The dataset is taken from the UCI Repoistory MSNBC website.

<https://archive.ics.uci.edu/ml/datasets/msnbc.com+anonymous+web+data> – (**I Table Data)**

|  |  |
| --- | --- |
| **Front page** | **1** |
| **News** | **2** |
| **Tech** | **3** |
| **Local** | **4** |
| **Opinon** | **5** |
| **On-air** | **6** |
| **Misc** | **7** |
| **Weather** | **8** |
| **Health** | **9** |
| **Living** | **10** |
| **Business** | **11** |
| **Sports** | **12** |
| **Summary** | **13** |
| **Bbs** | **14** |
| **Travel** | **15** |
| **msn-news** | **16** |
| **msn-sports** | **17** |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** |  |  |  |  |  |  |  |  |  |
| **2** |  |  |  |  |  |  |  |  |  |
| **3** | **2** | **2** | **4** | **2** | **2** | **2** | **3** | **3** |  |
| **5** |  |  |  |  |  |  |  |  |  |
| **1** |  |  |  |  |  |  |  |  |  |
| **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** |
| **1** | **1** | **11** | **1** | **1** | **1** |  |  |  |  |
| **12** | **12** |  |  |  |  |  |  |  |  |
| **1** | **1** |  |  |  |  |  |  |  |  |
| **8** | **8** | **8** | **8** | **8** |  |  |  |  |  |

Each record represent the user’s visited sequence.

**Problem Statement**:

**Predict the most frequent page** visited by the user in a specified website.

***Methodology***

First apply fuzzy clustering to group the clusters of similar pattern and then apply Aprioiri algorithm to predict the next cluster group. Among the cluster group apply Page rank algorithm to identify the most important page which need to be personalized or post advertisement…

I’ve again preprocessed the above dataset like this. **(Dataset – given as an attachment in this mail)**

No. of pages in the website can be taken as columns. Each row indicates the number of times that particular column (Ex: News) visited in their sequence. Otherwise mark as zero.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| FrontPage | News | Tech | Local | Opnion | On-air | Misc | Weather | Health | Living | Business | Sport | Summary |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 5 | 3 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

If we apply the same methodology in this dataset, whether the result will be same?