**MATLAB Apriori implementation**

To implement Aprior in MATLAB, the *FindRules* function by Narine Manukyan was downloaded from [Matlab central](http://www.mathworks.com/matlabcentral/fileexchange/42541-association-rules). *FindRules* takes in the data as well as minimum support, minimum confidence, as well as the number of rules the user wishes it to find. The output consists of data structures holding the rules found as well as each rule’s support and confidence. *FindRules* needs the data records to be in the form of a MATLAB matrix of doubles 0’s and 1’s indicating the presence or absence of each item, so additional code was created to iterate through the data records and line by line add the records in the proper form.

**MATLAB Apriori results**

The MATLAB code duplicated the results of the other algorithms. A minimum confidence threshold was used in pruning rules due to the low support observed for all rules generated. The rules found are given in Table X and match with the rules of the other implementations of Apriori.

In execution, adding the data lines was the step that dominated execution time, while the Apriori mining itself was nearly instantaneous. Creating the Matlab data structure took \_\_\_ while Apriori was completed in only 37 seconds.

Table 1: Top rules found by MATLAB Apriori in decreasing order of confidence

|  |  |
| --- | --- |
| Rule | Support, Confidence) |
| misc,summary -> on-air | (0.47524%, 79.4192%) |
| weather,bbs -> sports | (0.013942%, 76.6667%) |
| msn-sports,summary -> on-air | (0.035057%, 73.2068%) |
| health,bbs -> news | (0.020105%, 72.6277%) |
| summary,bbs -> on-air | (0.017074%, 72.2222%) |
| health,bbs -> on-air | (0.019802%, 71.5328%) |
| living,bbs -> news | (0.018488%, 70.9302%) |
| weather,bbs -> news | (0.012831%, 70.5556%) |
| opinion,travel -> news | (0.066982%, 70.2331%) |
| summary,bbs -> news | (0.016367%, 69.2308%) |
| weather,bbs -> on-air | (0.012325%, 67.7778%) |
| opinion,health -> news | (0.23873%, 66.9595%) |
| business,bbs -> news | (0.018892%, 66.3121%) |
| weather,travel -> living | (0.066578%, 65.966%) |
| tech,bbs -> news | (0.018185%, 65.9341%) |
| tech,bbs -> on-air | (0.018185%, 65.9341%) |
| living,bbs -> on-air | (0.017074%, 65.5039%) |
| health,summary -> on-air | (0.31117%, 64.8148%) |
| weather,summary -> on-air | (0.13184%, 64.7%) |
| tech,travel -> news | (0.17235%, 64.6702%) |

**Python Sequential Analysis**

The foundation of the code used for the sequential mining of the data was taken from the implementation of GSP in [pymining](https://github.com/bartdag/pymining). The changes were made within the Scientific Python Development Environment (Spyder) IDE. Each data record was read as a sequence of elements containing one element each.

Some parts of the Pymining code had to be modified in order to adapt it to the problem of interest. Initially the code used the support count in order to prune candidate rules, but this needed to be changed so that a fraction could be input as the minimum support instead of the integer count of supporting data records. Additionally, reading the data into a python data structure posed a problem since the algorithm was designed to read sequences in the form of a string of characters. For example, the sequence of items ‘a’, ‘b’, and ‘c’ would be in the form of: [‘abc’]. The topics in the data range from 1 to 17, so a sequence of ‘117’ could either be 1 followed by 17 (1 17 in the data), or it could be two ones followed by a 7 (1 1 7). To account for this issue with the data, the categories in the data were matched to a letter code from ‘a’ to ‘q’ internally so the records could be used by the pymining function. The output data structures holding the frequent sequences and their support were decoded and the end to match the categories given in the data and stored in a CSV file for analysis in Excel.

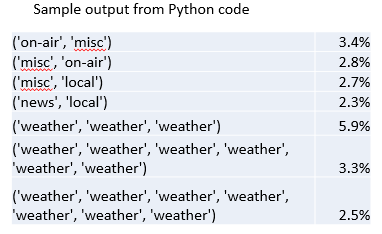
**Python Sequential Results**

The execution of the algorithm completed in 17 seconds, with the sequence mining step accounting for most of the computational load (14 seconds). Even the highest sequences had a low support (<5%) so a minimum support of 1% was used to ensure interesting patterns would be found by the algorithm.

