Mining Frequent Pattern

by Binanto Iwan

Submission date: 19-Apr-2021 02:13PM (UTC+0700)

Submission ID: 1367616810

File name: fpGrowth-ICICEL-B-final.pdf (526.56K)

Word count: 3117

Character count: 16920

ICIC Express Letters
Part B: Applications
Volume **11,** Number **9**, September **2020**

ICIC International ©2020 ISSN 2185-2766
pp. 1-8

Mining Frequent Pattern on Liquid Chromatography-Mass Spectrometer Data of Rodent Tuber to Find The Association Rules of Compounds for Machine Learning

Iwan Binanto[1,2], Harco Leslie Hendric Spits Warnars[1], Nesti Fronika Sianipar [3,4], Widodo Budiharto[1]

[1]Computer Science Department, BINUS Graduate Program, Doctor Computer Science Bina Nusantara University, Jakarta, Indonesia

[2]Informatics Department, Sanata Dharma University, Yogyakarta, Indonesia

[3]Food Technology Department, Faculty of Engineering, Bina Nusantara University, Jakarta, Indonesia

[4]Research Interest Group Biotechnology, Bina Nusantara University, Jakarta, Indonesia

Received March 2021; accepted April 2021

ABSTRACT. *Liquid Chromatography-Mass Spectrometry (LC-MS) data contains a lot of measurement data, which contains rodent tuber plant measurement. This paper aims to obtain frequent patterns on LC-MS data utilizing the FP-Growth_method, where frequent patterns will be used for interpretation and identification of chemical compound in a biological sample. For faster computation, it needed representative sampling data and Linear Systematic Sampling was utilized for this purpose. In conclusion, FP-Growth can be applied successfully on Rodent Tuber's sample LC-MS data and generate an association rule of chemical compound formula. The result showed that C49H79O13P was the dominant compound in the LC-MS data with the highest support value of 41.67%. But it depends on the other, namely [C43H49O23]+. It can be said that when [C43H49O23]+ appears, there is possibility that C49H79O13P will also appear. Also, there is a perfect rule that [C43H49O23]+ implies C49H79O13P but not conversely. These results can be utilized for rule-based machine learning.*

Keywords: LC-MS, Frequent Pattern Growth, Rodent Tuber, Data Mining, Linear Systematic Sampling, Machine Learning

1. Introduction. Raw data of Liquid Chromatography-Mass Spectrometry (LC-MS) contains millions of data points, there are hundreds to thousands of chromatographic peaks, after integration and peak extraction [1]. This raw data provide highly complex biological samples [2]. Meanwhile, LC-MS is widely used, especially in the interpretation or identification of the content of chemical compounds in a biological sample [2]–[6]. It consists of hundreds of thousands of mass per charge (m/z), retention time, and intensity [7]. Likewise, the LC-MS data of Rodent Tuber from the studies of Sianipar et. al [8]–[12] which is used in this paper.

It is not trivial to processing this data to be useful information. Many techniques are used to process these data. It is challenging to extract the information contained in the LC-MS data. One of them is finding certain patterns in the data with Association Rule. This pattern can be used to find and recognize certain chemical compounds contained in other LC-MS data. Further, the association rule can be used in machine learning.

Association Rule is a standardized and well-researched technique for finding interesting relationships between variables [13], finding causalities [14] between variables, finding the frequent patterns [15], [16] in large databases, and feature selection for classification [17].

The Association Rule is obtained by searching the frequent itemset of all frequent itemsets in the database. The discovery of frequent itemset is an important step in obtain Association Rule [18]. Two algorithms are popular in this field, namely Apriori and FP-Growth which are mining frequent patterns from a set of transactions in horizontal data format [19], although utilize only one minimum support threshold [20].

In this paper, FP-Growth was chosen because of its popularity [16], among the best ones to extract association patterns of the frequent item sets [21], and advantages such as more efficient than Apriori and adopting divide and conquer strategy as discussed in [13], [15], [22]–[24].

This paper describes pattern mining on Rodent Tuber's LC-MS data to obtain Association Rules to determine the relationships and/or casualities between chemical compounds. The results of this study will help biological or pharmaceutical scientists to 1) analyze and identify the possible presence of other compounds, 2) finding interesting relationships and/or casualities between compounds, 3) finding the frequent patterns, and 4) can be used for further processing in machine learning.

