Population-based metaheuristics for Association Rule Text Mining

Iztok Fister Jr.
University of Maribor
Maribor, Slovenia
iztok.fister1@um.si

Suash Deb
Victoria University, Decision Sciences and Modeling Program
Melbourne, Australia
IT & educational Consultant, Ranchi, Jharkhand, India
suashdeb@gmail.com

Iztok Fister
University of Maribor
Maribor, Slovenia
iztok.fister@um.si

ABSTRACT

Nowadays, the majority of data on the Internet is held in an unstructured format, like websites and e-mails. The importance of analyzing these data has been growing day by day. Similar to data mining on structured data, text mining methods for handling unstructured data have also received increasing attention from the research community. The paper deals with the problem of Association Rule Text Mining. To solve the problem, the PSO-ARTM method was proposed, that consists of three steps: Text preprocessing, Association Rule Text Mining using population-based metaheuristics, and text postprocessing. The method was applied to a transaction database obtained from professional triathlon athletes' blogs and news posted on their websites. The obtained results reveal that the proposed method is suitable for Association Rule Text Mining and, therefore, offers a promising way for further development.

CCS CONCEPTS

• **Theory of computation** → **Evolutionary algorithms**; • **Information systems** → *Expert systems*;

KEYWORDS

association rule text mining, natural language processing, particle swarm optimization, optimization, triathlon

1 INTRODUCTION

Stochastic population-based nature-inspired metaheuristics offer a very effective way for Association Rule Mining (ARM). They are stochastic in their nature, and do not discover association rules using an exhaustive search as the other classical methods do. Numerous population-based methods exist for ARM that were developed in the past years. According to the literature review, most of the existing methods are intended for mining categorical features that are stored in transaction databases. On the other hand, some methods also exist that can deal with numerical data. Actually, this kind of mining is also called Numerical Association Rule Mining (NARM) [2].

Some examples of the ARM based on stochastic population-based nature-inspired algorithms include methods like MODENAR [1], ARMGA [12], and ARM-DE [7]. MODENAR is an example of a very efficient multi-objective Differential Evolution for mining numeric association rules, while ARMGA is a Genetic Algorithm for discovering association rules, where there is no necessary to specify the minimum support and minimum confidence by users. On the other hand, ARM-DE is a new approach for NARM problems, based on Differential Evolution.

In contrast, there is a lack of works for discovering the association rules in text [11]. Association Rule Text Mining (ARTM) results in many practical applications, e.g., for building text classifiers [3, 13]. On the other hand, methods of generating knowledge discovery in text mining using association rule extraction are helpful for the users who are able to find accurate and important knowledge quicker than browsing through the text manually [10]. ARTM is also very interesting in the Medical domain [4]. In the paper [8], association rules are used to derive a feature set from pre-classified text documents. Interestingly, the authors of a survey paper [11] discovered that the Apriori algorithm is also suitable for ARTM, and utilized mostly in various domains, especially in the domain of Medical Care [11].

To the authors‘ knowledge, no methods exist for ARTM that are based fully on stochastic population-based nature-inspired metaheuristics. In this paper, we tackle the problem of ARTM using Particle Swarm Optimization (PSO) [9]. We would like to get some answers on the following questions:

- Are stochastic population-based nature-inspired metaheuristic algorithms suitable for ARTM?
- Can we find a viable interpretation of discovered association rules in text?
- Is there a bright way of developing these algorithms in the future?

The paper is structured as follows: Sec. 2 depicts some features of ARTM, as well as outlines the main differences between conventional ARM tasks, while Sec. 3 deals with a detailed description of the proposed method. Experiments and results are presented in

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
ISMSI '20, March 21–22, 2020, Thimphu, Bhutan
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7761-4/20/03…$15.00
DOI:https://doi.org/10.1145/3396474.3396493

Sec. 4, while the paper is concluded with Sec. 5, where directions for the future work are outlined.

2 BASIC INFORMATION

This section is focused on the background information necessary for understanding the subject that follows. At first, the problem of discovering association rules is illustrated in detail, followed by the basics of the PSO algorithm.

