ISSN: 2277-9655
Impact Factor: 4.116

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

OPTIMIZATION OF FREQUENT REPRESENTATIVE PATTERN SETS USING NAIVE BAYES TECHNIQUE

Preeti C. Tarange*, Rajesh H. Kulkarni*
* Department of Computer Engineering, JSPM Narhe Technical Campus, Pune, India

DOI: 10.5281/zenodo.58653

ABSTRACT

Frequent item sets plays a vital part in abundant actions of data mining which continuously endeavor to determine fascinating arrangement through number of databases like association rules, sequences, classifiers based on correlations, episodes, clusters and so on. The time period required to generate periodic item sets enacts a crucial aspect. Various algorithms are developed, taking only time aspect into account. This turns up requirement of determining small number recurrent producing patterns. Here, fundamental problems on mining of periodic item sets as well as pattern sets are elaborated. Frequent pattern mining generally develops a large number of recurrent patterns that depict complex concerns on understanding, viewing and next investigation of derived patterns. This results in determining a small number of classical arrangements of patterns to approximate all other patterns to the best suitable extent. In latest practices in the research of recurrent mining of pattern to determine a minimal classical set of pattern having zero error algorithm known as MinRPset is implemented. MinRPset determines modest result which one can probably implement in experimental procedures for the mentioned situation and it takes ample time to conclude as soon as the count of periodic closed arrangement patterns is less than one million. MinRPset is highly utilizes memory space and time on numerous heavy datasets when the count of recurrent closed pattern is greater. To solve this difficulty, another algorithm called FlexRPset is utilized, that introduces one supplementary parameter K for facilitating users for making adjustment amongst efficiency and result size. An additional view to provide users to make intervene satisfaction is utilized. Some recurrent pattern mining usually develops a more count of periodic patterns that require a huge objective of understanding, visualizing and next investigation of the developed patterns. This increases the demand to find less count of repeatedly occurring patterns. In the proposed work, system classifies the patterns on using Machine Learning, which classifies the items browsed by user for finding exact pattern set. By using naive bayes technique, service provider can increase his sell and also the customers can be advised for exact items they need through exact pattern generation.

Keywords: Representative patterns, frequent pattern summarization, frequent item sets, pattern sets.

INTRODUCTION

Data mining is an integrative subdivision of a field of computer science. It is the computational mechanism to invent patterns in huge number of data sets known as big data. The extensive objective of technique of the data mining is to find out knowledge through a set of data and alter it in a simple design for further use [1].

Data mining emerges as a prominent technique of obtaining knowledge or information through huge data. It is trending nowadays because of increased data count as well as endless requirement of converting the lots of data to the information. It is extensively applicable in various sectors like bioinformatics, business analytics, marketing, security and so on. An inevitable step of data mining is searching frequently occurring patterns which involve in vital role in clustering, associations mining, correlations, etc. Recurrent patterns take place in a data set intermittently like item sets, subsequences or substructures [1]. An item set A (or subsequence, or substructure) is said to be frequent if it fulfils the predetermined count of minimum support, where

$$Supp(A) = \frac{number\ of\ occurences\ of\ A}{total\ number\ of\ itemsets\ in\ a\ database}$$

The yield standard of connection is conveyed in the terms of Aϕ. Guideline A, D affects the check of minimum sponsorship of AUB in D. This rule has assurance C in dataset of D.

$$Confidence(A - B) = P(\frac{B}{A})$$

It is a two-step process in which all recurrent item sets are determined and strong association rules are achieved from the recurrent item sets. Mining of intermittent patterns from various patterns is vital aspect in data mining. Other concepts of data mining can be evaluated with the help of these concepts. It is start of technical training of data mining as it denotes the remarkable concept of data mining.

1 Pattern Sets

A pattern is a template, model or form which helps to create or to generate parts of things. In information mining we say that an example is a specific conduct of information, course of action or development that may be of a business interest. A frequent pattern sets include item sets, sub-sequences, or substructures which appear in the data set in a recurrent manner with not less than a user stated threshold value. A substructure might appertain to various organizational frameworks like sub graphs, sub trees and sub lattices that can be integrated with item sets. If substructure is generated number of times within database, it is denoted as frequent pattern. Identifying recurrent patterns involves a crucial aspect in mining applications, interrelationships and various other related interconnections in data. Besides, it is beneficial in the applications of data indexing, segregating, clustering and alternative data mining processes [2].