This paper categorized as follow: section 2 describes the problem statement and preliminaries process for further processing. Section 3 describes the research method. Section 4 describes the result and discussion. Section 5 focuses on conclusions and future work.

2. Problem Statement and Preliminaries. LC-MS data contain a huge number of compounds. It is possible to identify compounds that frequently occurs and their association with other compounds. It can be established as a pattern for other purposes.

There are more than 700,000 records in each of LC-MS data of Rodent Tuber from studies of Sianipar et. al [8]–[12]. Their studies resulted in 10 datasets. We have calculated, to get one compound name manually, will take about 20 seconds at the fastest time. By a simple calculation, if about 700,000 records will take 20 x 700,000 seconds. That means 5 months to complete the labeling for each dataset.

Using large native data like this case would be expensive. Therefore data sampling will be carried out. This LC-MS data of Rodent Tuber is time-series data, so required sampling per period to obtain a representative sample that represents the actual dataset. Linear systematic sampling is chosen because it is simple and can represent the actual dataset.

Algorithm 1: LSS

```
Result: Sample File in type .xlsx
df = pd.read_excel(file)
listDF = df.values.tolist()
pivotDF = pd.pivot_table(df,index=["Retention Time"])
arrRT = pivotDF.index.array.to_numpy()

arrSample = []
arrTempRT = []
arrTempI = []
totalRec = 0
k = len(listDF)/18000
for n in range(len(arrRT)):
    arrTemp=[]
    arrTempInt=[]
    for i in range(len(listDF)):
        if listDF[i][2] == arrRT[n]:
            arrTemp.append(listDF[i][0])
            arrTempInt.append(listDF[i][1])
    indexes = np.arrange(random.randint(0,k),
                                len(arrTemp), step=k)
    for i in indexes:
        arrSample.append(arrTemp[i])
        arrTempRT.append(arrRT[n])
        arrTempI.append(arrTempInt[i])
df1.to_excel('sampleFile.xlsx', index=False)
```

Figure 1. Algorithm of Linear Systematic Sampling

To obtain a systematic sample, given that N is the number of population of elements, while n is the number of samples desired, so if N/n is an integer, then k = N/n; otherwise, let k be the next integer after N/n.

After that, find a random integer R between 1 and k, defining the sample as the unit numbered R, R + k, R + 2k, and so on. The initial unit R selected is called "random start" and k is called "sampling interval" [25], [26].

Based on that, we develop Algorithm 1 as shown in Figure 1 and implemented in Python. It applied to one of ten of the LC-MS dataset of Rodent Tuber which we have and the result will be used in this paper.

3. Research Method. Our research method consists of several stages as shown in Figure 2. Preliminaries above are the first stage of this method as preparation and initial stage by implementing Algorithm 1 as shown in Figure 1 utilizing Linear Systematic Sampling. This preliminary generate a representative sample of the whole dataset. It generates 18,271 records out of 985,924 records.

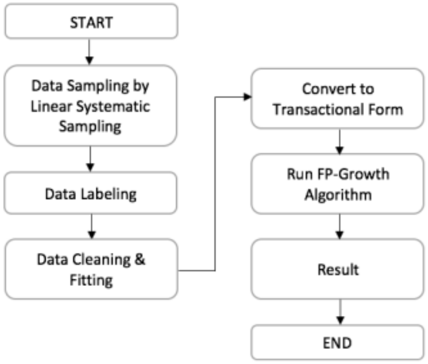
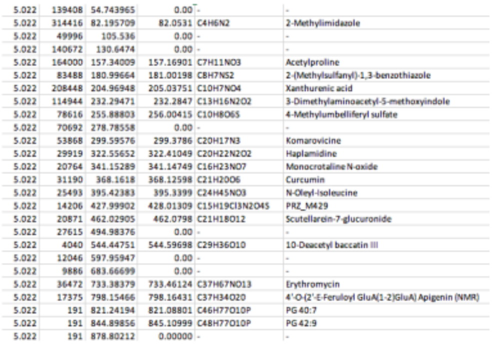


