Mining Sequential Patterns in Telecommunication Database Using Genetic Algorithm

Mourad Ykhlef, Yousuf Aldukhayyil and Muath Alfawzan

King Saud University, College of Computer and Information Sciences,
Information System Department
Kingdom of Saudi Arabia.

ykhlef@ksu.edu.sa yaldukhayyil@yahoo.com muathalfawzan@gmail.com

Keywords: Sequential Patterns, Genetic Algorithm, Telecommunication Database.

Abstract. Sequential pattern mining is the process of finding the relationships between occurrences of sequential events, to find if there exists any specific order of the occurrences. The extraction of sequential pattern is not polynomial in time of execution. 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.

1 Introduction

Data mining in telecommunication databases is the process of finding novel, interesting, and useful patterns in huge data using methods such as: sequential pattern mining, association rule extraction, clustering, etc... The ongoing rapid growth of online telecommunication data due to the Internet and the widespread use of database technology have created an immense need for data mining methodologies. Sequential pattern mining was first introduced by Agrawal and Srikant [2, 3], it is the process of finding the relationships between occurrences of sequential events, to find if there exist any specific order of the occurrences. 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 Database is described in section 4. Experimental results are shown in section 5.

2 Sequential Patterns Mining

The problem of identifying sequential patterns is introduced by Agrawal and Srikant in [2, 3]. The goal is to find all subsequences from the given sets of transactions; this approach is useful when the data to be mined have some sequential nature to deal with databases that have a time-series characteristics. Sequential Pattern can be defined as follows.

Table 1: Telecommunication Database

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

Definition 1: Let $I = \{x_1 ... x_n\}$ 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 $s=(X_1 ... X_l)$ is an order list of itemsets, and an itemset X_i $(1 \leq i \leq l)$ 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. [2, 3]

3 Genetic Algorithm

Genetic Algorithm (GA), presented in [4, 6, 8, 17], 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.

4 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 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) [10].

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, 12, 17]. 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, 17]. Mutation can then be as in Figure 3.

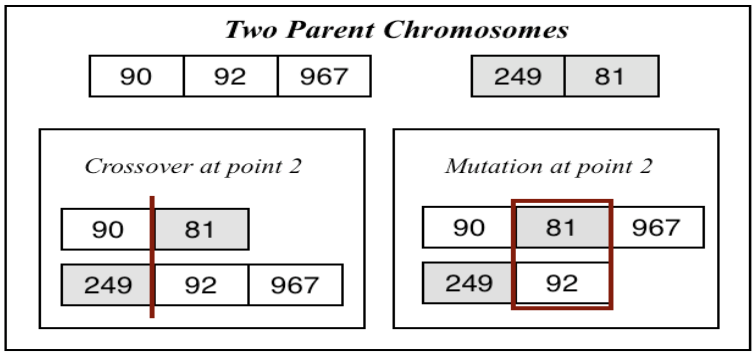

Fig. 3: Crossover and Mutation Example

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:

$$\mathbf{SIM(A \rightarrow C) = \min_{C_i \in C} \{(Confidence(A \mid C_i))^{\alpha}\} \times Support(AC)}$$

where ($\alpha \geq 0$) is a confidence priority that represents how important the frequency of the pattern is, C_i is sub-sequence 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
Confidence priority: α
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 ∈ P
    If A & C is given then
        Calculate the fitness F(i, A, C, α)
    Else If A is given then
        Calculate the fitness F(i, A, α)
    Else If C is given then
        Calculate the fitness F(i, C, α)
    Else
        Calculate the fitness F(i, α)
4. Mutate and crossover P.
5. IF (fitness ≥ minF)
    Select fittest rules from P
6. Set t = t +1
7. IF (t > G) then
    S = P
    Stop
  Else
    Go to Step 3
End

Fig. 4: Pseudo code of SPT-GA algorithm

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.

5. Experiment Results

Here, we present and analyze the results of the experiments on telecommunication database. All the experiments were performed on 1.86 GHz Intel® Centrino™ PC machine with 1.00 GB RAM, running on Windows XP platform. SPT-GA algorithm is written with MATLAB programming language. The user can tune population size (N), generations (G), confidence priority (α) and minimum fitness (minF).

A Telecommunication Database is taken from a Telecommunication Company and it has 1091 transactions and 60 country codes. The crossover probability used is 0.8, while the mutation is 0.001. The output of this experiment is a text file that includes the interesting rules that represent the most suitable telecommunication sequences. For example, the rule of Figure 5 told us that when country code 91 is called, that means (%40) of callers will call country code 92 afterwards.

91 → 92

Sequential Interestingness Measure (SIM) = 0.035134

Support = 0.089431

Confidence = 0.39286

Fig. 5: Sample of result

As we said, in SPT-GA algorithm, there are four parameters that must be determined by a user: number of generation (G), population size (N), confidence priority (α) and minimum fitness (minF). First experiment sets N = 20, α = 1 and minF = 0 with G = [20...200]. The second experiment sets G = 20, α = 1 and minF = 0 with N = [20...200]. The two experiments were done on two days of calls with 1091 transactions and 60 country codes.

According to the experiment, Figure 6 shows the time, in seconds, spent by the SPT-GA algorithm to extract the best rules while Figure 7 shows the average fitness of the final output. Both figures shows the result related to increasing the generations and population size.