The top rules of interest rules found are given in Table 1 with the complete results of the analysis shown in Table 2 in order of decreasing support. Most of the longer sequences simply represent users browsing within similar topics. For example, for ‘weather’, users were most likely to subsequently browse to other ‘weather’ pages.

Many 2-sequences of interest found indicate a few common associations and are illustrated in the graphs shown in Figure 2. ‘misc’ and ‘on-air’ were particular popular sequences. Additionally, a few longer sequences indicating popular patterns for moving between certain categories, such as the two sequences at the bottom of Figure 2 where users are switching between ‘misc’ and ‘on-air’.

Table 2: Representative rules found through sequential mining of the data set.



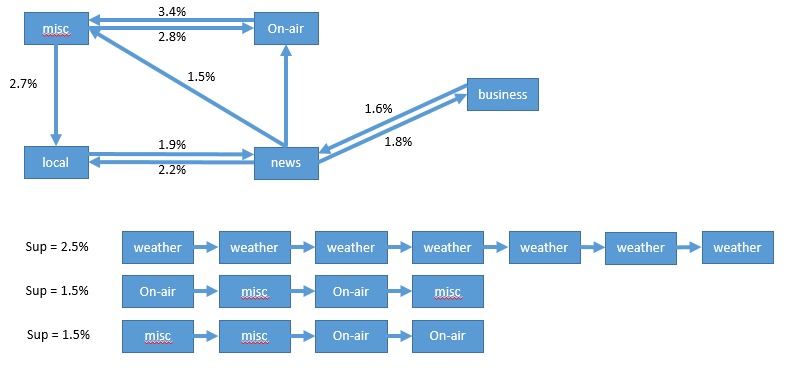


Figure 2: Visualization of the rules found through sequential mining of the data set.

Table 3: Top frequent sequences (Support > 0.01)