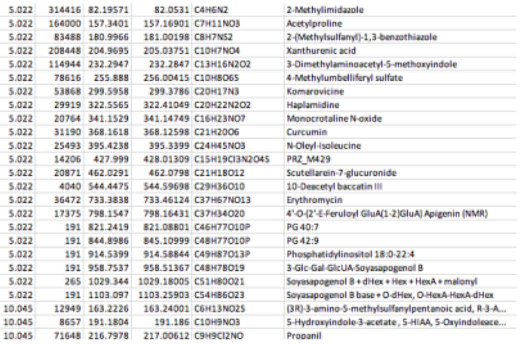
Figure 2. Diagram of research method

FP-Growth is commonly used for market basket analysis which usually has data label. The LCMS data to be analyzed does not have label, so it needs to be labeled. Labeling in this study is giving name of chemical compounds based on their mass per charge (m/z) in the data. After labeling, the LCMS data will be converted into a kind of transactional data like a market basket transaction. It is labeled with another previously developed tool in stage "Data Labeling".



5.022	139408	54.743965	0.00	-	-
5.022	314416	82.195709	82.0531	C4H6N2	2-Methylimidazole
5.022	49996	105.536	0.00	-	-
5.022	140672	130.6474	0.00	-	-
5.022	164000	157.34009	157.16901	C7H11NO3	Acetylproline
5.022	83488	180.99664	181.00198	C8H7NS2	2-(Methylsulfanyl)-1,3-benzothiazole
5.022	208448	204.96948	205.03751	C10H7NO4	Xanthurenic acid
5.022	114944	232.29471	232.2847	C13H16N2O2	3-Dimethylaminoacetyl-5-methoxyindole
5.022	78616	255.88803	256.00415	C10H8O6S	4-Methylumbelliferyl sulfate
5.022	70692	278.78558	0.00	-	-
5.022	53868	299.59576	299.3786	C20H17N3	Komarovicine
5.022	29919	322.55652	322.41049	C20H22N2O2	Haplamidine
5.022	20764	341.15289	341.14749	C16H23NO7	Monocrotaline N-oxide
5.022	31190	368.1618	368.12598	C21H20O6	Curcumin
5.022	25493	395.42383	395.3399	C24H45NO3	N-Oleyl-Isoleucine
5.022	14206	427.99902	428.01309	C15H19Cl3N2O4S	PRZ_M429
5.022	20871	462.02905	462.0798	C21H18O12	Scutellarein-7-glucuronide
5.022	27615	494.98376	0.00	-	-
5.022	4040	544.44751	544.59698	C29H36O10	10-Deacetyl baccatin III
5.022	12046	597.95947	0.00	-	-
5.022	9886	683.66699	0.00	-	-
5.022	36472	733.38379	733.46124	C37H67NO13	Erythromycin
5.022	17375	798.15466	798.16431	C37H34O20	4'-O-(2'-E-Feruloyl GluA(1-2)GluA) Apigenin (NMR)
5.022	191	821.24194	821.08801	C46H77O10P	PG 40:7
5.022	191	844.89856	845.10999	C48H77O10P	PG 42:9
5.022	191	878.80212	0.00000	-	-

