The Traveling Analyst Problem, Orienteering applied to exploratory data analysis

Alexandre Chanson, Nicolas Labroche, Patrick Marcel, Vincent T'Kindt

Université de Tours, Laboratoire d'Informatique Fondamentale et Appliquée (EA 6300)
ERL CNRS 7002 ROOT, Tours, France
{nom.prenom}@univ-tours.fr

Mots-clés : *Orienteering, Mathematical model, Exploratory Data Analysis*

1 Introduction

We introduce a variation of the Orienteering Problem stemming from the Database community [3], and its associated mathematical model with a polynomial number of variables and constraints. Exploratory data analysis (EDA) [7], one of the main tasks of data workers, is a tedious and time consuming process particularly challenging for novice users or data enthusiasts unfamiliar with querying languages. While several graphical commercial tools are capable of producing EDA sessions (Tableau, Saiku, PowerBI, etc), they still require a heavy involvement from the user. Recently ad-hoc solutions for generating automatically EDA sessions were proposed, without formally defining the problem [11, 4, 2].

We formulate the automated construction of EDA sessions, i.e., sequences of interesting database queries, as a variation of the orienteering problem. The orienteering problem belongs to the broad category of routing problems with profits. It is inspired by one of the many orienteering sports where a contestant must find and reach control points, obtaining a reward for each point visited [9]. Adding a service time would mean that the contestant performs a time consuming action to obtain the reward from control points, which does not change the problem formulation as mentioned by [10] since one can merge service time and travel time.

However, in our particular application, the distance is not analogous with time but with a distance between database queries (e.g., [1]). It must be modeled separately from the service time, since the latter corresponds to the run time of the queries on the given database system. Furthermore the score associated with each query is a specific type of reward known as interestingness [5] and will be referred as such from now on. Lastly, the number of queries to choose from is very large even for small database instances [3].

2 Mathematical model

Given a set of n queries Q over a database instance I, a distance matrix C where $c_{i,j}$ denotes the distance from q_i to q_j, an execution time t_i for each query and an interestingness measure v_i, we aim to find a sequence of queries in Q which total execution time is below a given budget, the total distance covered by the sequence is bounded and the overall interestingness is maximized. To this extent, we propose a mathematical model based on the formulations for the orienteering problem found in [6, 8]. This model uses $n^2 + 4n$ variables and $n^2 + 3n + 4$ constraints.

data

Given a set of n queries noted q_1 through q_n

$c_{i,j}$the distance between queries, q_i and q_j. (Positive integer)
$t_i$denotes the execution time of q_i (Positive integer)
v_i is the interestingness score associated with q_i (real number in [0,1])

variables

$x_{i,j}, (i,j) \in 1..n, x_{i,j} = 1$ if q_i comes directly before q_j in the solution, 0 otherwise
$x_{0,i}, i \in 1..n, x_{i,j} = 1$ if q_i is the first query of the solution, 0 otherwise
$x_{i,n+1}, i \in 1..n, x_{i,j} = 1$ if q_i is the last query of the solution, 0 otherwise
$s_i, i \in 1..n$: boolean variables denoting the presence of q_i in the solution.
$u_i, i \in 1..n$: integer variables used in subtour elimination constraints.

objective

$$max \sum_{i=1}^{n} v_i s_i \tag{1}$$

under constraints

$$\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} c_{i,j} x_{i,j} \leq \epsilon_d \tag{2}$$

$$\sum_{j=1, j \neq i}^{n+1} (x_{i,j}) - s_i = 0, \forall i \in 1..n \tag{5}$$

$$\sum_{i=1}^{n} t_i s_i \leq \epsilon_t \tag{3}$$

$$\sum_{j=1}^{n} x_{0j} = \sum_{i=1}^{n} x_{i,n+1} = 1 \tag{6}$$

$$\sum_{i=0, j \neq i}^{n} (x_{i,j}) - s_j = 0, \forall j \in 1..n \tag{4}$$

$$2 \leq u_i \leq n, i \in 1..n \ , \ u_i - u_j + 1 \leq (n-1)(1 - x_{ij}), (i,j) \in 1..n \tag{7}$$

This model will be used to study the influence of epsilon constraints on synthetic and real instances, and results will be presented at the conference. Further, this model will serve as a basis for matheuristics.

Références

[1] Julien Aligon, M. Golfarelli, P. Marcel, S. Rizzi, and Elisa Turricchia. Similarity measures for olap sessions. *Knowledge and Information Systems*, 39 :463–489, 2013.

[2] Ori Bar El, Tova Milo, and Amit Somech. Automatically generating data exploration sessions using deep reinforcement learning. SIGMOD '20.

[3] Alexandre Chanson, B. Crulis, N. Labroche, P. Marcel, V. Peralta, S. Rizzi, and P. Vassiliadis. The traveling analyst problem :definition and preliminary study. DOLAP '20.

[4] V. Dibia and Ç. Demiralp. Data2vis : Automatic generation of data visualizations using sequence-to-sequence recurrent neural networks. *IEEE C.G.A.*, 39(5), 2019.

[5] Liqiang Geng and Howard J. Hamilton. Interestingness measures for data mining : A survey. *ACM Comput. Surv.*, 38(3), September 2006.

[6] A. Gunawan, H. C. Lau, and P. Vansteenwegen. Orienteering problem : A survey of recent variants, solution approaches and applications. *E.J.O.R.*, 255(2), 2016.

[7] Stratos Idreos, Olga Papaemmanouil, and Surajit Chaudhuri. Overview of data exploration techniques. SIGMOD '15.

[8] Imdat Kara, Papatya Sevgin Bicakci, and Tusan Derya. New formulations for the orienteering problem. *Procedia Economics and Finance*, 39 :849–854, 2016.

[9] T. Tsiligirides. Heuristic methods applied to orienteering. *The Journal of the Operational Research Society*, 35(9) :797–809, 1984.

[10] Pieter Vansteenwegen and Aldy Gunawan. *Orienteering Problems : Models and Algorithms for Vehicle Routing Problems with Profits.* EURO Advanced Tutorials on Operational Research. Springer International Publishing.

[11] A. Wasay, M. Athanassoulis, and S. Idreos. Queriosity : Automated data exploration. In *2015 IEEE International Congress on Big Data*, pages 716–719, 2015.