2 Frequent Item Sets

Another idea is an intermittent thing set which has a place with a classification of example set. A repetitive thing set is a parameter that is determined by the client in the database. The parameter is called as a backing of a thing set. Every subset from a repetitive thing set is an intermittent example as well. This characteristic is also known as Apriori or downward closure characteristic. It explains that we do not demand to find number of the item set if subset is not frequent. This will get to be conceivable because of the counter monotone property of backing. Visit thing sets ought to fulfil the base backing of client limit. The backing for a thing sets never better than backing for a subset. In the event that we separate the whole database in a few segments, an arrangement of things might be repetitive alone when it is incessant in least one segment. To locate a successive thing set we ought to experience all sub things sets which they themselves are incessant in light of Downward Closure trademark. The recurrent item sets are found to reduce the problem of association rules [3]–[5].

3 Classification based on Bayesian Theorem

Arrangement taking into account Bayesian hypothesis alludes to a superintended preparing system and a measurable methodology for characterizing reason. Take into account an elemental model based on probability and it allows us to seize uncertainty related to the model in a noble technique by determining probabilities of the results. This may figure out distinguishing and prognostic problems. Classification based on Bayesian theorem facilitates functional training algorithms and prior information and obtained data can be logged. Classification based on Bayesian theorem facilitates an effective perspective for realizing and deciding various training methods. It determines accurate possibilities for interpretation and its sturdy to noise in the input data.

4 Naive Bayesian Classification

It is established on Bayesian Theorem and is especially appropriate in case the extent of the input arguments is very large. Naive Bayes design models employ the technique of maximal tendency to estimate the parameters. Despite the over abridged prediction, it ever executes more magnificently in ample complicated real world problems. The major benefit of this segregation is that it requires very less learning information to evaluate the parameters.

PROBLEM STATEMENT

Mining frequent patterns from different arrangements is the most vital concept of mining of data as well as it is hard to find frequent patterns. So to overcome this problem important algorithms of data mining are used in different systems.

Some of the ascertaining issues are:

- Establishing a consolidated concept of mining of data
- Increasing for huge sized data and streams of huge data speed
- Prospecting sequential information and data of time domain
- Prospecting complicated information from convoluted data
- Information mining in network setups

- Dispersed data excavation and finding Multi-Agent information
- Data determination for Biological and Environmental issues
- Excavation of data in process issues
- Surveillance, Confidentiality and Data integration
- Handling dynamic, unstable and cost susceptible data acceptance.

MOTIVATION

1) It is the procedure of determining new patterns from huge information source. Few information excavation methods like neural network based on artificial intelligence, trees of decision and adjacent nearby technique are intermittently applied. Each methods estimates information in various forms. Excavation of informationhas significance with respect to determining the arrangements, interpreting and exploration of information
2) A Mining limited recurrent sets of item from segregated unprobabilistic information.
3) Excavating recurrent variables and sets of item from streams of segregated information.
4) Excavating possibly recurrent arrangements of sequence in unprobabilistic sources of information.
5) Large numbers of probable sequence arrangements are concealed in databases.
6) A method should
   - determine the detailed bunch of arrangements, whenever feasible, fulfilling the minimal recurrence threshold value.
   - be extremely potential, extensive involving very low iterations of database scanning.
   - be integrate number of types of user customized restrictions.

LITERATURE SURVEY

Vivek B. Satpute proposed that Design mining as of late accomplished significance in the information digging group for the reason of its capacity of being utilized as imperative instrument for the learning disclosure and its appropriateness in the other information mining occupations like order and bunching. Affiliation principles are dependably of enthusiasm to the both database group and in addition information mining clients. Here a review is given of past studies made here and perceive some imperative holes accessible in the present information [1].

Javeriya Naaz Ishtiyaque Syed and Rajeshri R. Shelke proposed continuous sets of items that assume a vital part in numerous information mining undertakings that try to find out appealing examples from databases, such as affiliation rules, relationships, arrangements, scenes, classifiers, groups and some more. Numerous scientists concocted thoughts to produce the successive thing sets. The time required for creating recurrent sets assumes a vital part. A few calculations are outlined, considering just the time variable. Their study incorporates profundity investigation of calculations and examines a few issues of producing incessant item sets (design sets) from the calculation. The binding together element among the inward working of different mining calcu lations is investigated. Some Frequent example mining regularly creates countless examples, which forces an incredible test on imagining, understanding and further investigation of the produced designs. This rises the requirement for discovering little number regular happen- ing designs. In their paper, they clarified the fundamental successive thing set, design sets mining issues. They depicted the principle systems used to take care of these issues and give a far reaching review of the most powerful calculations that were proposed amid the most recent decade [2].

