: 'Hungary', 1125: 'ImageCompose
r', 1039: 'Internet Service Providers', 1034: 'Internet Explorer', 1265: 'So
urce Safe Support', 1271: 'mdsn', 1129: 'ActiveX Data Objects', 1018: 'isap
i', 1193: 'Office Developer Support', 1219: 'Corporate Advertising Content',
1030: 'Windows NT Server', 1182: 'Fortran', 1100: 'MS in Education', 1210:
'SNA Support'}
```

In [ ]:

```python
cases = []
votes = []
vote = []
case_id = 0
for i, line in other_lines.iterrows():
    if line[0]=='C':
        if len(vote)!=0:
            votes.append(vote)
            cases.append(case_id)
        vote = []
        case_id = line[1]
    else:
        vote.append(dic[line[1]])
votes.append(vote)
cases.append(case_id)
print(len(cases))
print(len(votes))
```

```
32711
32711
```

In [ ]:

```python
print(len(attribute_lines))
# 清洗数据，处理缺失值等
attribute_lines = attribute_lines.dropna()
print(len(attribute_lines))
```

```
294
294
```

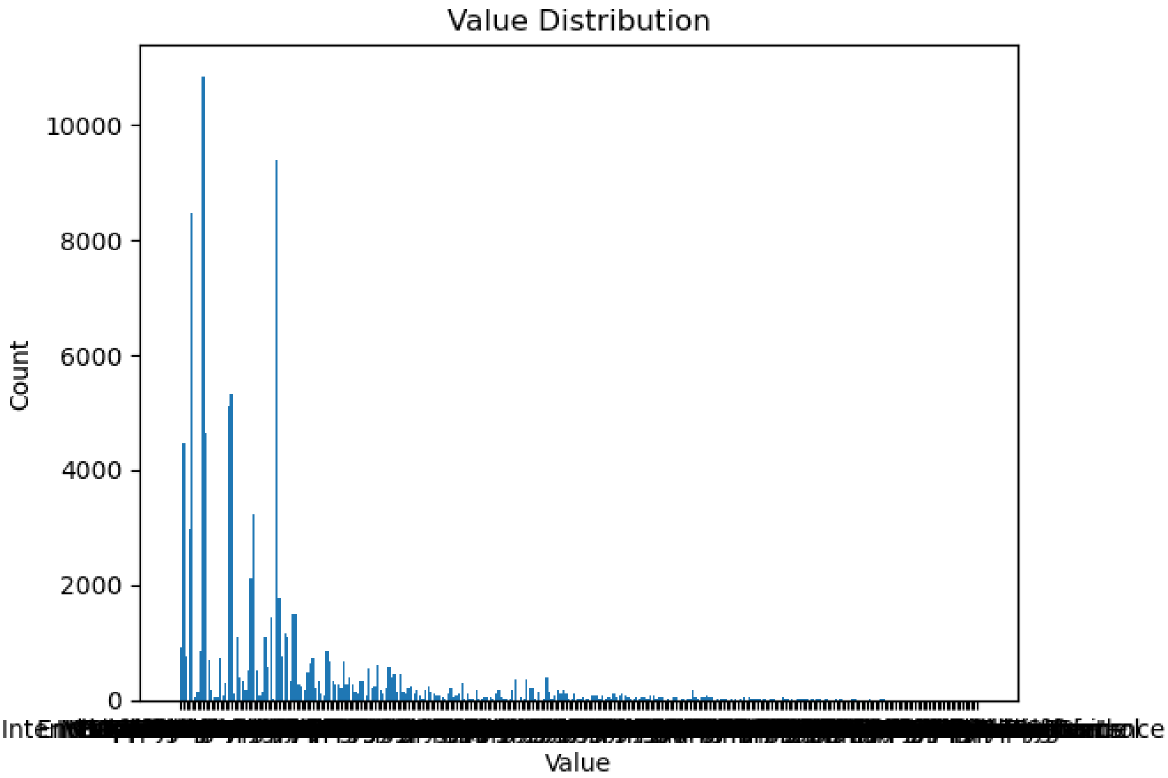
In [ ]:

```python
# 2. 数据探索性分析
# 分析最常被访问的页面

counter = Counter()
# 统计每个子列表中出现的值
for sublist in votes:
    counter.update(sublist)

