Классификация демографических последовательностей на основе узорных структур

Дмитрий Игоревич Игнатов[1]⋆ dignatov@hse.ru
Данил Кутдусович Гизадтуллин[1] gizdatullindanil@gmail.com
Екатерина Сергеевна Митрофанова[1] emitrofanova@hse.ru
Анна Александровна Муратова[1] anyamuratova@yandex.ru
Жауме Башерье[2] jbaixer@lsi.upc.edu

[1]Москва, Национальный исследовательский университет Высшая школа экономики

[2]Барселона, Политехнический университет Каталонии

В работе представлены результаты первых экспериментов применения узорных структур на последовательностях к анализу демографических данных в России. Использованы данные об 11-ти поколениях с 1930 по 1984 для панели из трех волн, имевших место в 2004, 2007 и 2011. Основная задача состояла в поиске таких закономерностей, которые являются (замкнутыми) частыми префиксами без "разрывов". Эти ограничения – естественное требование демеографов, необходимое для изучения первых событий на этапе взросления. Для решения этой задачи использованы узорные структуры неразрывных последовательностей и модифицированные FP-деревья. Наилучшие результаты в терминах TPR-FPR были получены при больших значений параметра роста (с некоторым числом отказов от классификации).

Статья подготовлена в ходе проведения исследования № 16-05-0011 «Разработка и апробация методик анализа демографических последовательностей» в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2016 г. и с использованием средств субсидии на государственную поддержку ведущих университетов Российской Федерации в целях повышения их конкурентоспособности среди ведущих мировых научно-образовательных центров, выделенной НИУ ВШЭ.

[1] *Ignatov, D.I., Mitrofanova, E., Muratova, A. and Gizdatullin, D.* Pattern Mining and Machine Learning for Demographic Sequences // Knowledge Engineering and Semantic Web: 6th International Conference (KESW 2015), Proceedings, Cham: Springer, 2015. — p. 225–239. http://dx.doi.org/10.1007/978-3-319-24543-0_17.

Pattern-based classification of demographic sequences

Dmitry I. Ignatov [1]⋆ dignatov@hse.ru
Danil Gizdatullin[1] gizdatullindanil@gmail.com
Ekaterina Mitrofanova [1] emitrofanova@hse.ru
Anna Muratova[1] anyamuratova@yandex.ru
Jaume Baixeries[2] jbaixer@lsi.upc.edu
[1]Moscow, National Research University Higher School of Economics
[2]Barcelona, Universitat Politècnica de Catalunya

This paper presents the first results of studies in application of sequence-based pattern structures and emerging patterns to analysis of demographic sequences in Russia. This study is performed on data of 11 generations from 1930 till 1984 for the panel of three waves of the Russian part of Generation and Gender Survey, which took place in 2004, 2007, and 2011. The main goal is to develop methods of extracting emerging patterns (EP) with the following restrictions: the obtained patterns need to be (closed) frequent gapless prefixes of the input sequences. These constraints were required by demographers since it is necessary for proper interpretation and understanding of early life course events that lead to adulthood. To solve this problem, we used pattern structures of gapless prefixes and modified FP-trees. After extraction of EP we use CAEP classifier to predict gender of respondents using their demographic sequences of the first life course events. The best results in terms of TPR-FPR have been obtained for large values of minimum growth-rate parameter (with some objects left without classification).

The paper was prepared within the framework of the Academic Fund Program at HSE in 2016 (grant № 16-05-0011 "Development and testing of demographic sequence analysis and mining techniques") and supported within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program.

[1] *Ignatov, D.I., Mitrofanova, E., Muratova, A. and Gizdatullin, D.* Pattern Mining and Machine Learning for Demographic Sequences // Knowledge Engineering and Semantic Web: 6th International Conference (KESW 2015), Proceedings, Cham: Springer, 2015. — p. 225–239. http://dx.doi.org/10.1007/978-3-319-24543-0_17.