Sequential Pattern mining: Genetic Algorithm

*Anita Rani, Lecturer, Computer Sc., VSPP College, Ahlupur
**Rajni, Lecturer, Computer Sc., VSPP College, Ahlupur
***Manju Bala, AP Computer Sc., JCDM, Sirsa

Abstract

Mining Sequential Patterns in large databases has become an important data mining task with broad applications. It is an important task in data mining field, which describes potential sequenced relationships among items in a database. There are many different algorithms introduced for this task. Conventional algorithms can find the exact optimal Sequential Pattern rule but it takes a long time, particularly when they are applied on large databases. Nowadays, some evolutionary algorithms, the algorithms for performing sequential pattern mining can assure optimum solutions but they do not take into consideration the time taken to reach such solutions. In this paper we propose a new algorithm based on genetic concepts which gives, may be a non-optimal solution but in a reasonable time (polynomial) of execution.

I INTRODUCTION:

MINING Sequential Patterns in large databases has become an important data mining task with broad applications, including business analysis, web mining, security and bio-sequences analysis. It extracts patterns that appear more frequently than a user-specified minimum support while maintaining their item occurrence order. In this task, time is the most important factor, especially when the results are needed in a limited period of time. Mining Sequential Pattern algorithms takes a long time to find the rules especially when they are applied on large databases. On the other hand, the evolutionary algorithms can find good Sequential Pattern rules within a short time. Nowadays, some evolutionary algorithms were proposed and have been applied in a timely manner. Genetic Algorithm (GA), which is an evolutionary algorithm, can be used to discover Sequential Pattern rules in a short time. It is a general purpose search algorithm which use principles inspired by natural genetic populations to evolve solutions to problems.

Related Work: Existing approaches to find appropriate sequential sets for sequential pattern mining are mainly classified into two categories. The first is concerned with conventional methods and the other employs evolutionary based approaches. Under the first category, Agrawal and Srikant have introduced the Sequential Pattern mining problem [1], [2]. In addition, there were many fitness measures that were applied to be used in discovering Sequential Patterns, as in [8], and Piatesky-Shapiro [7] suggested some principles and property to choose the most appropriate measure depending on the problem. This paper developed an efficient approach and applied these principles to choose the appropriate measure for the proposed algorithm. Differently, evolutionary based methods adjusted the time problem of conventional sequential methods according to an optimization scheme.

Genetic Algorithm has been introduced in [7]. GA has many chromosome representations in these representations have been studied and the most appropriate one has been chosen based on David advice. For example, for a telecommunication database where each transaction includes a caller phone number, date and time of the call and destination country code, as in Table 1. Sequential Pattern Mining can provide us with this rule: when country code 91 is called, (%40) of callers will call country code 92 afterwards. Extracting such patterns can help Telecommunication companies to know the country codes that have a relation between them. So, the telecommunication companies can estimate the countries that have a specific order of the occurrences and give a discount on the calls to these countries. The challenge of extracting sequential patterns from telecommunication database draws upon research in databases, machine learning, optimization, and high-performance computing, to deliver advanced intelligent solutions. The algorithms for performing sequential pattern mining are not polynomial (NP-Complete). The complexities mainly arise in exploiting huge taxonomies (a telecommunication database may stock 20 millions of phone numbers), and dealing with the large amounts of transaction data that may be available. Many algorithms have been proposed for sequential pattern mining [2, 3]. These algorithms assure the optimum solution despite the time taken in finding it. In this paper we propose a new algorithm based on genetic framework which gives good solutions in a reasonable time of execution without assuring always the optimum solution. The rest of the paper is organized as follows. Sequential Patterns Mining and Genetic Algorithm are defined in section 2 and 3. Our approach of applying GA to Sequential Patterns in Telecommunication databases is described.

II. SEQUENTIAL PATTERNS

Sequential Pattern mining addresses the problem of discovering the existent maximal frequent sequences in a given database. The problem was first introduced by Agrawal and Srikant [1], [2], where the basic concept involved in pattern detection has been established. It seeks similar patterns in data transaction; this approach is useful when the data to be mined has some sequential nature to deal with databases that have time-series characteristics, i.e. when each piece of data is an ordered set of elements [3]. For example, it can be said that 60% of patient who takes medicine X will take medicine Y afterward, regardless of the time gap. Given a pharmacy database, where each transaction includes a patient ID, prescription time and its medicine, as in Table1, Sequential Pattern can be defined as follows.

