Поиск закономерностей в индивидуальных демографических траекториях

Данил Гиздатуллин[1], Дмитрий Игнатов[1], Екатерина Митрофанова[1], Анна Муратова[1] и Жауме Башерье (Jaume Baixeries)[2]

[1] Национальный исследовательский университет Высшая школа экономики, Москва, Россия
{dgizdatullin,dignatov,emitrofanova,amuratova}@hse.ru
[2] Политехнический университет Каталонии, Барселона, Каталония
jbaixer@lsi.upc.edu

Аннотация В данной работе представлены результаты применения узорных структур (pattern structures) и "контрастных" закономерностей (emerging patterns) в анализе демографических последовательностей для данных по России. Панельные данные Российской части исследования GGS (Generation and Gender Survey) на основе трех волн опроса в 2004, 2007, и 2011 описывают 11 поколений респондентов, начиная с 1930 по 1984. Основная задача заключалась в реализации методов для извлечения "контрастных" закономерностей (EP) при наличии дополнительных ограничений: полученные закономерности должны быть (замкнутыми) частыми неразрывными префиксами входных последовательностей. Такие ограничения необходимы демографам для приемлемой интерпретации результатов и выявления событий на ранних этапах жизни, которые ведут к взрослению. Для удовлетворения ограничениям были применены FP-деревья[3] на основе узорных структур и неразрывных префиксов. После извлечения EP, мы используем схему классификатора CAEP[4] для предсказания пола респондентов на основе их демографических последовательностей событий их ранней жизни. Лушчие результаты в терминах TPR-FPR кривых были получены для больших значений параметра минимального темпа роста (однако, некоторые респонденты остались неклассифицированы).

Ключевые слова: демографические последовательности, узорные структуры (pattern structures), анализ последовательностей, "контрастные" закономерности (emerging patterns), "контрастные" последовательности, машинное обучение

1 Введение

Анализ демографических последовательностей – популярное и многообещающее направление в демографических исследованиях [2,3]. Весьма упрощен-

[3] чтобы отличить нашу версию от оригинальной [1] мы расшифровываем аббревиатура FP как frequent prefix, а не frequent pattern
[4] Classification by Aggregating Emerging Patterns

но жизнь людей можно рассматривать как последовательность демографических событий. Исследователям в области демографии интересен переход от анализа отдельных событий и их взаимосвязей к анализу полных последовательностей событий. Однако, этот переход замедляют технические сложности работы с такимим последовательностями. Сейчас демографы и социологи не имеют доступных и удобных инструментов для такого анализа. Некоторые демографы, обладающие навыками программирования, успешно используют анализ последовательностей [4,5,6] и статистические методы [7,8,9,10,11], но большинству исследователей в этой области остается толкьо возможность сотрудничать с коллегами из других областей для извлечения знаний из демографических данных. Обычно демографы полагаются на простую статистику, а сложные методы анализа последовательностей только начинают появляться в этой области [12]. Поскольку традиционные статистические методы не могут удовлетворить новым потребностям демографии, демографы начинают проявлять большой интерес к методам машинного обучения и поиска закономернойстей [13,14].

Демографическое поведение может сильно разниться среди людей разных поколений, пола, уровня образования, религиозных взглядов и т.д. Однако скрытое сходство в их поведении может быть выявлено и обобщено с помощью специально предложенных подходов. И хотя уже предложено немало методов для решения этой задачи, ее формулировка 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 напрямую и требуют специальной адаптации под нужды исследователей из других областей.

В своей предыдущей работе [14], мы использовали SPMF[5] [15] для поиска частых последовательностей и нашу разработку для поиска “контрастных” паттернов среди частых последовательностей. Однако, реализованные методы извлечения последовательностей основаны на определении подпоследовательности: это дает результаты, которые трудно интерпретировать. По просьбе коллег-демографов, мы будем искать неразрывные префиксные последовательности, т.е. подпоследовательности траекторий жизненного пути людей без пропусков.