2.1 Association Rule Mining

ARM can formally be defined as follows: Let us assume a set of objects $O = \{o_1, \ldots, o_M\}$ and transaction dataset $T_D = \{T\}$ are given, where each transaction T is a subset of objects $T \subseteq O$. Then, an association rule is defined as an implication:

$$X \Rightarrow Y, \tag{1}$$

where $X \subset O, Y \subset O$, and $X \cap Y = \emptyset$. In order to estimate the quality of mined association rule, two measures are defined: A confidence and a support. The confidence is defined as:

$$conf(X \Rightarrow Y) = \frac{n(X \cup Y)}{n(X)}, \tag{2}$$

while the support as:

$$supp(X \Rightarrow Y) = \frac{n(X \cup Y)}{N}, \tag{3}$$

where function $n(.)$ calculates the number of repetitions of a particular rule within D_T, and N is the total number of transactions in D_T. Let us emphasize that two additional variables are defined, i.e., the minimum confidence C_{min} and the minimum support S_{min}. These variables denote a threshold value limiting the particular association rule with lower confidence and support from being taken into consideration.

2.2 Basics of the PSO algorithm

Particle Swarm Optimization (PSO) is a member of an SI-based algorithm family that was developed by Eberhard and Kennedy in 1995 [9]. It is inspired by the social behavior of bird flocking and fish schooling. This algorithm works with a swarm (i.e., a population) of particles representing candidate solutions $\mathbf{x}_i^{(t)}$ of the problem to be solved. The particles fly virtually through the problem space, and are attracted by regions reached with food. When the particles are located in the vicinity of these regions, they are rewarded with the better values of fitness function by the algorithm.

Interestingly, the PSO algorithm exploits usage of an additional memory, where the particle's personal best $\mathbf{p}_i^{(t)}$ as well as the swarm's global best $\mathbf{g}^{(t)}$ locations in the search space are saved. In each time step t (i.e., generation), all particles change their velocities $\mathbf{v}_i^{(t)}$ towards their personal and global best locations according to the following mathematical formula:

$$\begin{aligned} \mathbf{v}_i^{(t+1)} &= \mathbf{v}_i^{(t)} + C_1 \cdot \text{rand}(0,1) \cdot \left(\mathbf{g}^{(t)} - \mathbf{x}_i^{(t)}\right) + C_2 \cdot \text{rand}(0,1) \cdot \left(\mathbf{p}_i^{(t)} - \mathbf{x}_i^{(t)}\right), \\ \mathbf{x}_i^{(t+1)} &= \mathbf{x}_i^{(t)} + \mathbf{v}_i^{(t)}, \end{aligned} \tag{4}$$

where C_1 and C_2 present social and cognitive weights typically initialized to 2, and rand(0, 1) is a random value drawn from uniform distribution in the interval $[0, 1]$.

The pseudo-code of the original PSO is illustrated in Algorithm 1,

Algorithm 1 The original PSO algorithm

```
1:  procedure ParticleSwarmOptimization
2:      t ← 0;
3:      P(t) ←Initialize;                    ▷ initialization of population
4:      while not TerminationConditionMeet do
5:          for all x_i^(t) ∈ P(t) do
6:              f_i^(t) = Evaluate(x_i^(t));    ▷ evaluation of candidate
7:              if f_i^(t) ≤ f_best_i^(t) then
8:                  p_i^(t) = x_i^(t); f_best_i^(t) = f_i^(t);
9:              end if                       ▷ preserve the local best solution
10:             if f_i^(t) ≤ f_best^(t) then
11:                 g^(t) = x_i^(t); f_best^(t) = f_i^(t);
12:             end if                       ▷ preserve the global best solution
13:             x_i^(t) = Move(x_i^(t));    ▷ move candidate w.r.t. Eq. (4)
14:         end for
15:         t = t + 1;
16:     end while
17: end procedure
```

from which it can be seen that the PSO is distinguished from the classical EAs by three specialties:

- does not have survivor selection,
- does not have the crossover operator,
- the mutation operator is replaced by the move operator changing each element of particle $\mathbf{x}_i^{(t)}$ with the probability of mutation $p_m = 1.0$,
- does not have the selection operator.

