International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

Research Article

Fast and Robust Hybrid Particle Swarm Optimization and Tabu Search Algorithm for Web Data Association Rule Mining

Parmjeet Kaur[Á*], Usvir Kaur[Á] and Dheerendra Singh[B]

[Á]Department of Computer Science and Engineering Sri Guru Granth Sahib World University Fatehgarh Sahib, Punjab, India
[B]Shaheed Udham Singh College of Engineering and Technology, Tangori,Punjab, India

Accepted 05 Sept 2014, Available online 01 Oct 2014, **Vol.4, No.5 (Oct 2014)**

Abstract

Web search portals contains large amounts of web search data which includes keywords, links and other information. Web data association rules algorithm/s is the technique to deal with the web search data to produce the best results by analyzing the information in various combinations. In this paper, a novel web data association rule mining based hybrid algorithm called HPSO-TS-ARM has been proposed. This algorithms is based three well known high-level procedures: Particle Swarm Optimization, Tabu Search and Apriori Algorithm for Association Rule Mining. Where PSO will fetch the web search data in its optimized form, which is further computed by Tabu Search to prepare balance data arrangement followed by Association rule mining on processed web search data. The proposed algorithms have outperformed HBSO-TS and BSO-ARM on the basis of elapsed time and fitness function.

Keywords*: Association Rule Mining, Particle Swarm Optimization, Tabu Search, Apriori.*

Introduction

From recent couple of year explosive growth in amount of information or data come into notice. Data could be simple numerical figures, multimedia data, web data, text data and spatial data, is also being stored in files databases and data repositories. Finding ubiquitous model in a large amount of data is one of the key problems. For this particular reason data mining is attracted by information business and the world and is required to turn data into useful information and knowledge. Data mining is the process of fetching the desired information from large databases. Extracted information is used for different areas like business analysis, client maintenance, identifying the frauds, and scientific discoveries (Han, J. *et al*, 2006).

There exist diverse models of data mining such as classification, clustering, decision tree and neural networks from which association rule mining is also an important model. Association rules are used to extract the frequent patterns or casual structure among the set of items from given database. The pattern and rule discovered are based on the majority of commonly repeated items in dataset. Nowadays Association rule mining is broadly used in many different areas such as telecommunication networks, market and risk management, inventory control mobile mining, graph mining, educational mining, etc. The traditional application of association rule mining is market basket analysis that considers the buying habits of customers. Market basket analysis also examines that how the items are purchased by the customer. Typical example of super market with large number of transactions for association rule mining is:

$$\text{bread} \rightarrow \text{jam [sup=10\%, conf=80\%]},$$

10% support states that of customer purchase bread and jam simultaneously, and 80% confidence means 80% customers purchased bread also buy the jam.

Formal statement of association rule mining is defined as: Let $I = \{i1, i2,\ldots, im\}$ be a set of items. Let T be a set of transactions (the database), where each transaction t (a data case) is a set of items such that $t \subseteq I$. An association rule is an implication of the form, $X \rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \varnothing$. The rule $X \rightarrow Y$ holds in the transaction set T with confidence c if c% of transactions in T that support X also support Y. The rule has support s in T if s% of the transactions in T contains $X \cup Y$.

Support and confidence are two basic parameters of association rules to generate the interesting association rules. To discover the interesting association rules domain experts specifies the minimum support (minsup) and minimum confidence (minconf) from given set of transactions itemsets are called interesting if have greater support and confidence from minsup and minconf. To mine the association rules firstly find all the itemsets having specified threshold support, secondly generate association rules from these itemsets (Liu *et al*, 1999).

Literature Survey

(R. Aggarwal *et al*, 1993) shows an algorithm that generates all significant association rules between items in

*Corresponding author: **Parmjeet Kaur**

database. The algorithm includes buffer management and new estimation and pruning techniques. Experimental result shows the effectiveness by applying to large retailing company.

(Bing Liu et.al, 1999) proposes a novel technique to solve the rare item problem to resolve the combinatorial explosion. Proposed model allows the user to specify the multiple minimum supports to show the nature of items and their varied frequencies in database and found rare item rules without producing meaningless rules with frequent items .