Definition 1: Let I = {x1..xn} be a set of items. An itemset is a non-empty subset of items, and an itemset with k items is called k-itemset. A sequence 1 s = (X1..Xl) is an ordered list of itemsets and an itemset Xi(1 ≤ i ≤ l) in a sequence is called a transaction. In a set of sequences, a sequence s is maximal if s is not contained in other sequences.

Phone Number	Date and Time	Destination Country Code
446215872 446215872	14-12-2008 18:46:26 14-12-2008 22:23:12	92 63
446212212 446212212 446212212	13-12-2008 06:55:24 14-12-2008 05:32:25 14-12-2008 12:09:34	249 973 92
446241822 446241822	14-12-2008 11:03:44 14-12-2008 17:14:11	964 91

Table 1: Telecommunication Database

III. GENETIC ALGORITHM

Genetic Algorithm (GA) is general purpose search algorithm which use principles inspired by natural genetic populations to evolve solutions to problems. All GAs typically starts from a set, called population, of random solutions (candidate). These solutions are evolved by the repeated selection and variations of more fit solutions, following the principle of survival of the fittest. The elements of the population are called individuals or chromosomes, which represent candidate solutions. Chromosomes are typically selected according to the quality of solutions they represent. To 1 Each sequence in sequential patterns is considered as a rule to measure the quality of a solution, fitness function is assigned to each chromosome in the population. Hence, the better the fitness of a chromosome, the more possibility the chromosome has of being selected for reproduction and the more parts of its genetic material will be passed on to the next generations. Genetic Algorithms are very easy to develop and validate, which makes them highly attractive, if they applied. The algorithm is parallel; it can be applied to large populations efficiently, so if it begins with a poor original solution it can rapidly progress to good solutions. Use of mutation makes the method capable of identifying global optimal, even in very difficult problem domains. The method does not require knowledge about the distribution of the data, this way Gas can efficiently explore the space of possible solutions. This space is called search space, and it contains all the possible solutions that can be encoded [6]. Genetic algorithms are good at taking large, potentially huge search space and navigating them, looking for optimal combination of this solutions one might not Otherwise find in a lifetime.

Genetic Algorithm (GA), presented in [4, 6, 8], is a part of evolutionary computing, which is a rapidly growing area of artificial intelligence. Genetic algorithm starts with a set of solutions (represented by chromosomes) called population. Solutions from one population are taken and used to form a new population by mutation and crossover. This is motivated by a hope, that the new population will be better than the old one.

Best solutions which are selected to form new solutions (offspring) are selected according to their best fitness.

This is repeated until some condition (for example number of populations or improvement of the best solution) is satisfied. To measure the quality of a solution, fitness function is assigned to each chromosome in the population.

IV MINING SEQUENTIAL PATTERNS IN TELECOMMUNICATION DATABASE USING GA

In this section, we describe our Genetic Algorithm for mining Sequential Patterns in Telecommunication Database, this algorithm is called SPT-GA algorithm. Firstly, we present our chromosome structure and encoding schema, genetic operators, and then we define the fitness assignment and selection criteria. Finally, we give the structure of SPT-GA algorithm.

4.1 Chromosome. In this section, we discuss the used structure of GA chromosomes and how it is represented in this paper.

4.1.1 Structure. In Telecommunication Database, country code values are used for creating the chromosomes. In our algorithm, we used a fixed length chromosome, and its length is equal to number of country codes that are available in the database as in Figure 1.

Destination Country Code 1	Destination Country Code 2	Destination Country Code 3

Fig. 1: Chromosome Structure

4.1.2 Representation. In Genetic Algorithm, there are many alternatives to represent a chromosome based on other problem like binary and integer representation. To decide which representation is better to be used for
Sequential Pattern rules, we should use the short, low-order schemata are relevant to the underlying problem and relatively unrelated to schemata over other fixed positions. Also we should select the smallest alphabet that permits a natural expression of the problem, presented in [8]. In SPT-GA algorithm, we choose the binary representation because it is the most suitable for our algorithm and it needs less space and it represents the needed information (element occurred or not).

For example, using the Telecommunication Database in Table.1, if a sequence is equal to <249, 973,
91>, it can be represented as in Figure.2.

63	91	92	249	964	973
0	1	0	1	0	1

Fig. 2: Chromosome Representation

Additionally, as you can see in Figure.2, order cannot be extracted directly. To solve this problem, we decided to associate the transactions sequence as a metadata with each chromosome. For that, we use Vertical Bitmap Representation, presented in [5] that makes SPT-GA algorithm to take less time and space to be executed.