Thabet Slimani and Amor Lazzez introduced the sequence of information digging produces different examples from a given information source. The most perceived infor- mation mining errands are the procedure of finding continuous thing sets, incessant successive examples, regular consecutive guidelines and regular affiliation rules. Various proficient calculations have been proposed to do the above procedures. Successive example mining has been an engaged theme in information mining research with a decent number of references in writing and thus an essential advancement has been made, shifting from performant calculations for continuous thing set mining in exchange databases to complex calculations, for example, consecutive example mining, organized example mining, connection mining. Affiliation Rule mining (ARM) is one the very pinnacle of current information mining systems intended to gathering questions together from substantial databases expecting to extricate the intriguing connection and connection among tremendous measure of information. In their article, they give a brief survey and examination of the momentum status of incessant example mining and talk about some encouraging exploration bearings. Also, this paper incorpo- rates a similar study between the exhibitions of the depicted methodologies [3].

In the paper by K. Kavitha and C. Anand, they inferred that, an immense check of incessant example sets is utilized to locate the suc- cessive shut example sets. An agent design sets are created by an incessant shut example. To discover delegate design set two calculations MinRPset and FlexRP set are utilized. MinRP set and FlexRP set is utilized to keep record of the incessant shut example set. It concedes the best inexact result. MinRP set turns moderate when enormous measure of successive shut example sets devours more memory space and it is costly. To take care of this issue, FlexRP set is utilized, which gives one extra parameter K to locate the base number of agent example sets by utilizing all regular shut example sets [5].

EXISTING ALGORITHMS

1 MinRPset Algorithm

The clients here are conceded to unwind the conditions in the issue definition to advance lessen the number of delegate examples. Henceforth this methodology is an exceptionally adaptable way to deal with find the delegate designs. MinRPset produces the little arrangement that one can have by and by under the given issue and it requires adequate measure of investment to complete when the quantity of incessant shut examples is underneath one million. MinRPset is exceptionally memory space devouring and tedious on some thick datasets when the quantity of regular shut example is higher [4].

Consider,

- F: the arrangement of successive examples in a dataset D as for limit min sup,
- ˆ F: be the arrangement of examples with backing no not exactly min sup ∗ (1 - e) in D.
- Obviously, F ⊆ˆF. Given an example X ∈ˆF, we utilize C(X) to signify the arrangement of successive examples that can be secured by X. Here, C(X) ⊆ F. On the off chance that X is regular, then X ∈ C(X).

Algorithm

1) Mine examples with backing ≥ min sup∗ (1 - €) and store them in a CFP-tree;
2) DFS Search CXs (root);
3) Remove non-shut passages from C(X)s;
4) Apply the insatiable set spread calcula- tion on C(X)s to discover delegate exam- ples and yield them;

When €=0, the agent examples are shut incessant examples.

2 FlexRP Set Algorithm

Coordinate all question tokens Only records that contain all the inquiry tokens are incor- porated into the coordinated rundown. Here, co-event measures utilizing page numbers are characterized. The most effective method to concentrate bunches of examples from pieces to speak to various semantic relations that exist between two words is appeared [4].

Algorithm

Data:

- Input: cnode is a CFP-tree hub; /cnode is the root hub at first.
- K is the base number of times that a regu- lar shut example should be secured;
- Output: C(X)s; Description

1) for every section E cnode from left to right do
2) if E is not set apart as non-shut then
3) if E.child 6= NULL then
4) Flex Search CXs (E.child);
5) if E is more regular than its kid sections then
6) if (E is incessant AND E is secured not as much as K times) OR (∃ a progenitor passage E' of E such that E' is successive, E' can be Q-shrouded by E and E' is secured not as much as K times) then
7) X=E.pattern;
8) (X) = Search CX (root, X, E.support);

PROPOSED WORK

1 Dataset Training

To prepare the dataset to the pursuit in view of the example to recover an accumulation of reports identified with the question design. After an inquiry is submitted to an internet searcher, a rundown of Web bits is come back to the client. Expect that if a catchphrase/expression exists regularly in the Web-pieces of a pre-determined question, it speaks to an essential idea identified with the inquiry since it exists together in close closeness with the question in the top records.

2 Pattern Generation

All words or products or items from the content are not considered as Pattern. Normally a few words or products happen of tentimes in the majority of the records. Due to this property, their segregation force is

ISSN: 2277-9655
Impact Factor: 4.116

immaterial. These sorts of words are called stop words and these words can be sifted through amid patternization.