Let us mention that the selection is implemented in the PSO implicitly, i.e., by improving the personal best solution permanently. However, when this improving is not possible anymore, the algorithm gets stuck in the local optima.

3 PROPOSED METHOD

At a glance, there is no simple analogy between text and market basket analysis that is a good example of an ARM task. Items in a market basket are self-contained, and well organized in transaction databases, where they are easy to employ by algorithms for ARM. However, working with pure text is a totally different story, because here, information is hidden in unstructured text. As stated by Brownlee [5], "a problem with modeling text is that it is messy, and techniques like machine learning algorithms prefer well defined fixed-length inputs and outputs".

Consequently, the unstructured text needs complex preprocessing treatment, where those unstructured data must be transformed into a structured transaction database. Obviously, the transaction database is appropriate for discovering the association rules using stochastic population-based nature-inspired algorithms. For the needs of this study, the PSO algorithm was employed for solving the ARTM (i.e., PSO-ARTM).

In general, the proposed PSO-ARTM consists of the following three phases:

- text preprocessing,
- optimization,

- postprocessing.

In the remainder of the paper, the aforementioned phases are presented in detail.

3.1 Text preprocessing

The purpose of this phase is to generate a transaction database from the raw data obtained from triathlon athletes' blogs and news posted on their websites, and consists of three steps, as follows:

- tokenizing,
- stop word removal,
- term frequencies' calculation.

Punctuation marks are removed in the first step. As a result, only words delimited by space remain in the document. Some words, like definite and indefinite articles (e.g., the, a, an), connective words (e.g., and, also, then), conjunctions (e.g., but, when, because), and verbs (e.g., is, done), represent the so-called stop words, and must be removed in the second step. The result of this step is a sequence of terms. The terms undergo term frequency calculation, where occurrences are not only determined, but also weighted. Thus, a Term Frequency/Inverse Term Frequency (TF/ITF) weighting scheme is used that penalizes the rare occurring terms with higher weights.

The TF/ITF weighting scheme is defined as follows: For a given term z_j, for $j = 1, \ldots, M$, occurring in document d_i, for $i = 1, \ldots, N$, the term frequency is expressed as:

$$TF_{i,j} = \frac{n(d_i, w_j)}{|d_i|}, \tag{5}$$

where $n(d_i, w_j)$ denotes the number of occurrences of term w_j in document d_i, and $|d_i|$ is the total number of terms in document D_i. On the other hand, the inverse term frequency is expressed as:

$$ITF_j = \left| \log \frac{n(d|w_j)}{N} \right|, \tag{6}$$

where $n(d|w_j)$ denotes the number of document d containing term w_j, and N is the total number of documents.

Furthermore, the weighted frequency of term z_j in document d_i is represented as a vector of weights $\mathbf{w}_i = \{w_{i,1}, \ldots, w_{i,n}\}$, where each element $w_{i,j}$ is expressed as:

$$w_{i,j} = TF_{i,j} \cdot ITF_j, \quad \text{for } j = 1, \ldots, n. \tag{7}$$

Finally, the transaction database is generated from the relevant documents by moving each vector $\mathbf{w}_i$, representing weighted frequencies for all terms in the corresponding document, to a transaction in D_T. In this way, the transaction database is very similar to the market basket, except that the weighted frequencies are put into D_T instead of the value of one.

3.2 Optimization

An ARTM problem can be defined formally as follows: Let us assume a set of documents $D = \{d_1, \ldots, d_N\}$ and set of terms $Z = \{z_1, \ldots, z_M\}$ are given, where N denotes the maximum number of documents, and M the maximum number of terms, respectively, to which also the matrix of weights $\mathbf{W}$ of dimension $N \times M$ is assigned, where each element $w_{i,j}$ represents a frequency weight of term z_j in document d_i, calculated according the TF-ITF weighting scheme. Then, the task of optimization is to select the binary vector $\mathbf{y} = (y_1, \ldots, y_M)^T$, determining the presence or absence of the corresponding term in the solution, such that the scalar product

$$AWS = \sum_{j=1}^{M} \sum_{i=1}^{N} w_{i,j} \cdot y_j \tag{8}$$

subject to

$$\sum_{j=1}^{M} y_j \leq K, \tag{9}$$

is maximum. Let us mention that variable K denotes the maximum number of terms in association rule. Actually, the selected elements of vector $\mathbf{y}$ form the set $Y = \{y_j | y_j = 1, \text{ for } j = 1, \ldots, M\}$ that is a subset of Z, in other words $Y \subset Z$. Let us notice that the values of vector are initially set to zero.