Итак, одна из целей этого исследования – это разработка методов извлечения контрастных закономерностей со следующими ограничениями: полученные паттерны должны быть замкнутыми частыми непрерывными префиксными подпоследовательностями. Для осуществления этой цели мы используем узорные структуры [16], которые предназначены для анализа данных сложной структуры на основе замкнутых описаний и уже неплохо за-

[5] Sequential Pattern Mining Framework: `http://www.philippe-fournier-viger.com/spmf/`

рекомендовали себя в анализе последовательностей (например, траекторий лечения) [17], а также “контрастные” закономерности [18].

Основной задачей является поиск интересных и интерпретируемых закономерностей, которые могут характеризовать отдельные классы, то есть таких закономерностей, которые были бы характерны для одного класса, но не для всех остальных. Классификация сама по себе скорее является средством, а не целью в данном исследовании. Таким образом, хорошие результаты классификации только заверяют нас, что классификатор на основе префиксов применим к проблеме. С этой точки зрения подходы типа “черный ящик”, такие как SVM (хотя можно было бы взглянуть на опорные векторы), и искусственные нейронные сети не соответствуют задаче; они могут показывать лучшие результаты в точности предсказаний, но не дают интерпретируемых закономерностей.

Помимо введения, материал статьи представлен в трех основных разделах. В разделе 2 описаны данные и постановка задачи. Результаты экспериментов обсуждаются в разделе 3. Раздел 4 завершает статью. В силу предъявляемых ограничений мы не даем математических определений и не разъясняем работу алгоритмов, но адресуем читателя к четырем базовым статьям [18,16,17,14,19].

2 Постановка проблемы и описание демографических данных

Набор данных был предоставлен научно-учебной группой “Модели и методы анализа демографических последовательностей” [6]. Мы использовали панель из трех волн обследования “Родители и дети, мужчины и женщины в семье и обществе”, которые проходили в 2004, 2007 и 2011 годах[7]. База данных содержит ответы 4857 респондентов (1545 мужчин и 3312 женщин). Гендерный дисбаланс набора данных вызван панельной природой данных.

Для каждого респондента была представлена следующая информация: дата рождения, пол, поколение, уровень образования, тип местности проживания (город, поселок городского типа, сельская местность), верующий ли человек, частота посещения церкви. Также указаны даты значимых событий в их жизни таких как: первый опыт работы, дата завершения обучения, отделение от родителей, первое партнерство, первое замужество/женитьба, дата рождения первого ребенка. В данных представлены 11 разных поколений в период между 1930 – 1984.

Существует ряд типовых вопросов, на которые демографы хотели бы получить ответы, например:

– В чем отличие между мужчинами и женщинами с точки зрения демографического поведения?

[6] `http://www.hse.ru/en/demo/family/`

[7] часть опроса GGS “Parents and Children, Men and Women in Family and in Society”, всероссийская панельная выборка: `http://www.ggp-i.org/`

– В чем отличие между различными поколениями с точки зрения демографического поведения?

Поэтому для ответов на большое количество подобных вопросов нужны подходящие средства выявления закономерностей в данных.

3 Эксперименты и результаты

Для проведения экспериментов с классификацией мы использовали Python и библиотеку Contiguous Sequences Analysis, написанную первым автором[8].

3.1 Классификация по полу

После обсуждения с демографами, мы установили минимальную относительную поддержку равной 0,09. Седующие неразрывные префиксные последовательности встречаются как минимум у 9% всех респондентов:

Таблица 1. Женские паттерны

Pattern	$rsupp$
$\langle\{work\}\rangle$	0,287
$\langle\{work\},\{education\}\rangle$	0,120
$\langle\{separation\}\rangle$	0,283
$\langle\{education\}\rangle$	0,239
$\langle\{education\},\{work\}\rangle$	0,168
$\langle\{separation\},\{education\}\rangle$	0,110
$\langle\{sep...\},\{edu...\},\{work\}\rangle$	0,097

Таблица 2. Мужские паттерны

Pattern	$rsupp$
$\langle\{work\}\rangle$	0,329
$\langle\{work\},\{education\}\rangle$	0,155
$\langle\{separation\}\rangle$	0,266
$\langle\{education\}\rangle$	0,276
$\langle\{education\},\{work\}\rangle$	0,103
$\langle\{separation\},\{education\}\rangle$	0,199
$\langle\{sep...\},\{edu...\},\{work\}\rangle$	0,099

Результаты дают понять, что начало жизненного пути не сильно зависит от пола.

