Поиск закономерностей в базе данных демографических последовательностей на основе узорных структур[1]

Гиздатуллин Данил Кутдусович
НИУ ВШЭ

Анализ демографических последовательностей - популярное и многообещающее направление в изучении демографии. Жизнь людей можно рассматривать как последовательность событий, происходящих в их жизни. Исследователям в области демографии интересен переход от анализа отдельных событий и их взаимосвязей к анализу полных последовательностей событий.

Демографическое поведение может сильно разниться среди людей из разных поколений, с разным полом, уровнем образования, религиозными взглядами и т.д. Однако скрытые сходства могут быть найдены и обобщены с помощью специально изобретенных техник. И хотя уже изобретено множество методов для решения этой задачи она все еще далека от того, чтобы решаться стандартными методами анализа последовательностей, которые изучаются в майнинге данных. Использование методов майнинга данных открывает для демографов новые возможности для анализа результатов исследований. Но, как будет показано в работе, некоторые стандартные методы, которые используются в традиционном анализе последовательностей, не могут быть использованы на прямую и требуют специальной адаптации под нужды исследователей из других областей.

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 демографов, 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 значениях параметра роста (с некоторым числом отказов от классификации).

[1] Статья подготовлена в результате 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 в рамках государственной поддержки ведущих университетов Российской Федерации "5-100".