Figure 3A. Snippet unclean data

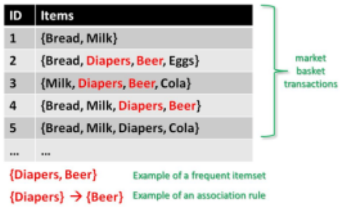


5.022	314416	82.19571	82.0531	C4H6N2	2-Methylimidazole
5.022	164000	157.3401	157.16901	C7H11NO3	Acetylproline
5.022	83488	180.9966	181.00198	C8H7NS2	2-(Methylsulfanyl)-1,3-benzothiazole
5.022	208448	204.9695	205.03751	C10H7NO4	Xanthurenic acid
5.022	114944	232.2947	232.2847	C13H16N2O2	3-Dimethylaminoacetyl-5-methoxyindole
5.022	78616	255.888	256.00415	C10H8O6S	4-Methylumbelliferyl sulfate
5.022	53868	299.5958	299.3786	C20H17N3	Komarovicine
5.022	29919	322.5565	322.41049	C20H22N2O2	Haplamidine
5.022	20764	341.1529	341.14749	C16H23NO7	Monocrotaline N-oxide
5.022	31190	368.1618	368.12598	C21H20O6	Curcumin
5.022	25493	395.4238	395.3399	C24H45NO3	N-Oleyl-Isoleucine
5.022	14206	427.999	428.01309	C15H19Cl3N2O4S	PRZ_M429
5.022	20871	462.0291	462.0798	C21H18O12	Scutellarein-7-glucuronide
5.022	4040	544.4475	544.59698	C29H36O10	10-Deacetyl baccatin III
5.022	36472	733.3838	733.46124	C37H67NO13	Erythromycin
5.022	17375	798.1547	798.16431	C37H34O20	4'-O-(2'-E-Feruloyl GluA(1-2)GluA) Apigenin (NMR)
5.022	191	821.2419	821.08801	C46H77O10P	PG 40:7
5.022	191	844.8986	845.10999	C48H77O10P	PG 42:9
5.022	191	914.5399	914.58844	C49H87O13P	Phosphatidylinositol 18:0-22:4
5.022	191	958.7537	958.51367	C48H78O19	3-Glc-Gal-GlcUA-Soyasapogenol B
5.022	265	1029.344	1029.18005	C51H80O21	Soyasapogenol B + dHex + Hex + HexA + malonyl
5.022	191	1103.097	1103.25903	C54H86O23	Soyasapogenol B base + O-dHex, O-HexA-HexA-dHex
10.045	12949	163.2226	163.24001	C6H13NO2S	(3R)-3-amino-5-methylsulfanylpentanoic acid, R-3-A...
10.045	8657	191.1804	191.186	C10H9NO3	5-Hydroxyindole-3-acetate , 5-HIAA, 5-Oxyindoleace...
10.045	71648	216.7978	217.00612	C9H9Cl2NO	Propanil

Figure 3B. Snippet cleaning and fitting result

The result of labeling is an unclean dataset as shown in Figure 3A, because there are several missing names of chemical compounds, in this case, we write it with characters "-". Therefore, "Data Cleaning and Fitting" were carried out. Data cleaning and fitting is intended to dispose empty data or has no label. It generate dataset as shown in Figure 3B.

After the dataset is clean and fit according to what we want, it needs to be processed again to adjust the use of the FP-Growth algorithm. We called it "Convert to Transactional Form".



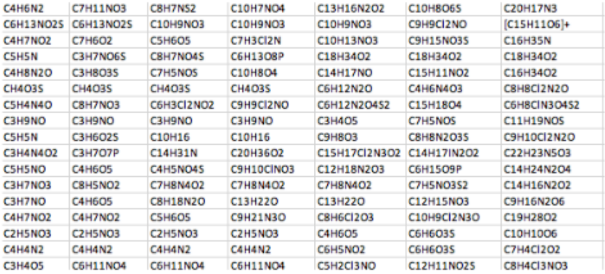
ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
...	...

{Diapers, Beer} Example of a frequent itemset

{Diapers} → {Beer} Example of an association rule

market basket transactions

Figure 4A. Illustration of Market Basket Transactions [27]



C4H6N2	C7H11NO3	C8H7NS2	C10H7NO4	C13H16N2O2	C10H8O6S	C20H17N3
C6H13NO2S	C6H13NO2S	C10H9NO3	C10H9NO3	C10H9NO3	C9H9Cl2NO	[C15H11O6]+
C4H7NO2	C7H6O2	C5H6O5	C7H3Cl2N	C10H13NO3	C9H15NO3S	C16H35N
C5H5N	C3H7NO6S	C8H7NO4S	C6H13O8P	C18H34O2	C18H34O2	C18H34O2
C4H8N2O	C3H8O3S	C7H5NO5	C10H8O4	C14H17NO	C15H11NO2	C16H34O2
CH4O3S	CH4O3S	CH4O3S	CH4O3S	C6H12N2O	C4H6N4O3	C8H8Cl2N2O
C5H4N4O	C8H7NO3	C6H3Cl2NO2	C9H9Cl2NO	C6H12N2O4S2	C15H18O4	C6H8ClN3O4S2
C3H9NO	C3H9NO	C3H9NO	C3H9NO	C3H4O5	C7H5NO5	C11H19NO5
C5H5N	C3H6O2S	C10H16	C10H16	C9H8O3	C8H8N2O3S	C9H10Cl2N2O
C3H4N4O2	C3H7O7P	C14H31N	C20H36O2	C15H17Cl2N3O2	C14H17IN2O2	C22H23N5O3
C5H5NO	C4H6O5	C4H5NO4S	C9H10ClNO3	C12H18N2O3	C6H15O9P	C14H24N2O4
C3H7NO3	C8H5NO2	C7H8N4O2	C7H8N4O2	C7H8N4O2	C7H5NO3S2	C14H16N2O2
C3H7NO	C4H6O5	C8H18N2O	C13H22O	C13H22O	C12H15NO3	C9H16N2O6
C4H7NO2	C4H7NO2	C5H6O5	C9H21N3O	C8H6Cl2O3	C10H9Cl2N3O	C19H28O2
C2H5NO3	C2H5NO3	C2H5NO3	C2H5NO3	C4H6O5	C6H6O3S	C10H10O6
C4H4N2	C4H4N2	C4H4N2	C4H4N2	C6H5NO2	C6H6O3S	C7H4Cl2O2
C3H4O5	C6H11NO4	C6H11NO4	C6H11NO4	C5H2Cl3NO	C12H11NO2S	C8H4Cl3NO3