|  |  |
| --- | --- |
| RULE | Support (%) |
| ('on-air',) | 24.7% |
| ('news',) | 19.9% |
| ('tech',) | 13.9% |
| ('local',) | 13.8% |
| ('sports',) | 13.5% |
| ('business',) | 12.7% |
| ('weather',) | 10.9% |
| ('msn-news',) | 10.2% |
| ('news', 'news') | 9.9% |
| ('misc',) | 9.1% |
| ('msn-sports',) | 8.7% |
| ('sports', 'sports') | 8.2% |
| ('on-air', 'on-air') | 8.1% |
| ('weather', 'weather') | 7.6% |
| ('misc', 'misc') | 6.7% |
| ('local', 'local') | 6.6% |
| ('living',) | 6.5% |
| ('news', 'news', 'news') | 5.9% |
| ('weather', 'weather', 'weather') | 5.9% |
| ('health',) | 5.8% |
| ('sports', 'sports', 'sports') | 5.6% |
| ('msn-sports', 'msn-sports') | 5.4% |
| ('business', 'business') | 5.3% |
| ('weather', 'weather', 'weather', 'weather') | 5.0% |
| ('msn-news', 'msn-news') | 5.0% |
| ('misc', 'misc', 'misc') | 4.6% |
| ('on-air', 'on-air', 'on-air') | 4.4% |
| ('weather', 'weather', 'weather', 'weather', 'weather') | 4.2% |
| ('tech', 'tech') | 4.1% |
| ('local', 'local', 'local') | 4.1% |
| ('sports', 'sports', 'sports', 'sports') | 3.9% |
| ('news', 'news', 'news', 'news') | 3.9% |
| ('on-air', 'misc') | 3.4% |
| ('msn-sports', 'msn-sports', 'msn-sports') | 3.4% |
| ('summary',) | 3.3% |
| ('weather', 'weather', 'weather', 'weather', 'weather', 'weather') | 3.3% |
| ('misc', 'misc', 'misc', 'misc') | 3.0% |
| ('on-air', 'misc', 'misc') | 3.0% |
| ('business', 'business', 'business') | 2.9% |
| ('opinion',) | 2.8% |
| ('misc', 'on-air') | 2.8% |
| ('sports', 'sports', 'sports', 'sports', 'sports') | 2.8% |
| ('local', 'local', 'local', 'local') | 2.8% |
| ('misc', 'local') | 2.7% |
| ('on-air', 'on-air', 'on-air', 'on-air') | 2.6% |
| ('on-air', 'misc', 'misc', 'misc') | 2.6% |
| ('msn-news', 'msn-news', 'msn-news') | 2.6% |
| ('news', 'news', 'news', 'news', 'news') | 2.6% |
| ('health', 'health') | 2.5% |
| ('misc', 'misc', 'on-air') | 2.5% |
| ('weather', 'weather', 'weather', 'weather', 'weather', 'weather', 'weather') | 2.5% |
| ('on-air', 'misc', 'on-air') | 2.4% |
| ('news', 'local') | 2.3% |
| ('local', 'local', 'local', 'local', 'local') | 2.1% |
| ('msn-sports', 'msn-sports', 'msn-sports', 'msn-sports') | 2.1% |
| ('on-air', 'misc', 'misc', 'on-air') | 2.1% |
| ('sports', 'sports', 'sports', 'sports', 'sports', 'sports') | 2.1% |
| ('on-air', 'on-air', 'misc') | 2.0% |
| ('misc', 'misc', 'misc', 'on-air') | 2.0% |
| ('news', 'tech') | 2.0% |
| ('on-air', 'news') | 2.0% |
| ('living', 'living') | 2.0% |
| ('tech', 'tech', 'tech') | 2.0% |
| ('local', 'misc') | 2.0% |
| ('weather', 'weather', 'weather', 'weather', 'weather', 'weather', 'weather', 'weather') | 1.9% |
| ('local', 'news') | 1.9% |
| ('msn-sports', 'sports') | 1.9% |
| ('news', 'on-air') | 1.9% |
| ('on-air', 'misc', 'misc', 'misc', 'on-air') | 1.8% |
| ('business', 'business', 'business', 'business') | 1.8% |
| ('misc', 'local', 'local') | 1.8% |
| ('news', 'news', 'news', 'news', 'news', 'news') | 1.8% |
| ('news', 'business') | 1.8% |
| ('opinion', 'opinion') | 1.7% |
| ('misc', 'on-air', 'on-air') | 1.7% |
| ('on-air', 'local') | 1.7% |
| ('local', 'local', 'local', 'local', 'local', 'local') | 1.7% |
| ('on-air', 'on-air', 'on-air', 'on-air', 'on-air') | 1.7% |
| ('misc', 'misc', 'local') | 1.7% |
| ('misc', 'misc', 'misc', 'misc', 'misc') | 1.7% |
| ('business', 'news') | 1.7% |
| ('msn-news', 'local') | 1.6% |
| ('on-air', 'misc', 'misc', 'misc', 'misc') | 1.6% |