Мы разделили все наши данные на две группы: обучающее множество и тестовое множество. 66,5%–33,5%.

Было выыбрано одинаковое минимальное значение поддержки для обоих классов, 0,004; это значит, что закономерность должна встретиться как минимуму у пяти мужчик и девяти женщин. Затем мы провели классификацию с различными минимальными значениями темпа роста $\{1.5, 2, 5, 7\}$ для мужчин и $\{1.5, 2, 5, 7, \infty\}$ для женщин.

Графики на Рис. 1 и 2 показывают результаты классификации с выделенной Парето-границей (skyline) в координатах TPR-FPR (true positive rate, false positive rate), TPR-NCPR (true positive rate, non-classified positive rate), NCPR-FPR (non-classified positive rate, false positive rate).

Так как нам важно было выявить интересные сильно отличительные паттерны, мы не пытались решить проблему того, что какая-то доля объектов из тестового множества остается без предположения о принадлежности

[8] `https://github.com/DanilGizdatullin/ContiguousSequencesAnalysis`

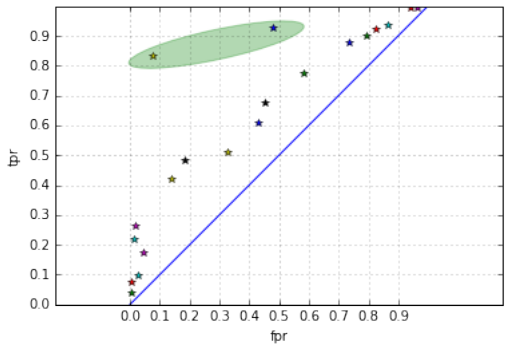

Рис. 1. График TPR-FPR с двумя Парето-оптимальными результатами из Парето-границы (в овале).

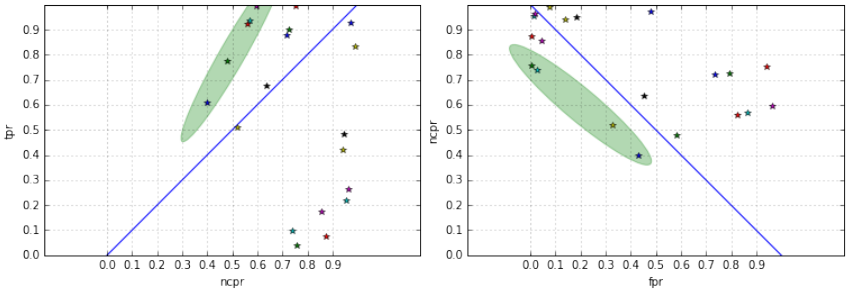

Рис. 2. Графики TPR-NCPR и NCPR-FPR с выделенными Парето-границами (в овалах).

одному из классов. Например в эксперименте с самым лучшим значением TPR-FPR меры классифицировано всего несколько больше 1% людей из тестовой выборки. При этом средние значения как точности (precision), так и полноты (recall) достигают $0,79\%$. То есть из результатов можно сделать вывод, что паттерны, которые являются сильно отличительными для какого-то класса относительно другого, имеют небольшое покрытие, а значит, варианты поведения мужчин и женщин не имеют больших отличий в общем, но присутствуют локальные группы обоих классов, которые ведут себя с очень большими отличиями.

Наилучшее качество классификации получилось достичь при минимальных значениях темпа роста $(7, \infty)$. Эти значения соответсвуют следующим "контрастным" закономерностям:

Мы получили 7 последовательностей, наиболее четко разделяющих мужчин и женщин: 3 последовательности, определяющих мужчин, и 4, определяющих женщин. Темп роста показывает, что все "женские" последовательности типичны только для женщин (growth rate $= \infty$), темп роста "мужских" последовательностей находится в диапазоне от 10,6 до 12,7, что означает,

Таблица 3. Закомерности для подвыборки женщин тестового множества

Pattern	Growth rate	$rsupp$
$\langle\{work, separation\}, \{marriage\}, \{children\}, \{education\}\rangle$	∞	0,006
$\langle\{separation, partner\}, \{marriage\}\rangle$	∞	0,005
$\langle\{separation, partner\}, \{marriage\}\rangle$	∞	0,005
$\langle\{work, separation\}, \{marriage\}, \{children\}\rangle$	∞	0,008
$\langle\{work, separation\}, \{marriage\}\rangle$	∞	0,009