Figure 4B. Snippet data

This process convert the dataset to the transactional form as a market basket transaction as illustrated in Figure 4A and the snippet of real result dataset of this conversion is shown in Figure 4B. In this case, Transaction ID (TID) is Retention Time and the item set is compound formula. It is done by Algorithm 2 as seen in Figure 5. Retention Time is a TID because it has many duplicates with different m/z , as in actual case, a grocery receipt consisting of many shopping items with one TID which actually it can be interpreted that there are many duplicates of TID because it is attached to one item which is then summarized into one.

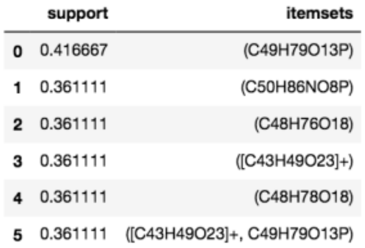
```
Algorithm 2 ConvertToMarketBasketTransaction
Get data from the dataset that has been made into an excel file and convert it to an
actual data list, namely listDF[].
Create a list that contains "Retention Time" column without duplication with the pivot
feature, namely arrRT[].
rows, cols = (len(arrRT), len(listDF))
arr=[]
for i in range(rows):
    row = []
    for j in range(cols):
        if listDF[j][4] == arrRT[i]:      ⇒ check same Retention Time
            row.append(listDF[j][1])  ⇒ append Formula to new list as row
    arr.append(row)                    ⇒ appen list ROW in list ARR
Write arr[] list into an excel file.      ⇒ result is 2D list
```

Figure 5. Algorithm Conversion to Market Basket Transaction

The next step is "Run FP-Growth Algorithm" which is processing the transactional form as a market basket transaction into FP-Growth algorithm. This aims to find the frequent pattern of item. This algorithm is implemented in the python library namely `mlxtend.frequent_patterns` [28]. We just used it without adjustment anything.

4. **Result and Discussion.** Linear Systematic Sampling successfully builds a sample of LC-MS data which at first contains 985,924 records to 18,271 records done by Algorithm 1. Using a smaller number of data sample but representing the entire dataset, labeling work is faster. Labeling was done by web scraping technique [29]–[31]. The next steps are cleaning and fitting, converting, and processed to FP-Growth algorithm. All processes implemented in python programming and fp-growth library.

The results of FP-Growth processing with minimum support of 35% is shown in Figure 6A and the association rule with a minimum threshold of 35% is shown in Figure 6B.

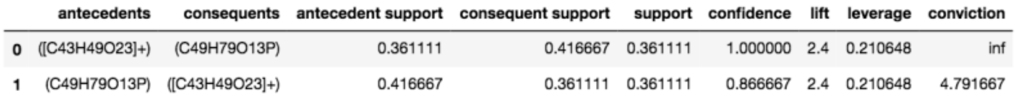


	support	itemsets
0	0.416667	(C49H79O13P)
1	0.361111	(C50H86NO8P)
2	0.361111	(C48H76O18)
3	0.361111	([C43H49O23]+)
4	0.361111	(C48H78O18)
5	0.361111	([C43H49O23]+, C49H79O13P)

Figure 6A. Result of FP-Growth

As seen in Figure 6A, the highest support is 0.416667 and it is a single item, which is C49H79O13P, namely Phosphatidylinositol 20:4-20:4 with exact mass is 906.52582; it means this item appears most frequently (41.6%) in the data. It is dominant than others.