The problem is solved using the PSO algorithm, that needs the following modifications referred to three components of the algorithm: (1) Representation of individuals, (2) Genotype-phenotype mapping, and (3) Evaluation function. In the remainder of the paper, the aforementioned modifications are discussed in detail.

3.2.1 Representation of individuals. The candidate solutions in the PSO algorithm are represented as real-valued vectors

$$x_i^{(t)} = (x_{i,1}^{(t)}, \ldots, x_{i,K}^{(t)}, x_{i,K+1}^{(t)}), \quad \text{for } i = 1, \ldots, Np, \tag{10}$$

where $x_{i,j}^{(t)} \in [0, 1]$ for $j = 1, \ldots, K$ encodes the selected terms in association rule, K the maximum number of terms in the particular association rule, Np is the population size, and t the generation counter. However, the last element of vector $x_{i,K+1}^{(t)}$ determines the cut point between antecedent and consequence in the rule.

3.2.2 Genotype-phenotype mapping. Each solution encodes a definite association rule in the genotype space that needs to be mapped into phenotype space before evaluation. This mapping is performed according to the following equation:

$$y_j^{(t)} = \begin{cases} 1, & \text{if } \left\lfloor \frac{x_{i,j}^{(t)}}{K} \right\rfloor = j, \\ 0, & \text{otherwise,} \end{cases} \quad \text{for } j = 1, \ldots, K. \tag{11}$$

The cut point denoted by the last element $x_{i,K+1}^{(t)}$ is calculated according to the following equation:

$$cp = \left\lfloor \frac{x_{i,j}^{(t)}}{K - 1} \right\rfloor, \tag{12}$$

which selects one of the $K - 1$ cut points between K elements of the association rule.

Let us mention that the mapping in Eq. (11) does not ensure the injective mapping of each element of a vector to the specific term. This means that it is possible that more elements of vector $x_{i,j}^{(t)}$ are mapped to the same term $y_j^{(t)}$. In this case, the number of terms in the association rule is less than K, but this is admissible according to Eq. (9).

As a result of genotype-phenotype mapping, the association rule $X \Rightarrow Y$ is obtained, where the cut point delineates the antecedent part from the consequence.

3.2.3 Evaluation function. Evaluation function in the PSO algorithm for ARTM estimates the quality of the association rule according to the following equation:

$$f(X \Rightarrow Y) = \frac{\alpha \cdot supp(X \Rightarrow Y) + \beta \cdot conf(X \Rightarrow Y) + \gamma \cdot AWS}{\alpha + \beta + \gamma}, \quad (13)$$

where α, β, and γ represent the weights of the particular terms. In our study, the values of weights are fixed to one. This means that each term in Eq. (13) is treated equally. Let us mention that the value of the fitness function needs to be maximized, in other words $f^*(X \Rightarrow Y) = \max f(X \Rightarrow Y)$.

3.3 Postprocessing

The aim of this step is the interpretation of the results. Typically, the results of the Association Rule Mining are illustrated in tabular form. This way of presentation is also applied in our study.

4 EXPERIMENTS AND RESULTS

Experiments are based on the dataset that represents the blog/website posts of various world triathletes. Interestingly, almost all of the observed websites are organized as blogs. WordPress is the most popular platform, while a lot of blogs are hosted on BlogSpot or Wix.com platforms. Datasets were scraped and extracted automatically into a transaction database. In summary, the transaction database in the experiments consists of 4,271 feeds [1].

In this database, the following elements of RSS feeds are included:

- title,
- description,
- link,
- date, and
- content.