| ('tech', 'news') | 1.6% |
| ('news', 'sports') | 1.6% |
| ('sports', 'sports', 'sports', 'sports', 'sports', 'sports', 'sports') | 1.6% |
| ('news', 'misc') | 1.6% |
| ('health', 'health', 'health') | 1.5% |
| ('weather', 'weather', 'weather', 'weather', 'weather', 'weather', 'weather', 'weather', 'weather') | 1.5% |
| ('misc', 'misc', 'on-air', 'on-air') | 1.5% |
| ('on-air', 'misc', 'on-air', 'on-air') | 1.5% |
| ('msn-news', 'msn-news', 'msn-news', 'msn-news') | 1.5% |
| ('local', 'misc', 'local') | 1.4% |
| ('on-air', 'on-air', 'misc', 'misc') | 1.4% |
| ('on-air', 'on-air', 'misc', 'on-air') | 1.4% |
| ('misc', 'on-air', 'misc') | 1.4% |
| ('news', 'news', 'local') | 1.4% |
| ('news', 'health') | 1.4% |
| ('on-air', 'summary') | 1.4% |
| ('misc', 'news') | 1.4% |
| ('local', 'local', 'local', 'local', 'local', 'local', 'local') | 1.4% |
| ('msn-sports', 'msn-sports', 'msn-sports', 'msn-sports', 'msn-sports') | 1.4% |
| ('on-air', 'misc', 'misc', 'on-air', 'on-air') | 1.3% |
| ('news', 'news', 'news', 'news', 'news', 'news', 'news') | 1.3% |
| ('living', 'news') | 1.3% |
| ('misc', 'local', 'local', 'local') | 1.3% |
| ('msn-news', 'misc') | 1.3% |
| ('health', 'news') | 1.3% |
| ('opinion', 'opinion', 'opinion') | 1.3% |
| ('msn-sports', 'msn-sports', 'sports') | 1.3% |
| ('misc', 'misc', 'misc', 'misc', 'misc', 'misc') | 1.3% |
| ('business', 'business', 'business', 'business', 'business') | 1.3% |
| ('on-air', 'tech') | 1.3% |
| ('travel',) | 1.3% |
| ('misc', 'misc', 'misc', 'on-air', 'on-air') | 1.2% |
| ('on-air', 'misc', 'on-air', 'misc') | 1.2% |
| ('on-air', 'news', 'news') | 1.2% |
| ('on-air', 'msn-news') | 1.2% |
| ('sports', 'sports', 'sports', 'sports', 'sports', 'sports', 'sports', 'sports') | 1.2% |
| ('misc', 'misc', 'on-air', 'misc') | 1.2% |
| ('weather', 'weather', 'weather', 'weather', 'weather', 'weather', 'weather', 'weather', 'weather', 'weather') | 1.2% |
| ('msn-sports', 'sports', 'sports') | 1.2% |
| ('local', 'on-air') | 1.2% |
| ('local', 'local', 'local', 'local', 'local', 'local', 'local', 'local') | 1.2% |
| ('on-air', 'health') | 1.2% |
| ('sports', 'news') | 1.2% |
| ('news', 'news', 'on-air') | 1.2% |
| ('on-air', 'misc', 'misc', 'misc', 'on-air', 'on-air') | 1.1% |
| ('local', 'news', 'news') | 1.1% |
| ('on-air', 'on-air', 'misc', 'misc', 'misc') | 1.1% |
| ('business', 'tech') | 1.1% |
| ('news', 'local', 'local') | 1.1% |
| ('summary', 'summary') | 1.1% |
| ('misc', 'local', 'misc') | 1.1% |
| ('on-air', 'on-air', 'on-air', 'on-air', 'on-air', 'on-air') | 1.1% |
| ('misc', 'misc', 'local', 'local') | 1.1% |
| ('news', 'living') | 1.1% |
| ('misc', 'misc', 'misc', 'misc', 'on-air') | 1.1% |
| ('news', 'misc', 'misc') | 1.1% |
| ('business', 'local') | 1.1% |
| ('tech', 'tech', 'tech', 'tech') | 1.1% |
| ('on-air', 'misc', 'misc', 'on-air', 'misc') | 1.1% |
| ('news', 'news', 'business') | 1.1% |
| ('news', 'sports', 'sports') | 1.0% |
| ('opinion', 'opinion', 'opinion', 'opinion') | 1.0% |
| ('news', 'news', 'tech') | 1.0% |
| ('weather', 'news') | 1.0% |
| ('local', 'msn-news') | 1.0% |
| ('local', 'misc', 'misc') | 1.0% |
| ('local', 'misc', 'local', 'local') | 1.0% |
| ('health', 'health', 'health', 'health') | 1.0% |
| ('local', 'local', 'local', 'local', 'local', 'local', 'local', 'local', 'local') | 1.0% |
| ('misc', 'misc', 'misc', 'misc', 'misc', 'misc', 'misc') | 1.0% |
| ('msn-news', 'news') | 1.0% |