Таблица 4. Закомерности для подвыборки мужчин тестового множества

Pattern	Growth rate	$rsupp$
$\langle\{education\}, \{marriage\}, \{work\}, \{children\}, \{separation\}\rangle$	10.6	0,006
$\langle\{education\}, \{marriage\}, \{work\}, \{children\}\rangle$	12,7	0,007
$\langle\{edu...\}, \{work\}, \{partner\}, \{marriage\}, \{sep...\}, \{children\}\rangle$	10,6	0,006

что данные последовательности показательны для мужчин в 10-12 раз чаще, чем для женщин.

В “женских” последовательностях первое событие “отделение” происходит одновременно с другими: “работа” (3 случая из 4) и “партнер” (1 случай из 4). Второе событие для женщин “замужество”, третье, если есть, это “дети” и четвертое “образование”.

Полученные выше результаты показывают, что женщины склонны начинать взрослую жизнь с отделения от семьи. Только в одном случае отделение связано с рождением ребенка, в остальных случаях мы видим образ независимой женщины, у которой есть работа и которая отделилась от родителей. Второй шаг во всех случая – замужество. Мы видим, что финансово независимая женщина создает свою семью и рожает ребенка. Самая длинная последовательность содержит событие “получение высшего образования”. Таким образом, только после 4 важных социально-экономических и социально-демографических событий наша “типичная” женщина завершает образование.

Рассмотрим “мужские” последовательности. В них первое событие для мужчин – образование. В отличие от женщин, мужчины раньше получают образование. Это показывает не только жизненные приоритеты мужчин и женщин, но и разницу в наивысшей ступени получения образования. Второе событие для мужчин – это женитьба (2 случая из 3) и работа (1 случай из 3). Как и женщины, мужчины склонны создавать семью достаточно рано, но в отличие от женщин, которые к моменту создания семьи уже имеют работу и независимы, мужчины к этому моменту обладают только образованием. Мужчины, для которых женитьба является вторым шагом, затем получают работу и становятся отцами. На последнем этапе они покидают родитель-

ский дом, что является окончательнм шагом к взрослению. Мужчины, для которых работа – второй шаг, имеют другой набор последующих: у них появляется первый партнер, затем они женятся, покидают родительский дом и на последнем шаге становятся родителями.

3.2 Классификация по поколению

В этом эксперименте мы пытаемся найти "контрастные" паттерны для различных поколений одного пола. Класс 0 будут иметь люди рожденные между 1924 и 1959. Класс 1 будут иметь люди рожденные между 1960 и 1984.

Сначала найдем закономерности для женщин разных поколений. У нас есть 940 женщин класса 0 и 1364 женщины класса 1. Нам нужно подобрать два параметра: мнимальную поддержку и минимальный темп роста.

Подберем минимальную поддержку (Таблица 5).

Таблица 5. Подбор параметра минимальная поддержка (minimal support) для женщин

minsup	accuracy	TPR	FPR	NCR non-classification rate
0,001	0,682	0,707	0,331	0,255
0,004	0,683	0,703	0,316	0,333
0,01	0,668	0,710	0,332	0,399
0,025	0,660	0,648	0,298	0,540
0,04	0,660	0,616	0,278	0,606
0,05	0,652	0,646	0,312	0,641
0,1	0,651	1,0	1,0	0,884

Как видно из результатов, минимальная поддержка может влиять в значительной мере только на долю неклассифицированых объектов, и незначительно влиять на долю правильно классифицированных, TPR и FPR.

Мы выбрали 0,004 в качестве минимальной поддержки и начали искать оптимальный минимальный темп роста.

Выбор minGrowthRate=2 адекватен, так как в этом случае покрывается 66% тестовой выборки и обеспечиваются хорошие результаты по мерам accuracy, TPR и FPR.

Так как мы получили достаточно много контрастных паттернов из этих данных, мы рассмотрим только паттерны с наибольшими значениями темпа роста и поддержки.

Как мы увидели из Таблиц 7–8, главное отличие в поведении женщин разных поколений – это тенденция к получению образования, работы и лишь затем отделение от родителей.