It is worth mentioning that some experiments were already conducted on the initial version of the database, while the initial findings were published in paper [6]. In that paper, the active lifestyle of triathon athletes was analyzed, where deep analytic methods were proposed for analyzing these feeds. The results of the analysis were presented as social networks that serve as a basis for the decision-making process, from which the real triathlon trainer can extract some characteristics and information about the triathlon athlete's lifestyle.

However, the transaction database also has some limitations, referring especially to cleaning and removing some inappropriate feeds. It seems that some websites were probably hacked, and some inappropriate bots posted, which caused the improper and strange content. Obviously, such feeds were also removed from the database. Similarly, the same procedure was applied also by preparing the database in the present study.

The parameter setting of the PSO for ARTM is proposed in Table 2, from which it can be seen that the population size was limited to $Np = 200$, the termination condition to the maximum number of fitness function evaluations $nFEs = 10,000$, and there were five independent runs of the PSO algorithm conducted.

Interestingly, the PSO algorithm is limited to process only the 1,000 most frequently occurring terms from the transaction database. The 15 of these are depicted in the histogram in Fig. 1, where

[1] updated in December 2019

Table 1: Parameter settings of the PSO algorithm

Parameter	Abbreviation	Value
Population size	Np	200
Social component	C_1	2.0
Cognitive component	C_2	2.0
Inertia weight	w	0.7
Number of independent runs	NUM_RUNS	5
Number of fitness evaluations	$nFEs$	10,000

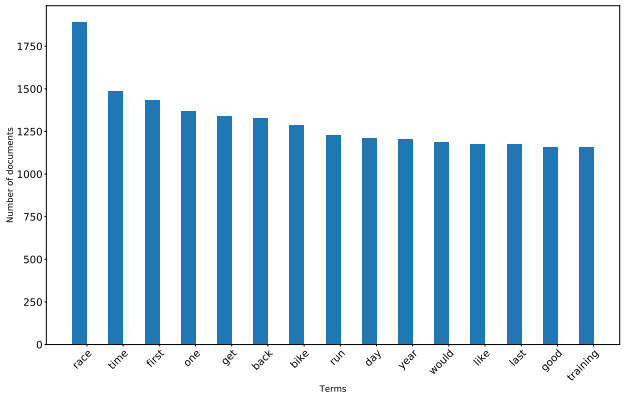

Figure 1: Histogram of the most frequent words in a database.

Table 2: Statistics of rules found on different K settings

K	5	6	7	8
No. Rules	4594	1947	282	273
Avg Ant.	1.693	2.776	2.148	2.520
Avg Cons.	2.306	2.223	3.851	4.479

the terms are presented according to the number of occurrences. As can be seen from the histogram, the most frequently used term in the posted feeds is "race".

4.1 The results

The association rules are mined according to the number of terms K, that was varied in the interval $K \in [5, 8]$ in steps of one. In this way, four instances were obtained, while the statistics of the mined rules are presented in Table 2, from which it is evident that the maximum number of rules is mined by the lower number of the parameter K. Interestingly, the average number of antecedents and consequences do not follow these trends, because the aforementioned values for $K = 5$ are lower than the same for $K = 8$.

Finally, the five more interesting association rules mined using the PSO algorithm for ARTM are illustrated in Table 3. The meaning of these mined rules can be interpreted as follows: The first rule is referring to the relationship between cycling and running in a

Table 3: Examples of some interesting solutions found by the proposed approach.

Rule	Antecedent	Consequence
1	amazing ∧ ride ∧ next	running ∧ hurt ∧ hopefully ∧ fine
2	championship ∧ skills	race ∧ technical
3	great	year ∧ news ∧ mph ∧ course ∧ start ∧ always
4	one ∧ race	hard ∧ bike ∧ finish ∧ week ∧ amount
5	triathlete ∧ people	right ∧ family ∧ sprint

triathlon. From this rule, it is evident that, if athletes ride the cycles well, they are also good in running. The second rule asserts that the athlete who wants to be the champion, needs also to train some technical skills. The third rule is added intentionally by us to show that the interpretation of some rules is not so easy. The fourth rule is referring to the hard cycling courses, while the last one exposes the importance of family support in triathlon racing.