Our tests showed that when the generation increases the time will be around each other but when the population increases the time will be increased. However, comparing the two experiments, as shown in Figure 6, GA takes less time when increasing the generation than increasing the population. Moreover, our tests also showed that when the generation and population increase, the best fitness will be around.

From the experiments, it is observed that increasing of generation will take less time than increasing of population size. But, either ways, GA does not take a long time; it is only a matter of seconds. In addition, increasing the generation and the population will not guarantee a large improvement of average fitness.

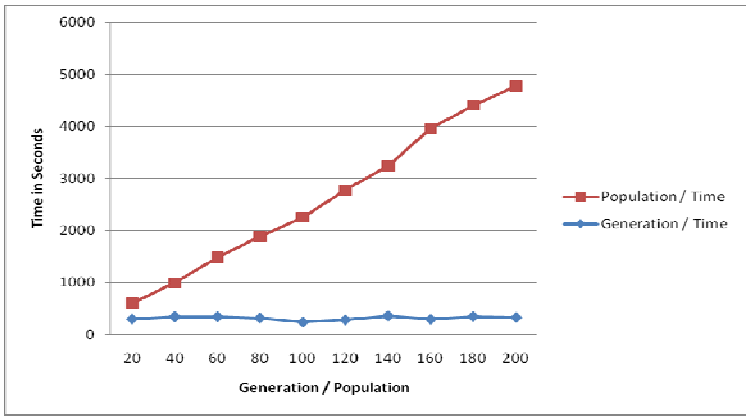

Fig. 6: Operation time related to generation and population size

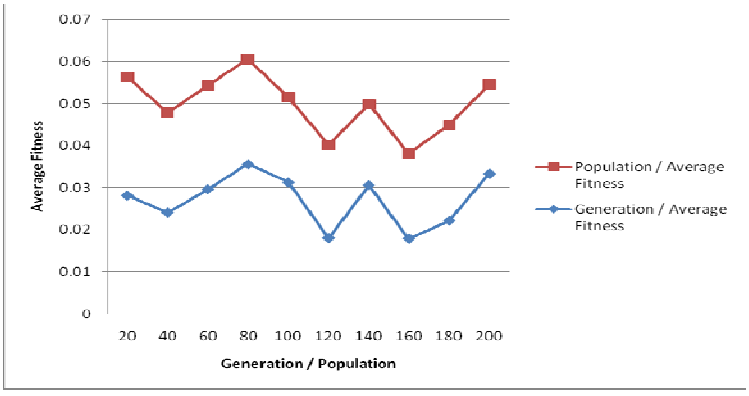

Fig. 7: Operation average fitness related to generation and population size

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] C. Antunes and A. L. Oliveira, "Sequential Pattern Mining Algorithms: Tradeoffs between Speed and Memory", Instituto Superior Técnico / INESC-ID.

[2] R. Agrawal and R. Srikant, "Mining Sequential Patterns", *IBM Almaden Research Center,* 650 Harry Road, San Jose, CA 95120-6099.

[3] R. Agrwal & R. Srikant, "Mining Sequential Patterns: Generalizations and Performance Improvements ", 1996, *IBM Almaden Research Center,* 650 Harry Road, San Jose, CA 95120.

[4] Saleh Al Kodhair, Mining Association Rules using Genetic Algorithm, 2008, Master Thesis.

[5] Jay Ay res , Johannes Gehr ke, Tomi Yiu, and Jason Flannick,"Sequential Pattern Mining using A Bitmap Representation", 2002, SIGKDD '02 Edmonton, Alberta, Canada.

[6] A. Freitas, "A Review of Evolutionary Algorithms for Data Mining",*Computing Laboratory*, University of Kent, UK.

[7] Alex A. Freitas, Data Mining and Knowledge Discovery with Evolutionary Algorithms, 2002, Springer-Verlag, Berlin.

[8] D. Goldberg, *Genetic Algorithms*, Addison Wesley, 1989. ISBN: 0-201-15767-5.

[9] T. C. Huang, "A Fuzzy Mining Process for Discovering Sequential Patterns", 1995, *Department of Information Management, National Central University,* Jungda Rd, Chung-Li City, Taoyuang, Taiwan , 32001, R.O.C.

[10] S. Sakurai, Y. Kitahara, and R. Orihara, "A Sequential Pattern Mining Method based on Sequential Interestingness", Fall.2008, International Journal of Computational Intelligence, pp.252-260.

[11] W. Spears, "Crossover or Mutation?", *Navy Center for Applied Research in Artificial Intelligence*, Naval Research Laboratory, Washington, D.C. 20375-5320.

[12]W. Spears and V. Anand, "A Study of Crossover Operators in Genetic Programming", Navy Center for Applied Research in AI, Washington, D.C 20375-5000.

[13]D. Taniar, *Data Mining and Knowledge Discovery Technologies,* 2008, New York: Hershey.

[14] J. Wook and S. Woo, "New Encoding/Converting Methods of Binary GA/Real-Coded GA", June 2005, *IEICE Trans*, Vol.E88-A, No.6, pp.1545-1564.

[15] S. Yuan-Wei Li, "Paper Survey on Sequential Pattern Data mining", December 14, 2004.

[16]Y. Zhou, "Study on Genetic Algorithm Improvement and Application", May 2006, Master thesis.

[17] www.en.wikipedia.org, Genetic algorithm, Wikipedia.