(H. Drias *et al*, 2010) designed Bees Swarm Optimization algorithm named BSO-IR to explore the excessive number of documents to find the information desired by user. Experimental results shows better quality and runtime are compared between BSO and exact algorithms.

(Y. Djenouri *et al*, 2012) proposes two new Association Rule Mining algorithms based on Genetic Meta-heuristic and Bees Swarm Optimization. Classical algorithms are not capable to cope with large data in lesser respond time. Proposed model achieves better while compared with IARMGA, AGA in both fitness criterion and CPU time.

(Y. Djenouri *et al*, 2013) proposes a novel hybrid algorithm HBSO-TS. It based on two meta-heuristics that are Bees Swarm optimization and Tabu Search Experimental result shows better results as comparing to traditional approaches. They also planned the computation on GPU.

(Mohammed J. Zaki 1999) surveys the state of the art in parallel and distributed association-rule-mining algorithms and uncovers the field's challenges and open research problems also exposes that lot of exciting work remains to be done in system design, implementation, and deployment.

(T. Fukuda *et al*, 1996) discusses the data mining based on association rules for two numeric attributes and one Boolean attribute. They consider two classes of regions, rectangles and admissible (i.e. connected and x-monotone) regions. They had implemented algorithms for admissible regions, and constructed a system for visualizing the rules.

(D. Martens *et al*, 2010) surveys two popular domains: swarm intelligence and data mining. Data mining has been a popular academic topic for decades, swarm intelligence is new subfield of artificial intelligence, based on social behaviour that can be observed in nature, such as ant colonies, flocks of birds, fish schools and bee hives. Framework that categorizes the swarm intelligence based data mining algorithms into two approaches: effective search and data organizing.

(D. Karaboga *et al*, 2009) surveyed the algorithms based on intelligence in bee swarms and their applications. And surveyed algorithm VBA, ABC, BA developed for numerical problems can be expanded for combinatorial types problems by suitable modifications.

Experimental Design

We have proposed a new web data association rule mining algorithm with improved results than the existing ones. The new algorithm is called HPSO-TS-ARM which is based on a hybridization of Particle Swarm Optimization, Tabu Search and Association Rule Mining. The three major components of the proposed algorithm are the particle swarm optimization, tabu search and association rule mining.

Particle Swarm Optimization

As mentioned earlier, Particle Swarm Optimization simulates the activities of bird's flocking. Assume that a faction of birds are erratically searching foodstuff in a paricular area. Only single piece of foodstuff is existed in the area, which is being searched by the birds. All of the birds in the group are not aware about the actual position of the food. Hence, the question is "what is the best strategy to find the food?" The most useful technique is to find the bird closest to the foodstuff.

PSO cultured from the latter scenario is used to solve such optimization problems. Every solution is the bird in the area of interest or area of search. It is called a particle. Each particle carries a fitness value which is evaluated by the fitness function. The birds have different velocities which direct the flight of the particles. The particles take flight through the area of interest by succeeding the existing optimal particles.

PSO is initialized with a group of random particles or solutions, followed by the search for optimal value or optima by updating generations. Every solution is rationalized by finding two best values. First is optimal solution or fitness. It is also called pbest. Second values is followed by the particle swarm optimizer is the second best value, attained by any particle in the bird population. It is called global best or gbest. When computations are performed on a particle taking part in the population as per its neighbors in the topology, best values becomes local best of lbest. After obtain the optimal values, the solution updates velocity and positions of itself with following equation (i) and (ii).

v[] = v[] + CL1 * rand() * (pbest[] - present[]) + CL2 * rand() * (gbest[] - present[]) **(i)**

present[] = present[] + v[] **(ii)**

Where v[] denotes the particle velocity, present[] denotes current solution. pbest[] and gbest[] are fitness and global best respectively. rand () is a random number between (0,1). CL1, CL2 are learning factors (Nasser Lotfi *et al*).

