Apriori Algorithm for Frequent Pattern Mining for Public Librariesin United States

Muhammad Muhajir, Ayundyah Kesumawati, Satibi Mulyadi
Departement of Statistics, Faculty of Mathematics and Natural Science, Islamic University of Indonesia

Keywords: Data Mining, Apriori, Frequent Pattern Mining, Map.

Abstract: This paper shows how the different approaches achieve the objective of frequent mining along with the complexities required to perform the job and also give the descriptive by map representation about the number of public library in United States. In addition, we demonstrate the use of Big ML tool for association rule mining using Apriori algorithm. There are three attributes of public libraries is used among other City, Location type, and State. The result shows that there are six most recommended rules with confidence value ≥ 0.8 that are, Rules {City includes Chicago} => {Location type=Branch library}; Rules {City includes Brooklyn} => {Location type=Branch library}; Rules {City includes Losangeles} => {Location type=Branch library}; Rules {City includes Baltiomore} => {Location type=Branch library}; Rules {City includes Sandiego} => {Location type=Branch library} and Rules {City includes Miami} => {Location type=Branch library}. The six rules means that the most visited public libraries by population of legal service area over one million people is a branch library.

1 INTRODUCTION

One of service that could be accessed by the general public is public library, where its management is done by the librarian and professional library. There are some of the characteristics of public libraries among others: library funded by the community through taxes, the public can access library collection, library service provided to the public free of charge. The main objective of public libraries is prepared to meet the needs of individuals and group in education, provide information resources and services to the public (Rubin, 2010).

The development of information currently provides a huge impact for librarian. The librarian are required to follow every the development of information that was going on , so information about who gained may applied in management of libraries. librarian is has a lot of digging pieces of data derived from human activity from various sources. The data has a very large and widely in terms of quantity , variatif and acceleration. The term massive data can be called as *big data* that changed the way human understanding the world that had a huge impact and will keep creating ripples through all aspect of human life (Nath, 2015)

The united states was one of the countries establish technology the big data in management of systems public libraries. However, there are the problems faced by public libraries there which is approximately 80 % out of a total of public libraries available data could only present serving 25,000 a person from year to year, while the rest would belong serve around 1,000,000 a person from year to year (ALA Library, 2014).

Increase in the information or data , we need to proper management that public library have its existence as providers information that updates and easily accessible. Hence , the application of the big data concept in public libraries should be introduced that data continues to grow library can be set systematically. Data mining is technology that is very useful for assist public library could find important information from the data the warehouse. One of the methods to find knowledge in a relational large database is association rules. (Agrawal et al. 1993).

Here we have discussed the mechanism for Big ML to use Apriori algorithm in Public Libraries United States. Due to some problems that mentioned before, the problems which will be discussed is to mapping the type of libraries that most visited in United States. Before do the mapping, the libraries

Muhajir, M., Kesumawati, A. and Mulyadi, S.
Apriori Algorithm for Frequent Pattern Mining for Public Librariesin United States.
DOI: 10.5220/0008517200600064
In *Proceedings of the International Conference on Mathematics and Islam (ICMIs 2018)*, pages 60-64
ISBN: 978-989-758-407-7

in United States will be clustered in order to ease the process of mapping. The expected benefit of writing this paper is to know the types of libraries that most visited in United States.

2 METHOD

2.1 Association Rule

Association rule mining is a technique data mining to find rules associative between a combination items. One of the stages in association analysis that attracts attention many researchers to produce the efficient algorithms is frequent pattern mining (Jiawei, and Micheline, 2006). Support is presentation combination the items in a database, where if have items X and items Y so support is the proportion of transactions in a database that are containing X and Y (Han, and Kamber, 2006).

Definition 1. The value of an item support calculated as (Srikant, 1996):

$$Support\ (X => Y) = P(X \cup Y).$$

Confidence is the size of the accuracy of a rule, the proportion of transactions in a database containing containing X and Y. Confidence can also measure the strength of the relationship between the items in association rule

Definition 2. The value confidence than two items are as follows (Srikant, 1996):

$$Confidence\ (X => Y) = P(Y/X) = \frac{P(X \cup Y)}{P(X)},$$

where P(Y/X): conditional probability of occurrence of Y when X events occurred, $P(X \cup Y)$: probability of occurrence of X and Y simultaneously, $P(X)$: probability of occurrence X.

The lift ratio is a measure to analyze the strength of association rules formed. The value of lift ratio usually used as a determining whether the association rules valid or invalid.

Definition 3. To calculate the lift ratio, we used the formula as follows (Srikant, 1996):

$$Lift\ rasio\ (X => Y) = \frac{P(X \cup Y)}{P(X)P(Y)}$$

$$= \frac{Confidence\ (X=>Y)}{P(Y)}$$

$$= \frac{Confidence\ (X=>Y)}{Support(Y)},$$

Association rule mining can be broken down into the following two subproblems among others first,generating all itemsets that have support greater than, or equal to, user specified minimum support; second, generating all rules that have minimum confidence (Brin et al., 1997).