5 CONCLUSION

This paper deals with the problem of Association Rule Text Mining, as well as proposing a new method for solving this problem using the stochastic population-based nature-inspired metaheuristics. The proposed PSO-ARTM method was evaluated on a transaction database consisting of the website/blog feeds of many world triathlon athletes. The results suggest that we can infer some useful information from these blogs.

Actually, we posted three questions at the beginning of the experimental study, and revealed that: (1) The stochastic population-based nature-inspired metaheuristics are suitable tools for solving ARTM, (2) The interpretation of the discovered associated rules is not trivial, especially for rules with either one antecedent or one consequence, and (3) There is a bright way for the future development of these algorithms.

The last finding is justified as follows: Indeed, the future opens a broad path for a further enhancement of the method. For instance, the development of a better evaluation function or development of a multi-objective version of these methods might be one of the first steps for its improvement. Nevertheless, the evaluation of this method on the other well-known transaction databases could also represent a big challenge for the future.

6 ACKNOWLEDGMENT

I. Fister Jr. acknowledge the financial support from the Slovenian Research Agency (Research Core Funding No. P2-0057). I. Fister acknowledge the financial support from the Slovenian Research Agency (Research Core Funding No. P2-0041).

7 REFERENCES

[1] Bilal Alatas, Erhan Akin, and Ali Karci. 2008. MODENAR: Multi-objective differential evolution algorithm for mining numeric association rules. *Applied Soft Computing* 8, 1 (2008), 646–656.

[2] Elif Varol Altay and Bilal Alatas. 2019. Performance analysis of multi-objective artificial intelligence optimization algorithms in numerical association rule mining. *Journal of Ambient Intelligence and Humanized Computing* (2019), 1–21.

[3] M-L Antonie and Osmar R Zaiane. 2002. Text document categorization by term association. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.* IEEE, 19–26.

[4] Svetla Boytcheva. 2018. Indirect Association Rules Mining in Clinical Texts. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*. Springer, 36–47.

[5] Jason Brownlee. 2017. Machine learning mastery. *URL: http://machinelearningmastery.com/gentle-introduction-bag-words-model/* (2017).

[6] Iztok Fister Jr., Dušan Fister, Samo Rauter, Uroš Mlakar, Janez Brest, and Iztok Fister. 2017. Deep analytics based on triathlon athletes' blogs and news. In *23rd International Conference on Soft Computing*. Springer, 279–289.

[7] Iztok Fister Jr., Andres Iglesias, Akemi Galvez, Javier Del Ser, Eneko Osaba, and Iztok Fister. 2018. Differential Evolution for Association Rule Mining Using Categorical and Numerical Attributes. In *International Conference on Intelligent Data Engineering and Automated Learning*. 79–88.

[8] SM Kamruzzaman, Farhana Haider, and Ahmed Ryadh Hasan. 2010. Text classification using association rule with a hybrid concept of naive Bayes classifier and genetic algorithm. *arXiv preprint arXiv:1009.4976* (2010).

[9] J Kennedy and R Eberhart. 1995. Particle swarm optimization. In *Neural Networks, 1995. Proceedings., IEEE International Conference on*, Vol. 4. IEEE, 1942–1948.

[10] Manasi Kulkarni and Sagar Kulkarni. 2016. Knowledge Discovery in Text Mining using Association Rule Extraction. *International Journal of Computer Applications* 143, 12 (2016), 30–35.

[11] J Manimaran and T Velmurugan. 2013. A survey of association rule mining in text applications. In *2013 IEEE International Conference on Computational Intelligence and Computing Research*. IEEE, 1–5.

[12] Hamid Reza Qodmanan, Mahdi Nasiri, and Behrouz Minaei-Bidgoli. 2011. Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence. *Expert Systems with applications* 38, 1 (2011), 288–298.

[13] Chowdhury Mofizur Rahman, Ferdous Ahmed Sohel, Parvez Naushad, and SM Kamruzzaman. 2010. Text classification using the concept of association rule of data mining. *arXiv preprint arXiv:1009.4582* (2010).