In Figure 6B, the highest confidence value is 1.00, this can be interpreted that if there is [C43H49O23]+, then absolutey C49H79O13P exists, but if conversely, the confidence value is smaller (86.67%) which is still a good value because of more than 50%. This result have an association with Lift value which is more than 1. They are dependent on each other as can be seen in the Leverage value which is more than 0.



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	([C43H49O23]+)	(C49H79O13P)	0.361111	0.416667	0.361111	1.000000	2.4	0.210648	inf
1	(C49H79O13P)	([C43H49O23]+)	0.416667	0.361111	0.361111	0.866667	2.4	0.210648	4.791667

Figure 6B. Result of Association Rule

The conviction will be infinity (∞) if a perfect rule and has a value of 1 if it is a completely uncorrelated rule [32]. [C43H49O23]+ implies C49H79O13P has conviction infinity which means that this is a perfect rule of the antecedent ([C43H49O23]+) to the consequent (C49H79O13P) but not conversely.

5.Conclusion. The FP-Growth algorithm using python library namely `mlxtend.frequent_patterns` can be applied well to the sample of LC-MS data of Rodent Tuber. Sampling was done by the Linear Systematic Sampling method. It serves to simplify a huge dataset but still reflects the actual dataset.

There is a huge number of compounds in LC-MS data of Rodent Tuber when closely observed. However, by processing the dataset with FP-Growth, it can be seen which compounds are more dominant than others. As seen from Figure 6A, the highest support value is 41.67% which is C49H79O13P. It can be interpreted that C49H79O13P is the dominant compound in this LC-MS data. But this compound depends on other compounds, namely [C43H49O23]+ which can be seen in Figure 6B with an infinity conviction value. It can also be said that when [C43H49O23]+ appears, absolutely C49H79O13P will also appear. These association rule can be utilized for processing with rule-based machine learning.

For future works, this study will be applied to other datasets from the existing LC-MS data of Rodent Tuber so that more comprehensive results are obtained.

REFERENCES

[1] P. Kharyuk *et al.*, "Employing fingerprinting of medicinal plants by means of LC-MS and machine learning for species identification task," *Sci. Rep.*, vol. 8, no. 1, pp. 1–12, 2018, doi: 10.1038/s41598-018-35399-z.

[2] M. Brown *et al.*, "Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets," *Bioinformatics*, vol. 27, no. 8, pp. 1108–1112, 2011, doi: 10.1093/bioinformatics/btr079.

[3] M. Gerlich and S. Neumann, "MetFusion: Integration of compound identification strategies," *J. Mass Spectrom.*, vol. 48, no. 3, pp. 291–298, 2013, doi: 10.1002/jms.3123.

[4] C. Guijas *et al.*, "METLIN: A Technology Platform for Identifying Knowns and Unknowns," *Anal. Chem.*, vol. 90, no. 5, pp. 3156–3164, 2018, doi: 10.1021/acs.analchem.7b04424.

[5] B. Zhou, J. F. Xiao, L. Tuli, and H. W. Ressom, "LC-MS-based metabolomics," *Mol. Biosyst.*, vol. 8, no. 2, pp. 470–481, 2012, doi: 10.1039/C1MB05350G.

[6] J. Listgarten and A. Emili, "Statistical and Computational Methods for Comparative Proteomic Profiling Using Liquid Chromatography-Tandem Mass Spectrometry," *Mol. Cell. Proteomics*, vol. 4, no. 4, pp. 419–434, 2005, doi: 10.1074/mcp.R500005-MCP200.

[7] F. Fernández-Albert, "Machine Learning Methods for the Analysis of Liquid Chromatography-Mass Spectrometry datasets in Metabolomics," UNIVERSITAT POLIT` ECNICA DE CATALUNYA, 2014.

[8] D. Laurent, N. F. Sianipar, Chelen, Listiarini, and A. Wantho, "Analysis of Genetic Diversity of Indonesia Rodent Tuber (Typhonium flagelliforme Lodd.) Cultivars Based on RAPD Marker)," in

The 3rd International Conference on Biological Science 2013 (The 3rd ICBS-2013), 2015, vol. 2, pp. 139–145.