2.2 The Apriori Algorithm

Apriori is popular algorithm and much used for identification mining frequent itemsets. Apriori algorithm steps described in the flow diagram the following (Agarwal et al., 1993):

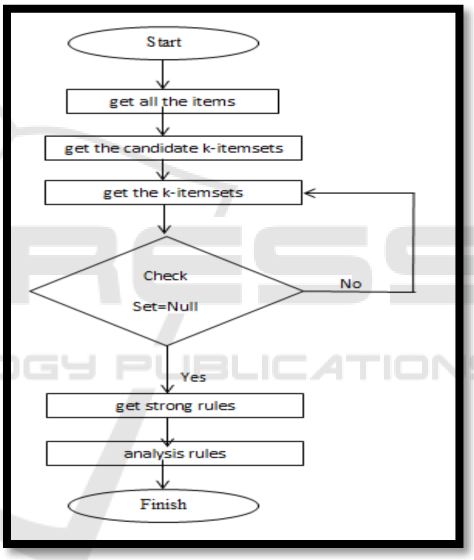

Figure 1: Flowchart of the Apriori Algorithm.

From Figure 1, the first step in association rule mining is to find frequent itemsets. K-itemset is defined as an itemsetas k items, Lk as the set of frequent k-itemsets, and Ck as the set of candidate k-itemset. Pseude-code following a priori algorithm used to generate all frequent itemsets and pruning frequent itemset unattractive in a transaction data base.

2.3 Geographic Information System

Geographic information system or information system based mapping and geography is management of information closely related to a mapping system. GIS technology integrating the operation of a database that are used today, as the collection of data, as well as statistic analysis with visualization. There are two kinds of data in geographic information system i.e. the data geospasial or usually called spatial data and non spatial (attributes) data (Berry, 1993).

3 EXPERIMENTAL RESULT

3.1 Desriptive

First, we identification the data with descriptive analysis using by using mapping, we can see for the distribution number of public libraries as follows:

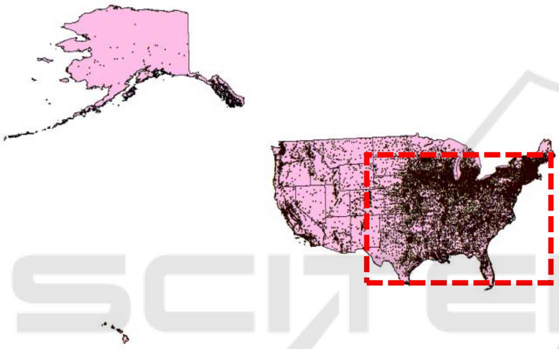

Figure 2: Distribution of Public Libraries in US

There are five kind of location type in Public Libraries is Bookmobile type, Books by mail type, Branch library, Central Library, and library system type. From the Figure 2, we can conclude that the spread of public libraries in the eastern US state of New York that region and surrounding areas.

3.2 Association Rules Apriori Algorithm

The result of the analysis using Big ML software can be shown in table 1. Based on this table, the minimum support for apriori algorithm is 0.0830 and the minimum confidence 0.2920. From these values, we can analyze the assosiation rules between three atrributes of public libraries data.

Based on table 1, all of the confidence values is greater than 0.8 for six association pattern are formed i.e.

a. Rule{City includes Chicago} => {Location type=Branch library}
Rules with support value is 0.0029, confidence is 0.9625 and lift ratio is 3.4006. It’s mean that 0.29% or about 77 Branch library of the whole location type happen in Chicago. Confidence value means that in case of Chicago is going to have Branch library with a confidence level of 96.25%. From the lift ratio value, we can see how strong the rule formed by the apriori pattern mining algorithms. The value of lift ratio is positive (≥ 0). If the lift ratio value equals to 1, then the rule {City includes Chicago} => {Location type=Branch library} often occurred together but independently.

Table 1: Result of a rules with Apriori Algorithm using BIG ML software.

Rules	City	Location Type	Support	Confi-dence	Lift Ratio
1	Chicago	Branch Library	0.0029	0.9625	3.4006
2	Brooklyn	Branch Library	0.0023	0.8857	3.1293
3	Losangeles	Branch Library	0.0021	0.9655	3.4113
4	Baltimore	Branch Library	0.0012	0.9167	3.2387
5	Sandiego	Branch Library	0.0013	0.8293	2.9299
6	Miami	Branch Library	0.0001	0.8000	2.8265

b. Rule{City includes Brooklyn} => {Location type=Branch library}
Rules with support value is 0.0023, confidence is 0.8857 and lift ratio is 3.1293. It’s mean that 0.23% or about 61 Branch library of the whole location type happen in Brooklyn, Brooklyn is going to have Branch library with a confidence level of 88.571%, and lift ratio value is 3.1293. From the lift ratio value, we can see how strong the rule formed by the apriori pattern mining algorithms. The value of lift ratio is positive (≥ 0). If the lift ratio value equals to 1, then the rule, then the rule {City includes Brooklyn} => {Location type=Branch library} often occurred together but independently.

