EFFICIENT FREQUENT PATTERN SEARCHING USING AMOEBA AND DECISION TREE TECHNIQUE

B. Mahalakshmi, A. T. Nandhini
K. S. Rangasamy College of Technology

Abstract

Data mining plays a vital role due to improvement in technologies, thereby extracting the hidden information or patterns of data from a huge database or collection of large data set. In this paper, concerning about finding the frequent pattern of words from a collected dataset using amoeba model. A new algorithm called AMOEBA, is used to find the chain of possible frequent patterns. All documents in dataset can be analyzed by reading the files. The semantic word in the document can be scanned with the help of wordnet tool. Every semantic word in the document is scanned for further processing. Now, the AMOEBA model is used for clustering both document and word simultaneously. The generated model is optimized by AMOEBA algorithm which provides efficiency. Finally the optimized word can be chosen by using decision tree, which helps to make clear result for the user search. Therefore this algorithm will win the space and time complexity by, in-comparison with Aprior and FP-Growth.

Introduction

Pattern mining is an efficient and scalable method for mining the complete set of frequent patterns by pattern fragment growth. Frequent patterns are itemsets, subsequences or a substructure that appears in a data set with frequency not less than user specified threshold. Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. Let $I = \{i_1, i_2, i_3, \ldots, i_n\}$ be a set of binary attributes called items. Let $D = \{t_1, t_2, t_3, \ldots, t_m\}$ be a set of transactions called the database. Each transaction in $D$ has a unique transaction ID and contains a subset of the items in $I$. A rule is defined as an implication of the form: $X \rightarrow Y$, where $X, Y \subseteq I$. In order to select interesting rules from the set of all possible rules, constraints on various measures of significance and interest are used. The best known constraints are minimum thresholds on support and confidence. Let $X$ is an itemset, $X \rightarrow Y$ an association rule and $T$ a set of transactions of a given database. Support: Support is an indication of how frequently the itemset appears in the dataset, $\text{Support}(X) = \frac{|\{t \in T | X \subseteq t\}|}{|T|}$. Confidence: Confidence is an indication of how repeatedly the rule has been found to be true. The confidence value of a rule, $X \rightarrow Y$, with respect to a set of transactions $T$, is the proportion of the transactions that contains $X$ which also contains $Y$, $\text{Conf}(X \rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)}$.

The FP-Growth Algorithm will mining the pattern by the complete set of
frequent patterns using pattern fragment growth, is an efficient and scalable method. For storing compressed and crucial information about frequent patterns, it uses an extended prefix-tree structure called frequent-pattern tree (FP-tree). The performance metric of this algorithm is better when compared with APRIORI. In this algorithm introducing frequent item sets without using candidate generations. This algorithm has been come up with a divide-and-conquer strategy. Apriori is one of the most commonly used association mining algorithm for finding the frequent patterns. An assigned support and calculated confidence factors is calculated for finding its frequent patterns.

A new algorithm named AMOEBA is proposed based on the characteristics of unicellular organism amoeba. This algorithm is planned to rise above the pre-calculations of some association mining algorithms.

Overview of the literature survey

Nobuo Suzuki et.al., discussed to get a more frequency resources using radio systems with frequency sharing, which is one of the critical technique. The characteristics by a series of data can be taken using frequent sequence mining technology.

Dinesh J. Prajapati et.al., describes, in a distributed environment a sales data is placed, from that data the consistent and inconsistent association rule can be identified. It can be performed by using mapreduce algorithm to provide a useful knowledge to the domain expert.

Songfeng lu et.al., uses an FP-growth algorithm for mining frequent itemset. The EFP-growth(Enhanced Frequent Pattern) is used to achieve the best quality of FP-growth. In a transaction database EFP-growth is used to discover the frequent pattern. Depends on this method the minimum supports are decreased under execution time.

Roshni Chandran et.al., discussed that, in a real-time data stream a discovery of knowledge is increased by using time-efficient Hadoop CanTree- GTree algorithm. It mines the complete frequent item sets from real time transactions with the help of sliding window technique.

Methodology

Amoeba is a unicellular organism which is irregular of its shape and belongs to phylum protozoa. The name "amibe" was specified to its by Bory de Saint-Vincent, from Greek amoibe, sense change. Amoeba moves by means of pseudopodia or "false feet". There are many hypothesis have been introduced to simplify the mechanism of AMOEBA movement, but still there is a mystery of exact association of AMOEBA. These attribute features of amoeba guide to the evolution of a new association mining algorithm AMOEBA. Amoeba moves in a route which is not detailed. This is due to the existence of false feet in amoeba. This characteristic was enabled for the evolution of this new algorithm. This can be termed as attribute value determining. This determination can be achieved by using functional dependency i.e. determining an attribute value by another attribute value. The determination also includes, at what percentage an attribute value determines
other attribute distinct values. This algorithm works mainly on two principles:
- Determining another attribute value in a data set using an attribute value. 
  (Or) Determining another attribute value in a data set which determined the attribute value.
- Probability of an attribute value being determined by an attribute value.

Extraction of documents

Constraint to cluster the document is created automatically by using NE extractor. Document is parsed to identify named entity. NE extractor, extract entity form documents which are provided by user. If there are overlapping NEs in two documents and the number of overlapping NEs is larger, and then an entity added as constraint for document clustering. Named-entity-based document constraints, is likely to integrate additional lexical constraints resulting from existing knowledge sources to further improve clustering results.

Document constraint using NE extractor

Mining Semantic Words

Constraint to cluster the word is created automatically by using WordNet which is lexical database for English. The semantic relatedness between words can be measured based on the word hierarchies in the Wordnet. Parse the document and compare word with WordNet to create constraint. Furthermore, while word knowledge can be transferred to the document side during co-clustering, with additional word constraints, it is achievable to further progress in document clustering as well.

Mining Semantic word on Wordnet tool

Retrieval of clustered word

The document constraint extracted from NE overlapping and word constraint created from WordNet. AMOeba is modeled for both document and word to perform the cluster simultaneously. AMOeba is used to formulate the prior information for both document and word latent labels.
Retrieval of Clustered Word

Determining the probability

AMOEBA generated model is optimize by amoeba algorithm for efficiency. EM algorithm is optimizing the latent labels in the model. There are two steps in the amoeba algorithm:

- Determining another attribute value in a data set which determined the attribute value.
- Probability of an attribute value being determined by an attribute value.

Optimizing model based on Co-clustering

Conclusion

Algorithm AMOEBA does not require the construction of transaction data set, calculation factors like support and confidence and assembling of frequent pattern trees. The restriction AMOEBA is, input data set must be discredited because determination through chance can be defined on discrete values. The probability of frequent items of these frequent items sets decreases with increase in size of frequent item set. Choice of initial attribute value, manipulate the evolution of frequent items chain for the algorithm AMOEBAB. This is due to that, if the determination values of other attributes by initial attribute value are lowest of its probability or zero then such initial attribute value becomes void for finding frequent items chain. Selection of such initial attribute value whose possibility of determining other attribute values is zero, results is to identify out the infrequent items in a data set. This attribute value cannot be integrated in frequent item set. A decision tree is a decision support tool that uses a tree like graph or model of decisions and their possible consequences, including chance event outcomes and its utility. It helps to identify a strategy most likely to reach a goal. The algorithms, Apriori and FP Growth cost more, when compared with the algorithm Amoeba in terms of disk usage.

References