[9] N. F. Sianipar, R. Purnamaningsih, and Rosaria, “Bioactive compounds of fourth generation gamma-irradiated Typhoniumflagelliforme Lodd . mutants based on gas chromatography-mass spectrometry,” 2016, doi: 10.1088/1755-1315/41/1/012025.

[10] N. F. Sianipar, R. Purnamaningsih, D. L. Gumanti, Rosaria, and M. Vidianti, “Analysis of Gamma Irradiated-Third Generation Mutants of Rodent Tuber (Typhonium flagelliforme Lodd .) Based on Morphology , RAPD , and GC-MS Markers,” *Pertanika J. Trop. Agric. Sci.*, vol. 40, no. 1, pp. 185–202, 2017.

[11] N. F. Sianipar and R. Purnamaningsih, “Molecular Detection Of Putative Mutant Clones Of Rodent Tuber (Typhonium Flaelliforme Lodd.) CV. Pekalongan Using RAPD Markers,” *Malays. Appl. Biol.*, vol. 47, no. 2, pp. 1–8, 2018.

[12] N. F. Sianipar, K. Assidqi, R. Purnamaningsih, and T. Herlina, “in Vitro Cytotoxic Activity of Rodent Tuber Mutant Plant (Typhonium Flagelliforme Lodd.) Against To Mcf-7 Breast Cancer Cell Line,” *Asian J. Pharm. Clin. Res.*, vol. 12, no. 3, pp. 185–189, 2019, doi: 10.22159/ajpcr.2019.v12i3.29651.

[13] M. S.Mythili and A. R. Mohamed Shanavas, “Performance Evaluation of Apriori and FP-Growth Algorithms,” *Int. J. Comput. Appl.*, vol. 79, no. 10, pp. 34–37, 2013, doi: 10.5120/13779-1650.

[14] S. Lee, Y. Cha, S. Han, and C. Hyun, “Application of association rule mining and social network analysis for understanding causality of construction defects,” *Sustain.*, vol. 11, no. 3, 2019, doi: 10.3390/su11030618.

[15] M. M. Kavitha and S. T. Tamil Selvi, “Comparative Study on Apriori Algorithm and Fp Growth Algorithm with Pros and Cons,” *Int. J. Comput. Sci. Trends Technol. (IJCS T)*, vol. 4, no. 4, pp. 161–164, 2013.

[16] S. Nasreen, M. A. Azam, K. Shehzad, U. Naeem, and M. A. Ghazanfar, “Frequent pattern mining algorithms for finding associated frequent patterns for data streams: A survey,” *Procedia Comput. Sci.*, vol. 37, pp. 109–116, 2014, doi: 10.1016/j.procs.2014.08.019.

[17] C. Huang *et al.*, “Sample imbalance disease classification model based on association rule feature selection,” *Pattern Recognit. Lett.*, vol. 133, pp. 280–286, 2020, doi: 10.1016/j.patrec.2020.03.016.

[18] S. Bagui, K. Devulapalli, and J. Coffey, “A heuristic approach for load balancing the FP-growth algorithm on MapReduce,” *Array*, vol. 7, no. February 2019, p. 100035, 2020, doi: 10.1016/j.array.2020.100035.

[19] A. Mokkadem, M. Pelletier, and L. Raimbault, “Recursive association rule mining,” *arXiv*, pp. 1–27, 2020.

[20] B. Huynh, C. Trinh, V. Dang, and B. Vo, “A parallel method for mining frequent patterns with multiple minimum support thresholds,” *Int. J. Innov. Comput. Inf. Control*, vol. 15, no. 2, pp. 479–488, 2019, doi: 10.24507/ijicic.15.02.479.

[21] S. Jamshidi Nejad, F. Ahmadi-Abkenari, and P. Bayat, “A Combination of Frequent Pattern Mining and Graph Traversal Approaches for Aspect Elicitation in Customer Reviews,” *IEEE Access*, vol. 8, pp. 151908–151925, 2020, doi: 10.1109/ACCESS.2020.3017486.