# 获取出现频率最高的前十个值及其出现次数
most_common_values = counter.most_common(10)

# 打印结果
for value, count in most_common_values:
    print(f"值：{value}，出现次数：{count}")
```

```
值：Free Downloads，出现次数：10836
值：Internet Explorer，出现次数：9383
值：Microsoft.com Search，出现次数：8463
值：isapi，出现次数：5330
值：Products ，出现次数：5108
值：Windows Family of OSs，出现次数：4628
值：Support Desktop，出现次数：4451
值：Internet Site Construction for Developers，出现次数：3220
值：Knowledge Base，出现次数：2968
值：Web Site Builder's Gallery，出现次数：2123
```

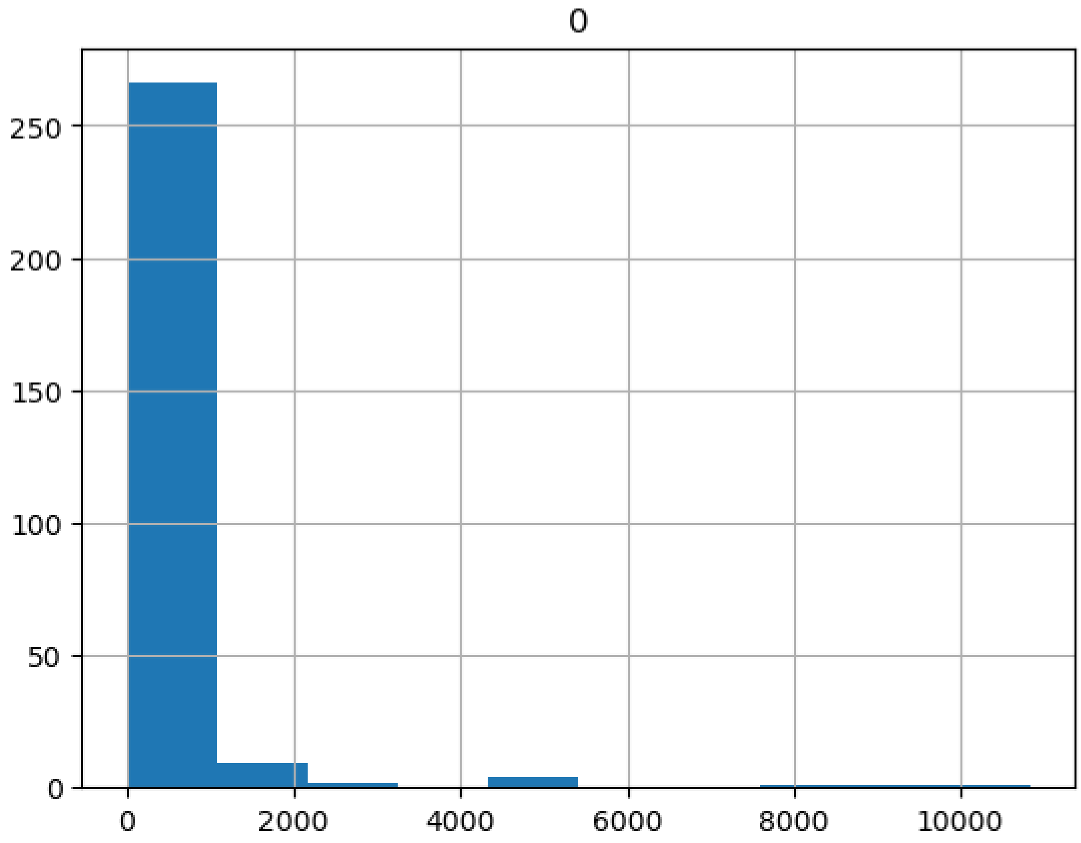
In [ ]:

```python
values = counter.keys()
counts = counter.values()

# 绘制柱状图
plt.bar(values, counts)

# 添加标题和轴标签
plt.title('Value Distribution')
plt.xlabel('Value')
plt.ylabel('Count')

# 显示图形
plt.show()
```



```
In [ ]: counts_df = pd.DataFrame(counts)
        summary = counts_df.describe()

        # 打印统计摘要信息
        print(summary)
        hist = counts_df.hist()
```

```
                  0
count    284.000000
mean     347.373239
std     1173.084272
min        1.000000
25%       10.000000
50%       46.000000
75%      187.500000
max    10836.000000
```



In [ ]:

```
# 3. 关联规则挖掘
# 转换数据格式为TransactionEncoder所需的布尔矩阵形式
te = TransactionEncoder()
te_ary = te.fit_transform(votes)
df_encoded = pd.DataFrame(te_ary, columns=te.columns_)

# 使用Apriori算法计算频繁项集
frequent_itemsets = apriori(df_encoded, min_support=0.07, use_colnames=True)

# 使用关联规则算法计算关联规则
association_results = association_rules(frequent_itemsets, metric="confidenc
```

In [ ]:

```
# 4. 结果评估
# 打印关联规则的支持度、置信度和提升度
print("关联规则: ")
association_results = association_results[['antecedents', 'consequents', 'su
print(association_results)
```

```
关联规则:
                antecedents           consequents   support  confidence  \
0          (Free Downloads)  (Internet Explorer)  0.160802    0.485419   
1       (Internet Explorer)     (Free Downloads)  0.160802    0.560588   
2  (Windows Family of OSs)     (Free Downloads)  0.077925    0.550778   
3                  (isapi)     (Free Downloads)  0.073064    0.448405   

       lift  
0  1.692267  
1  1.692267  
2  1.662652  
3  1.353616  
```

1. 结果分析与应用 根据分析结果提供导航结构优化建议等

针对以上实验结果，我们可以得到强相关规则：

1. Internet Explorer -> Free Downloads
2. Free Downloads -> Internet Explorer
3. Windows Family of OSs -> Free Downloads
4. isapi -> Free Downloads 可以看到Free Downloads的使用度很高，因此可以将其导航设置在更显眼更中心的位置；另外，Free Downloads与Internet Explorer关联度较高，可以将这两个导航设置地更靠近。