3 Flex RP Set

Coordinate all inquiry words. Only records that contain all the question words are incorporated into the coordinated rundown. Here, coevent measures utilizing page checks are characterized. Step by step instructions to concentrate groups of examples from scraps to speak to various semantic relations that exist between two words is appeared.

4 Naive Bayes Algorithm

Naive Bayes classifiers are profoundly adaptable, requiring various parameters direct in the quantity of variables (components/indicators) in a learning issue. In the proposed technique, the guileless Bayes classifiers is uti lized to classify the products on utilizing Machine Learning, which characterizes designs or patterns on search which could manage slant order.

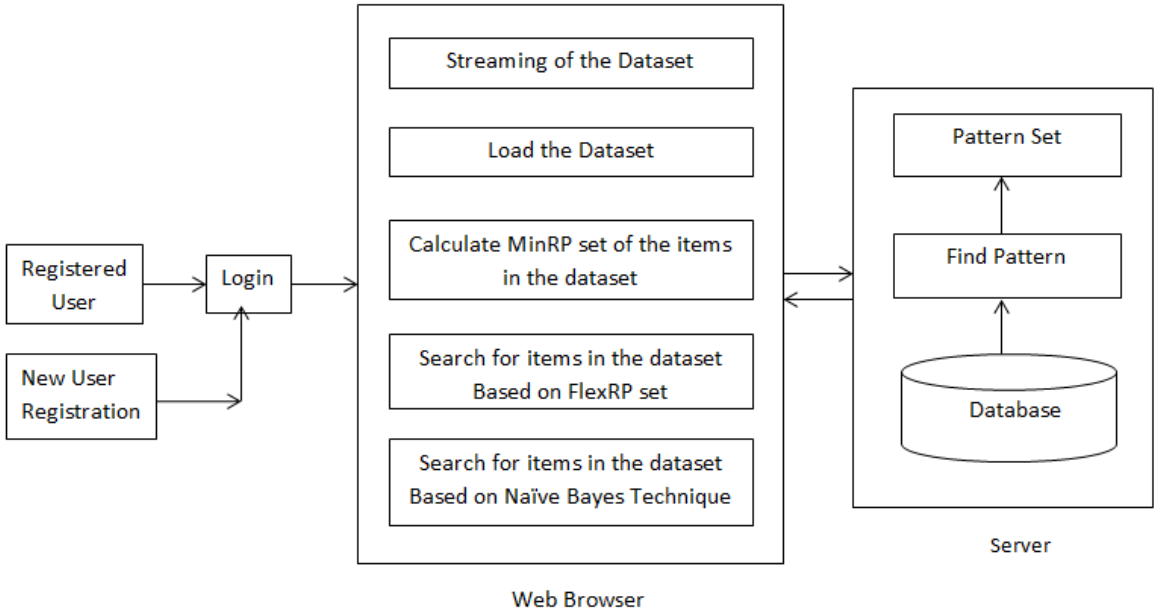

Figure 1: System Architecture Diagram

OBJECTIVE OF PROPOSED WORK

1) To manufacture an effective adaptable calculation a proficient and versatile cal- culation for incremental digging issue for continuous thing sets. In former method- ologies such as Apriori calculation and FP-Growth, entire procedure is to be reintroduced from base and is not suitable for element examination.
2) To study different example mining calculations that can be utilized for finding learning (designs) from human collaborations, additionally to study and think about different affiliation standard mining calculations.
3) To minimize the expense of discovering regular thing set utilizing the new ap- proach of Naive Bayes calculation.
4) Reduce running time and memory usage.
5) Have good efficiency.

MATHEMATICAL MODEL FOR PROPOSED SYSTEM

Let us consider, a set S
where S = {U, R, D, SER}
Here,
S is system that contains U as a set of users
Where U = {U1 , U2, U3, ..., Un }; SER: Server; R: Collection of Request
where R = {R1, R2 , R3, ..., Rn }; D: Database
U(R1) → S (browse (Items))
S → create (log(i1, i2, i3, ..., i)) → Pattern Set
Items (i1, i2, i3 , ...) → Naive Bayes (Clusters)

ISSN: 2277-9655
Impact Factor: 4.116

U (R1) → Display (Related items)x

RESULTS AND DISCUSSION

In this work the pattern set of items bought on the Food Mall website are found. Server will use data stored in browser, and Naive Bayes will classify the items browsed by user for finding exact pattern set. When next time user browses the data it will show only that items which are in pattern sets and mostly browsed. By this approach provider can increase his sell and also the customers can be advised for exact items they need through pattern generation. Here we are using minimum support factor which adds accuracy to the pattern set. Final result will be that customer or user will get maximum relevant items from food mall website. Following figure 2 and figure 3 shows the results calculation time in seconds of the food mall website data using MinRPset, FlexRPset and Naïve Bayes Technique. In this work, the food mall website dataset is obtained from KDD cup 2000.