[22] F. Xu and H. Lu, “The application of FP-growth algorithm based on distributed intelligence in wisdom medical treatment,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 31, no. 4, pp. 1–11, 2017, doi: 10.1142/S0218001417590054.

[23] P. Shendge and T. Gupta, “Comparitive Study of Apriori & FP Growth Algorithms,” *Indian J. Res.*, no. March, pp. 20–22, 2013.

[24] M. Narvekar and S. F. Syed, “An optimized algorithm for association rule mining using FP tree,” *Procedia Comput. Sci.*, vol. 45, no. C, pp. 101–110, 2015, doi: 10.1016/j.procs.2015.03.097.

[25] R. Arnab, “Chapter 4 - Systematic Sampling,” in *Survey Sampling Theory and Applications*, R. Arnab, Ed. Academic Press, 2017, pp. 89–115.

[26] S. L. Lohr, *Sampling: Design and Analysis*, Second Edi. Boston: Brooks/Cole, Cengage Learning, 2010.

[27] “Market Basket Analysis using R - DataCamp.” https://www.datacamp.com/community/tutorials/market-basket-analysis-r (accessed Nov. 27, 2020).

[28] “Mlxtend.frequent patterns - mlxtend.” http://rasbt.github.io/mlxtend/api_subpackages/mlxtend.frequent_patterns/ (accessed Nov. 27, 2020).

[29] M. Herrmann and L. Hoyden, “Applied Webscraping in Market Research,” in *First International Conference on Advanced Research Methods and Analytics*, 2016, p. 2016, doi: 10.4995/carma2016.2016.3131.

[30] M. Shreesha, S. B. Srikara, and R. Manjesh, “A Novel Approach for News Extraction Using Webscraping Technique,” in *The 3rd National Conference on Image Processing, Computing, Communication, Networking and Data Analytics (NCICCNDA 2018)*, 2018, no. NCICCNDA 2018, pp. 359–362, doi: 10.21467/proceedings.1.56.

[31] R. McAlister, “Webscraping as an investigation tool to identify potential human trafficking operations in Romania,” *Proc. 2015 ACM Web Sci. Conf.*, no. 2013, 2015, doi: 10.1145/2786451.2786510.

[32] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, “Dynamic Itemset Counting and Implication Rules for Market Basket Data,” *SIGMOD Rec. (ACM Spec. Interes. Gr. Manag. Data)*, vol. 26, no. 2, pp. 255–264, 1997, doi: 10.1145/253262.253325.

Mining Frequent Pattern

ORIGINALITY REPORT

10%	7%	7%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	eprints.binus.ac.id Internet Source	2%
2	silo.pub Internet Source	2%
3	Pavel Kharyuk, Dmitry Nazarenko, Ivan Oseledets, Igor Rodin, Oleg Shpigun, Andrey Tsitsilin, Mikhail Lavrentyev. "Employing fingerprinting of medicinal plants by means of LC-MS and machine learning for species identification task", Scientific Reports, 2018 Publication	1%
4	umpir.ump.edu.my Internet Source	1%
5	Submitted to Al Zaytoonah University Student Paper	1%
6	Sepideh Jamshidi Nejad, Fatemeh Ahmadi-Abkenari, Pyman Bayat. "A Combination of Frequent Pattern Mining and Graph Traversal Approaches for Aspect Elicitation in Customer Reviews", IEEE Access, 2020 Publication	1%

7	www.ijcttjournal.org Internet Source	1 %
8	Raghunath Arnab. "Systematic Sampling", Elsevier BV, 2017 Publication	1 %
9	Sanjaya, Christine, May liana, and Agus Widodo. "Revenue Prediction Using Artificial Neural Network", 2010 Second International Conference on Advances in Computing Control and Telecommunication Technologies, 2010. Publication	<1 %
10	"Biological Knowledge Discovery Handbook", Wiley, 2013 Publication	<1 %
11	Yu, Tianwei, Youngja Park, Shuzhao Li, and Dean P. Jones. "Hybrid Feature Detection and Information Accumulation Using High-Resolution LC–MS Metabolomics Data", Journal of Proteome Research, 2013. Publication	<1 %
12	docplayer.net Internet Source	<1 %
13	www.researchgate.net Internet Source	<1 %
14	www.duo.uio.no Internet Source	

<1%

15	www.omicsonline.org Internet Source	<1%

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		