c. Rule{City includes Losangeles} => {Location type=Branch library}
Rules with support value is 0.0021, confidence is 0.9655 and lift ratio is 3.4113. It’s mean that 0.21% or about 55 Branch library of the whole location type happen in Losangeles, Losangeles

is going to have Branch library with a confidence level of 96.55%, and lift ratio value is 3.4113. From the lift ratio value, we can see how strong the rule formed by the apriori pattern mining algorithms. The value of lift ratio is positive (≥ 0). If the lift ratio value equals to 1, then the rule {City includes Losangeles} => {Location type=Branch library} often occurred together but independently.

d. **Rule{City includes Baltimore} => {Location type=Branch library}**
Rules with support value is 0.0012, confidence is 0.9167 and lift ratio is 3.`2387. It's mean that 0.12% or about 33 Branch library of the whole location type happen in Baltimore, Baltimore is going to have Branch library with a confidence level of 91.67%, and lift ratio value of 3.2387. From the lift ratio value, we can see how strong the rule formed by the apriori pattern mining algorithms. The value of lift ratio is positive (≥ 0). If the lift ratio value equals to 1, then the rule {City includes Baltimore} => {Location type=Branch library} often occurred together but independently.

e. **Rule{City includes Sandiego} => {Location type=Branch library }**
Rules with support value is 0.0013, confidence is 0.8293 and lift ratio is 2.9299. It's mean that 0.13% or about 33 Branch library of the whole location type happen in Sandiego, Sandiego is going to have Branch library with a confidence level of 91.67%, and lift ratio value of 3.2387. From the lift ratio value, we can see how strong the rule formed by the apriori pattern mining algorithms. The value of lift ratio is positive (≥ 0). If the lift ratio value equals to 1, then the rule {City includes Sandiego} => {Location type=Branch library} often occurred together but independently.

f. **Rule{City includes Miami} => {Location type=Branch library}**
Rules with support value is 0.0001, confidence is 0.80 and lift ratio is 2.8265. It's mean that 0.1% or about 27 Branch library of the whole location type happen in Miami, Miami is going to have Branch library with a confidence level of 91.67%, and lift ratio value of 3.2387. From the lift ratio value, we can see how strong the rule formed by the apriori pattern mining algorithms. The value of lift ratio is positive (≥ 0). If the lift ratio value equals to 1, then the rule {City includes Miami} => {Location type=Branch library} often occurred together but independently.

4 CONCLUSION

The results of association rules analysis can be concluded that the most recommended rules are Rules{City includes Chicago} => {Location type=Branch library}; Rules{City includes Brooklyn} => {Location type=Branch library}; Rules{City includes Losangeles} => {Location type=Branch library}; Rules{City includes Baltiomore} => {Location type=Branch library}; Rules{City includes Sandiego} => {Location type=Branch library}; Rules{City includes Miami} => {Location type=Branch library}. From the six rules mentioned above, we known that the most visited library in these six states in US is Branch Library. It's because the location of the Branch Library usually near to people's houses. Therefore, to be visited by more people, the main branch library should be developed near to the people's house. Management of other types of libraries should refer to branch library management. It's shown by the result of the association rules analysis, which confirmed that the Branch Library is the library that most visited by population in that area.

REFERENCES

ALA Library, 2014. *The Nation's Largest Public Libraries: Top 25 Rankings.* American Library Association.

Agrawal, R., Imielinski, T., and Swami, A., 1993. Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD International Conferenceon Management of Data, SIGMOD '93,* New York, NY, USA, pp.207–216. ACM Press.

Berry, J. K., 1993. *Beyond Mapping: Concepts, Algorithms and Issues in GIS*, Fort Collins, CO: GIS World Books.

Brin, S., Motwani, R., and Silverstein., 1997. Beyond Market Baskets: Generalizing Association Rule to Correlations. *Proceedings ACM SIGMOD Conference on Management of Data (SIGMOD1997),* pp. 265 – 276.

Han, J. and Kamber, M., 2006. *Data Mining Concepts and Techniques Second Edition*. Morgan Kauffman: San Francisco, 2006.

Jiawei, H., and Micheline, K., 2006. *Data Mining: Concepts and Techniques.* MORGAN KAUFMANN PUBLISHER, An Imprint of Elsevier.

Nath, A. 2015. Big Data Security Issues and Challenges. International Journal of Innovative Research In Advanve Enggineering, 2(2), 15–20

Rubin, R. E., 2010. *Foundations of library and Information Science (3rd ed)*. Neal-Schuman Publishers: New York.

Srikant R, 1996. *Fast algorithms for mining association rules and sequential patterns*, University of Wisconsin.

Zhao, Y., and Yonghua, C., 2014. *Data Mining Applications with R*. Academic Press: UK, Amsterdam, the Netherlands.