Algorithm 1: The general Algorithm PSO

- *For each particle*
- *Initialize particle*
- *END*
- *Do*
- *For each particle*
- *Calculate fitness*
- *If fitness value is better than pBest in history*
- *set pBest equal to current value*
- *End*

- *Choose the pbest of all the particles as the gBest for all particle*
- *Calculate particle velocity according equation (a)*
- *Update particle position according equation (b)*
- *End*

Tabu Search

The main idea behind tabu search, which is a local search metaheuristic to rearrange the data in a particular manner to find the most effective solution called tabu list. Tabu search finds the best neighbors to maintain the tabu list L. The process runs the maximum number of iterations for a given condition to find the best solutions for the input data.(Y. Djenouri *et al*, 2013)

Algorithm 2 The General Algorithm TS

- *S Some Initial candidate solution.*
- *Best S.*
- *L { } a tabu list o f maximum length 1.*
- *I l.*
- **while** i < Max-Iter and not stop **do**
- *Enqueue S into L.*
- *S Best_neighbors(S).*
- *Alter Best if Quality(Best) < Quality(S).*
- **end while**

Apriori- Association Rule Mining

To perform web data association rule mining, we have used apriori algorithm. Apriori algorithm is primarily used with transactional databases. This algorithm analyzes the individual entities in the database and extends them to find the item/entity sets. These item sets are called frequent item sets, which are further used to determine to association rules to find the general trends in the database. Hence, this algorithm is easily adaptable to web data association rule mining (Han, J. *et al*, 2006).

Algorithm 2 The General Algorithm TS

Ck: Candidate item set of size k
Lk : frequent item set of size k
L1 = {frequent items};
for (k = 1; Lk !=; k++) **do**
- Ck+1 = candidates generated from Lk;
- **for each** transaction t in database **do**
increment the count of all candidates in Ck+1 that are contained in t;
endfor;
- Lk+1 = candidates in Ck+1 with min_support
endfor;
return k Lk;

Result Analysis

In our proposed algorithm HPSO-TS-ARM algorithm is developed using Particle Swarm Optimization, Tabu Search and ARM are combined together. It was a difficult task to choose the datasets and performance parameters. Various databases are tested under this research and the results are obtained in the form of elapsed time, elapsed time graph and fitness value. The result analysis confirms that HPSO-TS-ARM algorithm outperforms the HBSO-TS and BSO-ARM. We have performed all of the results on the the MATLAB v2011a installed on Windows PC with i3 processor and 2GBs of RAM. The PC was also equipped with 1GB GPU, but it was not utilized for the experimental computations. The proposed algorithm is taking almost half or less than half of the time than the existing HBSO-TS and BSO-ARM algorithm for web data association.

The datasets tested under this research project are Bolts, Sleep, Pollution and Basketball. These datasets are used because they were compact in size and the algorithm we developed is most applicable on the compact sized datasets. According to our experimental design analysis, we found that this algorithm will also perform better on the large sized datasets.

Table 1: Data Sets Description

Data set Name	Transaction Size	Item Size
Bolts	40	8
Sleep	56	8
Pollution	60	16
Basket Ball	96	5
Connect	607	43

In order to perform experimentations, several realtime scientific databases which are frequently used indata mining conununity like Frequent and Mining Data set Repository [13] , Bilkent University Function Approximation Repository [12] , were consulted.

Table 2: Our Approach to Other Approaches w.r.t. to Fitness

Data set	HBSO--TS	BSO-ARM	HPSO-TS-ARM	PSO-ARM
Bolts	1.0	1.0	1.477239	1.1015393
Pollution	1.0	1.0	1.520532	1.3817868
Basket Ball	0.97	0.97	1.524602	1.496168
Sleep	1.0	1.0	1.533414	1.250272
Connect	0.50	0.26	1.503142	1.426236

The main parameter tested is fitness function and overall time taken by algorithm and personal best value for each iteration of particle swarm optimization.

Fitness Function = X*confidence(s)+Y*support(s)

Association Rule Mining problem is to find all the rules satisfying minimum support and minimum confidence X and Y are weighted parameters set according to support and confidence's importance.

In HBSO-TS algorithm, BSO and TS managed together. Table 2 summarizes all the outcomes obtained by executing HPSO-TS, PSO-ARM, in terms of the fitness function described above. With comparison HPSO-TS and PSO-ARM shows better performance as compared to

HBSO-TS and BSO-ARM. The aim is to maximize this function.