4.2 Genetic Operators. SPT-GA uses genetic operators to generate the offspring of the existing population. Genetic algorithm will send chromosomes that represented by binary string where each bit corresponds to an element occurrence (0 or 1), the number of bits is equal to the number of items. After encoding of the solution domain, initially many chromosome solutions are randomly generated depending on population size. Genetic algorithm will select the chromosome regarding to our fitness function. The major measure that used by our algorithm is Sequential Interestingness measure (SIM). Then crossover takes place, it selects genes from parent chromosomes and creates a new offspring. The simplest way how to do this is to choose randomly some crossover point and everything before this point copy from a first parent and then everything after a crossover point copy from the second parent, presented in [4, 6, 8]. Crossover can be as Figure 3. After a crossover is performed, mutation takes place. This is to prevent falling all solutions in population into a local optimum of solved problem. Mutation changes randomly the new offspring. For binary encoding we can switch a few randomly chosen bits from 1 to 0 or from 0 to 1, presented in [4, 6, 8].

Mutation can then be as in Figure 3.

Fig. 3: Crossover and Mutation Example

GA will repeat the operation until finding the best result.

4.3 Fitness Function. Shigeaki, Youichi, and Ryohei, in [10], proposed a method that discovers sequential patterns corresponding to the interests of users without using background knowledge. They defined a new criterion called the sequential interestingness measure (SIM).

Definition 2: The sequential interestingness measure of a rule A->C is:

$$\mathbf{SIM(A\text{->}C) = min_{Ci}\in C\ \{(Confidence(A \mid Ci))\alpha\} \times Support(AC)}$$

where ($\alpha \geq 0$) is a confidence priority that represents how important the frequency of the pattern is, Ci is supsequence of C, it represents a condition of sequence C, and i = 1 … n where n is the number of conditions in C. The first term of the criterion evaluates that the frequencies of the sub-patterns are not frequent while the second term evaluates that the frequency of the pattern is frequent.

4.4 SPT-GA Algorithm. In this section, we present SPT-GA algorithm that we proposed. In Figure 4, the pseudo code of SPT-GA algorithm is presented.

SPT-GA Algorithm

Input:

Population size: N

Maximum number of generations: G

```
Else
      Calculate the fitness $F(i, \alpha)$
4. Mutate and crossover P.
5. IF (fitness $\geq$ minF)
         Select fittest rules from P
6. Set $t = t + 1$
7. IF ($t > G$) then
        $S = P$
        Stop
  Else
        Go to Step 3
  Confidence priority: $\alpha$
  Minimum fitness: minF
  Rule guidance: antecedent A or consequent C.
  Output:
  Interesting sequence pattern: S
  Begin
  1. Initialize counter $t = 0$
  2. Generate population P of size N
  3. For each chromosome $i \varepsilon P$
          If A & C is given then
                Calculate the fitness $F(i, A, C, \alpha)$
          Else If A is given then
                Calculate the fitness $F(i, A, \alpha)$
          Else If C is given then
       Calculate the fitness $F(i, C, \alpha)$
```

5. End

After the encoding of the dataset, using bitmap representation [5], the algorithm starts by selecting individuals to initial population. Then the following processes are repeated until the pre-specified maximum number of generations is achieved. The fitness values determined for each selected individual given the rule antecedent or consequent. The fittest rules that are larger than or equal minimum fitness in P will be selected. Giving the antecedent or consequent is not mandatory but it will reduce the time of search and extract more desired rules. Existing chromosomes are used in generating new ones by applying crossover and mutation operators. Chromosomes survive based on their fitness used in the process. This way, the interesting set is determined and the target is achieved.

6. Conclusion.

In this paper, we applied Genetic Algorithm to find frequent sequences in Telecommunication Database in order to help Telecommunication companies to know the country codes that have a relation between them. So, the telecommunication companies can estimate the countries that have a specific order of the occurrences and give a discount on the calls to these countries. SPT-GA algorithm utilizes the property of evolutionary algorithm that discovers best rules in a short time with meaningful results.

8. References

[1] Agrawal R. and Srikant R. Mining Sequential Patterns. IBM Almaden

[2] Antunes C. and Oliveira A. Sequential Pattern Mining.

[3] Ayres J. Gehrke J. Yiu T. and Flannick J. Sequential Pattern Mining

[4] Blum C. and Li X. Swarm Intelligence in Optimization. Natural

[5] Colombetti M. and Dorigo M. Training Agents to Perform Sequential Behavior.

[6] Geng L. and Hamilton H. Interestingness Measures for Data Mining

[7] Goldberg D. Genetic Algorithms.

[8] Herrera F. Lozano M. and Verdegay J. Tackling Real-Coded Genetic Algorithms: Operators and tools for the Behaviour Analysis.