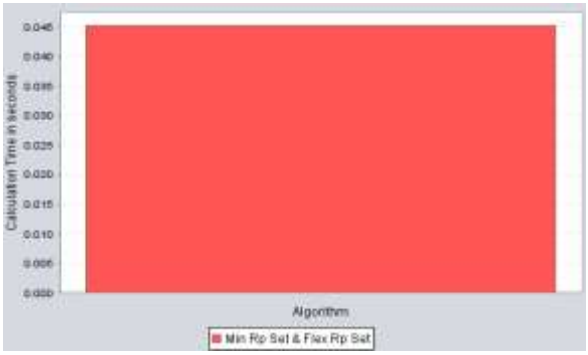

Figure 2: Result Using MinRPset and FlexRPset
Calculation Time = 0.045 seconds

Figure 3: Result Using Naïve Bayes Technique
Calculation Time = 0.025 seconds

CONCLUSION

An immense number of incessant example sets are expected to locate the regular shut example sets. A regular shut example finds a delegate design sets. For finding an agent design set two calculations MinRPset and FlexRPset are utilized. MinRPset and FlexRP set is used to keep record of the frequent closed pattern set which provides the best approximate result. MinRPset turns out to be moderate when vast measure of incessant shut example sets that expends more memory space. MinRPset is more expensive. To beat this issue, FlexRPset is utilized, which requires additional parameter k quality to locate the base number of agent example sets by covering all continuous shut example sets. In the proposed technique, Naive Bayesian Classification is especially suited when the dimensionality of the inputs is immense. Naive Bayes models utilize the strategy for most extreme probability for Parameter estimation. In dislike over-improved desire, it generally performs more magnificent in rich complex cer tifiable circumstances. The fundamental point of interest of this order is that it needs a little tally of preparing information to appraise the parameters.

ACKNOWLEDGEMENTS

With immense pleasure, It gives me proud privilege to complete this work under the valuable guidance of Prof. Rajesh H. Kulkarni (H.O.D, Computer Engineering) and I am also extremely grateful to Dr. D. M. Yadav (Director, JSPM NTC) for providing all facilities and help for smooth progress of work. I would also like to thank all the Staff Members of Computer Engineering Department, Management, friends and my family members, who have directly or indirectly guided and helped me for the preparation of this work and gave me an unending support right from the stage the idea was conceived.

REFERENCES

[1] Vivek B. Satpute, A Review on Frequent Pattern Mining International Journal of Engineering Research and General Science, Volume 2, Issue 6, October-November, 2014, ISSN 2091-2730.
[2] Javeriya Naaz, Ishtiyaque Syed, Rajeshri R.Shelke, Analysis of Frequent Item sets and Pattern Sets Mining Algorithms, International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169, Volume: 3 Issue: 2, page no: 249-253.
[3] Thabet Slimani, Amor Lazzez , Efficient Analysis of Pattern and Association Rule Mining Approaches,

College of Com- puter Science and Information Technology , Taif University, KSA.
[4] Guimei Liu, Haojun Zhang and Limsoon Wong, A Flexible Approach to Finding Representative Pattern Sets, IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 7, July 2014.
[5] K. Kavitha, C. Anand, A Novel Approach in Data Mining for Representative Pattern Sets, International Conference on Science, Technology, Engineering and Management [ICON- STEM15], Journal of Chemical and Pharmaceutical Sci- ences, ISSN: 0974-2115.
[6] A.Bykowski and C.Rigotti, A condensed representation to find frequent patterns, NewYork, NY, USA, 2001, 267-273.
[7] A. K. Poernomo and V. Gopal Krishnan, CP-summary: A concise representation for browsing frequent item sets, New York, NY, USA, 2009, 687-696.
[8] J. Wang, J. Han, Y. Lu and P. Tzvetkov, TFP: An efficient algorithm for mining top-k frequent closed itemsets, 652-664.2005.
[9] R. J. Bayardo, TFP: An efficient algorithm for mining top-k frequent closed itemsets, 652-664.2005.
[10] N. Pasquier, Y. Bastide, R. Taouil and L. Lakhal, Discover- ing frequent closed itemsets for association rules, Jerusalem, Israel, 1999, 398-416.