Table 3: Elapsed Time by HPSO-TS-Arm

Dataset Name	HPSO-TS-ARM	PSO-ARM
Bolt	14.3026	13.7676
Pollution	31.2269	31.1480
Basketball	16.6421	16.5710
Sleep	21.6184	21.3338
Connect	840.3135	756.0325

Table 3 shows the CPU time of the suggested algorithms HPSO-TS and PSO-ARM. PSO-ARM outperforms from all other algorithms in all large data sets. This can be explained by the fact that: in HPSO-TS, we use tabu search for the neighborhood computation. However, in PSO-ARM, we apply just a simple search of neighbors.

Conclusion

The proposed algorithm named as HPSO-TS-ARM algorithm. HPSO-TS-ARM has outperformed the existing HBSO-TS and BSO-ARM algorithms in terms of fitness value. The performance parameter of elapsed time has been tested on the same databases which yields good and acceptable results. An equal or better fitness function values with less elapsed time is a sign that proposed algorithm will also perform better results on large datasets. In future, this algorithm will be tested with more datasets and will be compared with the HBSO-TS and BSO-ARM in the terms of other performance parameters also. Its performance will be also tested and compared with other similar algorithms on the basis of various datasets and more performance parameters. Because the proposed algorithm is proved to be useful for the web data association rule mining, it will be enhanced to perform better than the proposed algorithm by combining it with different algorithms to develop new algorithms using new algorithmic combinations or newly developed algorithms.

References

Han, J., Kamber, M . , & Pei, J., (2006), Data mining: concepts and techniques. *Morgan kaufmann*.

Liu, B., Hsu, W. and Ma, Y., (August 1999), Mining association rules with multiple minimum supports, *KDD '99 Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, ISBN: 1-58113-143-7, pp 337-341, .

R. Agrawal, T. Imielinski and A. Swami, (May 1993), Mining association rules between sets of items in large databases, *SIGMOD '93 Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, ISBN: 0-89791-592-5, 22(2), pp 207-216,

Habiba Drias, Hadia Mosteghanemi, (September 2010), Bees Swarm Optimization based Approach for Web Information Retrieval, *International Conference on Web Intelligence and Intelligent Agent Technology IEEE/WIC/ACM*, ISBN: 978-0-7695-4191-4, 1, pp 6-13, .

Y. Djenouri, H. Drias, Z. Habbas, H. Mosteghanemi, (December 2012), Bees Swarm Optimization for Web Association Rule Mining, *International Conferences on Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM*, ISBN: 978-1-4673-6057-9, 3, pp 142-146

Youcef Djenouri, Habiba Drias, Amine Chemchem, (August 2013), A Hybrid Bees Swarm Optimization and Tabu Search Algorithm for Association Rule Mining, *World Congress on Nature and Biologically Inspired Computing (NaBIC) IEEE*, ISBN: 978-1-4799-1414-2, pp 120-125

Mohammed J. Zaki, (December 1999), Parallel and Distributed Association Mining: A Survey, *Concurrency IEEE*, ISSN: 1092-3063, 7(4) , pp 14-25

Takeshi Fukuda, Yasuhiko Morimoto, Shinichi Morishita, Takeshi Tokuyama, (June 1996), Data Mining Using Two-Dimensional Optimized Association Rules: Scheme, Algorithms, and Visualization, *ACM SIGMOD international conference on Management of data*, 25(2), pp 13-23

David Martens, Bart Baesens, Tom Fawcett, (January 2011), Editorial survey: swarm intelligence for data mining, *Machine Learning Springer*, ISSN: 0885-6125, 82(1), pp 1-42

Dervis Karaboga, Bahriye Akay, (June 2009), A survey: algorithms simulating bee swarm intelligence, *Artificial Intelligence Review Springer*, 31, pp 61 – 85.

Nasser Lotfi, Jamshid Tamouk, Mina Farmanbar, 3-SAT Problem: A New Memetic-PSO Algorithm, available at http://arxiv.org/abs/1306.5070

Guvenir, H. A ., &Uysal, I.(2000), Bilkent university functionapproximation repository. 201 2-03 - 1 2] . http ://funapp. cs.bilkent. edu. tr/DataSets.

Goethals, B., &Zaki, M. J. (2003), Frequent itemset miningimplementations repository. This site contains a wide-variety ofalgorithms for mining frequent, closed, and maximal itemsets,http://fimi. cs. helsinki. fi..