Теперь найдем "контрастные" паттерны для мужчин из разных поколений. И снова мы ищем оптимальное значение поддержки:

Таблица 6. Подбор параметра минимальный темп роста (minimal growth rate) для женщин

minGrowthRate	accuracy	TPR	FPR	NCR
1,5	0,683	0,655	0,297	0,102
2	0,692	0,703	0,316	0,333
3	0,766	0,747	0,217	0,684
5	0,751	0,821	0,347	0,848
7	0,777	0,848	0,333	0,891

Таблица 7. Закономерности для женщин старшего поколения

Pattern	$Growthrate$	$rsupp$
$\langle\{work\},\{separation\}\rangle$	1,85	0,38
$\langle\{work\},\{marriage, separation\}\rangle$	3,92	0,08

Снова минимальная поддержка сильно влияет только на число неклассифицированных объектов.

Зафиксируем 0,01 в качестве минимальной поддержки и найдем оптимальный темп роста.

Таблица 10. Настройка параметра оптимальный темп роста для мужчин

minGrowthRate	accuracy	TPR	FPR	NCR
1,2	0,638	0,510	0,242	0,050
1,5	0,670	0,591	0,260	0,171
2	0,723	0,671	0,232	0,442
3	0,754	0,627	0,144	0,664
5	0,744	0,625	0,152	0,845
7	0,836	0,808	0,138	0,901

Закономерности с наибольшими темпом роста и поддержкой выписаны в Таблицах 11,12.

Таблица 11. Закономерности для женщин старшего поколения

Pattern	Growth rate	$rsupp$
$\langle\{work\},\{marriage, separation\},\{education\}\rangle$	13,52	0,025
$\langle\{work\},\{marriage\},\{separation\}\rangle$	22,87	0,042
$\langle\{work\},\{marriage\},\{separation\},\{education\}\rangle$	∞	0,0208

Таблица 8. Закономерности для женщин молодого поколения

Pattern	$Growthrate$	$rsupp$
$\langle\{education\}\rangle$	1,84	0,26
$\langle\{education\},\{work\}\rangle$	3,92	0,08

Таблица 9. Подбор параметра минимальная поддержка (minimal support) для мужчин

minsup	accuracy	TPR	FPR	NCR
0,001	0,701	0,667	0,266	0,271
0,004	0,704	0,667	0,262	0,338
0,01	0,723	0,671	0,232	0,442
0,025	0,719	0,651	0,218	0,590
0,04	0,706	0,536	0,165	0,712
0,05	0,718	0,627	0,208	0,764
0,08	0,710	0,0	0,0	0,944

Таблица 12. Закономерности для женщин молодого поколения

Pattern	Growth rate	$rsupp$
$\langle\{education\},\{work\},\{separation\},\{marriage\},\{children\}\rangle$	10,58	0,020
$\langle\{education\},\{work\},\{separation, partner\},\{marriage\}\rangle$	8,65	0,016
$\langle\{education\},\{marriage, separation\}\rangle$	7,69	0,015

Как и в предыдущем эксперименте с подвыборкой женщин, основным отличием является тенденция к получению образования; таким образом, мужчины молодого поколения демонстрируют эту тенденцию.

4 Заключение

Основным результатом работы является применение различных методов майнинга данных к анализу демографических последовательностей. Можно сделать следующие выводы по результатам работы:

1. В данной работе было изучено применение методов анализа последовательностей к задачам демографического направления. В частности задаче поиска закономерностей, характеризующих отдельные классы данных.
2. Была разработана и реализована новая модификация метода анализа последовательностей специального типа (неразрывных и префиксных).
3. Получены и проинтерпретированы паттерны (закономерности) поведения для разных классов респондентов.

4. Разработан и протестирован классификатор на основе “контрастных” частых последовательностей и узорных структур.

Проделанная работа очень важна для дальнейшего развития применения методов майнинга данных в демографических задачах, в частности анализе демографических последовательностей.

Благодарности Выражем благодарность нашим коллегам, проф. Сергею Олеговичу Кузнецову, Алексею Бузмакову и Меди Кейту (Mehdi Kaytoue) за их советы и помощь по тематике узорных структур и анализу последовательностей, а также благодарим проф. Гуожу Донга (Guozhu Dong) за проявленный интерес к нашей работе по тематике “контрастных” закономерносте и коллег из Института демографии НИУ ВШЭ.

