Population-based metaheuristics for Association Rule Text Mining

Iztok Fister Jr., Suash Deb, Iztok Fister

University of Maribor

iztok.fister1@um.si

Introduction

- Stochastic population-based nature-inspired metaheuristics offer a very effective way for Association Rule Mining (ARM)
- Most of the existing methods are intended for mining categorical features that are stored in transaction databases
- In contrast, there is a lack of works for discovering the association rules in text (ARTM)
- 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

Research 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?

Particle swarm optimization

```
procedure ParticleSwarmOptimization
    t ← 0;
    P^(t) ← Initialize;                        ▷ initialization of population
    while not TerminationConditionMeet do
        for all x_i^(t) ∈ P^(t) do
            f_i^(t) = Evaluate(x_i^(t));        ▷ evaluation of candidate
            if f_i^(t) ≤ f_{best_i}^(t) then
                p_i^(t) = x_i^(t); f_{best_i}^(t) = f_i^(t);
            end if                             ▷ preserve the local best solution
            if f_i^(t) ≤ f_best^(t) then
                g^(t) = x_i^(t); f_best^(t) = f_i^(t);
            end if                             ▷ preserve the global best solution
            x_i^(t) = Move(x_i^(t));            ▷ move candidate
        end for
        t = t + 1;
    end while
end procedure
```

Proposed method

Text preprocessing

$$TF_{i,j} = \frac{n(d_i, w_j)}{|d_i|}, \tag{1}$$

$$ITF_j = \left|\log \frac{n(d|w_j)}{N}\right|, \tag{2}$$

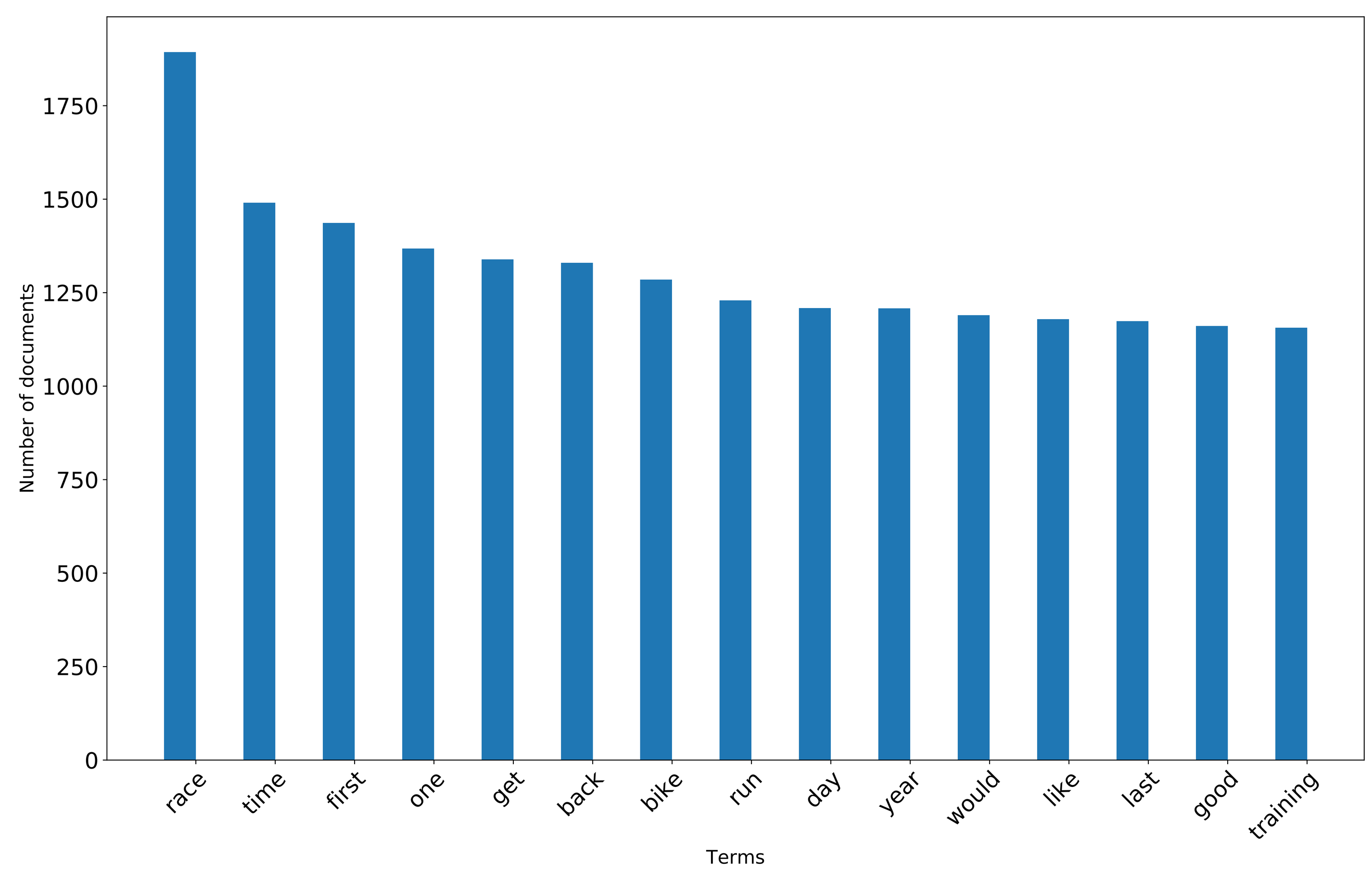
Optimization

$$AWS = \sum_{j=1}^{M} \sum_{i=1}^{N} w_{i,j} \cdot y_j \tag{3}$$

$$f(X \Rightarrow Y) = \frac{\alpha \cdot supp(X \Rightarrow Y) + \beta \cdot conf(X \Rightarrow Y) + \gamma \cdot AWS}{\alpha + \beta + \gamma}, \tag{4}$$

Test data

Dataset that represents the blog/website posts of various world triathletes.



(Histogram of the most frequent words in a database.)

Results

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

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

Conclusions and Future Work

- The stochastic population-based nature-inspired metaheuristics are suitable tools for solving ARTM
- The interpretation of the discovered associated rules is not trivial, especially for rules with either one antecedent or one consequence
- There is a bright way for the future development of these algorithms.