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”. Второй автор был частично поддержан Российским фондом фундаментальных исследований.

Список литературы

1. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: A frequent-pattern tree approach. Data Mining and Knowledge Discovery **8**(1) (2004) 53–87
2. Aisenbrey, S., Fasang, A.E.: New life for old ideas: The ‘second wave’ of sequence analysis bringing the ‘course’ back into the life course. Sociological Methods & Research **38**(3) (2010) 420–462
3. Billari, F.C.: Sequence analysis in demographic research. Canadian Studies in Population **28**(2) (2001) 439–458.
4. Aassve, A., Billari, F.C., Piccarreta, R.: Strings of adulthood: A sequence analysis of young british women’s work-family trajectories. European Journal of Population **23**(3/4) (2007) 369–388
5. Braboy Jackson, P., Berkowitz, A.: The structure of the life course: Gender and racioethnic variation in the occurrence and sequencing of role transitions. Advances in Life Course Research (9) (2005) 55–90
6. Worts, D., Sacker, A., McMunn, A., McDonough, P.: Individualization, opportunity and jeopardy in american women’s work and family lives: A multistate sequence analysis. Advances in Life Course Research **18**(4) (2013) 296–318
7. Abbott, A., Tsay, A.: Sequence analysis and optimal matching methods in sociology: Review and prospect. Sociological Methods & Research (2000)
8. Billari, F., Piccarreta, R.: Analyzing demographic life courses through sequence analysis. Mathematical Population Studies **12**(2) (2005) 81–106
9. Billari, F.C., Fürnkranz, J., Prskawetz, A.: Timing, Sequencing, and Quantum of Life Course Events: A Machine Learning Approach. European Journal of Population **22**(1) (2006) 37–65

10. Gauthier, J.A., Widmer, E.D., Bucher, P., Notredame, C.: How Much Does It Cost? Optimization of Costs in Sequence Analysis of Social Science Data. Sociological Methods & Research **38**(1) (2009) 197–231
11. Ritschard, G., Oris, M.: Life course data in demography and social sciences: Statistical and data-mining approaches. Adv. in Life Course Research **10** (2005) 283–314
12. Gabadinho, A., Ritschard, G., Müller, N.S., Studer, M.: Analyzing and Visualizing State Sequences in R with TraMineR. J. of Stat. Software **40**(4) (4 2011) 1–37
13. Blockeel, H., Fürnkranz, J., Prskawetz, A., Billari, F.C.: Detecting temporal change in event sequences: An application to demographic data. In: Principles of Data Mining and Knowledge Discovery, 5th Eur. Conf., PKDD 2001. (2001) 29–41
14. Ignatov, D.I., Mitrofanova, E., Muratova, A., Gizdatullin, D.: Pattern mining and machine learning for demographic sequences. In: Knowledge Engineering and Semantic Web, KESW 2015, Proceedings. (2015) 225–239
15. Fournier-Viger, P., Lin, J.C., Gomariz, A., Gueniche, T., Soltani, A., Deng, Z., Lam, H.T.: The SPMF open-source data mining library version 2. In: Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part III. (2016) 36–40
16. Ganter, B., Kuznetsov, S.O.: Pattern structures and their projections. In: Harry S. Delugach and Gerd Stumme, editors, Conceptual Structures: Broadening the Base, volume 2120 of Lecture Notes in Computer Science. (2001)
17. Egho, E., Raïssi, C., Calders, T., Jay, N., Napoli, A.: On measuring similarity for sequences of itemsets. Data Min. Knowl. Discov. **29**(3) (2015) 732–764
18. Dong, G., Li, J.: Efficient mining of emerging patterns: Discovering trends and differences. In: Proc. of the Fifth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. KDD '99, ACM (1999) 43–52
19. Kaytoue, M., Codocedo, V., Buzmakov, A., Baixeries, J., Kuznetsov, S.O., Napoli, A.: Pattern structures and concept lattices for data mining and knowledge processing. In: Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2015, Porto, Portugal, September 7-11, 2015, Proceedings, Part III. (2015) 227–231