Раздел I. Искусственный интеллект и нечеткие системы

УДК 004.853 DOI 10.18522/2311-3103-2020-4-6-21

В.В. Курейчик, В.В. Бова, Ю.А. Кравченко

МЕТОД ПОИСКА ПОСЛЕДОВАТЕЛЬНЫХ ПАТТЕРНОВ ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ В ИНТЕРНЕТ-ПРОСТРАНСТВЕ*

Одной из важных задач интеллектуального анализа данных является выделение закономерностей и обнаружение связанных событий в последовательных данных на основе анализа последовательных паттернов. В статье исследуются возможность применения последовательных паттернов для анализа событий поисково-познавательной деятельности пользователей при взаимодействии с Интернет-ресурсами открытой информационно-образовательной среды. Поиск последовательных паттернов является сложной вычислительной задачей, цель которой состоит в извлечении всех частых последовательностей, отражающих потенциальные связи внутри элементов из транзакционной базы данных последовательностей событий поисковой активности при заданной минимальной поддержке. Для ее решения в статье предлагается метод поиска закономерностей в последовательностях событий для обнаружения скрытых закономерностей, указывающих с возможные уровни уязвимости при выполнении задач информационного поиска в Интернет-пространстве. Описана математическая модель поведения пользователей в поисковой сессии, основанная на теории последовательных паттернов. Для повышения вычислительной эффективности метода разработан модифицированный алгоритм генерации последовательных паттернов, на первом этапе которого выполняется AprioriAll, формирующий частые последовательности-кандидаты всевозможных длин, а на втором – генетический алгоритм оптимизации входных параметров признакового пространства сгенерированного множества для поиска максимальных паттернов. Проведены серии вычислительных экспериментов на тестовых данных корпуса MSNBC, библиотеки интеллектуального анализа данных с открытым исходным кодом SPMF. Сравнительной анализ проводился с алгоритмами VMSP и GSP. Результаты исследований подтвердили эффективность поиска максимальных последовательных паттернов предложенным алгоритмом с точки зрения времени выполнения и количества извлеченных паттернов. Результаты проведенных экспериментальных исследований метода показали, что для увеличения стабильности и точности работы размер выборки, полученной в результате работы ГА, позволит сократить необходимое число сканирований базы данных паттернов, обеспечивая приемлемые вычислительные затраты, сопоставимые с алгоритмом VMSP и превосходящий по времени поиска последовательных паттернов алгоритм GSP в среднем более чем на 150 %.

Ппоиск последовательных паттернов; секвенциальный анализ; генетический алгоритм; транзакционная база данных; информационный поиск.

* Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-29-22019.

V.V. Kureychik, V.V. Bova, Y.A. Kravchenko

METHOD FOR SEARCHING SEQUENTIAL PATTERNS OF USER'S BEHAVIOR ON THE INTERNET

One of the important tasks of data mining is to isolate patterns and detect related events in sequential data based on the analysis of sequential patterns. The article examines the possibility of using sequential patterns to analyze the events of search and cognitive activity of users when interacting with Internet resources of an open information and educational environment. Searching for sequential patterns is a complex computational task whose goal is to retrieve all frequent sequences representing potential relationships within elements from a transactional database of sequences of search activity events with a given minimum support. To solve it, the article proposes a method for searching for patterns in sequences of events to detect hidden patterns that indicate possible levels of vulnerability when performing information search tasks in the Internet space. A mathematical model of user behavior in a search session based on the theory of sequential patterns is described. To improve the computational efficiency of the method, a modified algorithm for generating sequential patterns has been developed, at the first stage of which AprioriAll is performed, which forms frequent candidate sequences of all possible lengths, and at the second stage, a genetic algorithm for optimizing the input parameters of the feature space of the generated set to search for maximum patterns. A series of computational experiments were carried out on test data from the MSNBC corpus, the SPMF open source data mining library. The comparative analysis was carried out with the VMSP and GSP algorithms. The research results confirmed the efficiency of the search for maximum sequential patterns by the proposed algorithm in terms of the execution time and the number of extracted patterns. The results of the experimental studies of the method showed that to increase the stability and accuracy of the work, the sample size obtained as a result of the GA operation will reduce the required number of scans of the pattern database, providing acceptable computational costs comparable to the VMSP algorithm and the GSP algorithm that exceeds the search time for sequential patterns. an average of more than 150 %.

Search for sequential patterns; sequential analysis; genetic algorithm; transactional database; information search.

Введение. В настоящее время для поиска закономерностей при решении задач, связанных с необходимостью извлечения новых знаний при обработке данных об активности пользователей при взаимодействии с Интернет-ресурсами открытой информационно-образовательной среды, получили методы машинного обучения: ассоциативные правила и последовательные паттерны [1–3]. Задача поиска закономерностей в таких данных связана с необходимостью обнаружения и анализа возможных ассоциативных последовательностей событий поисковой сессии при выполнении поставленных перед пользователем задач [3–6]. Исследуемые объекты данных описываются числовыми признаками, поэтому для поиска ассоциативных зависимостей и построения на их основе последовательных паттернов необходимо осуществлять предварительную обработку таких атрибутов с учетом значимости признаков [4, 7].

Для решения задач повышения эффективности взаимодействия пользователей с информационными поисковыми системами ряд исследователей предлагают анализировать паттерны (шаблоны) активности пользователей, т.е. повторяющиеся последовательности действий [8–10]. Однако существующие методы анализа ориентированы на поиск лишь предопределенных исследователями шаблонов, основанных на заранее собранных рекомендациях и гипотезах относительно того, как пользователь взаимодействует с Интернет-пространством [6, 11–13]. Данные методы не используют существующие методы интеллектуального анализа ассоциативных правил и последовательных паттернов для поиска новых (ранее неизвестных) зависимостей в данных о сетевой активности пользователей.

При решении задачи извлечения новых знаний об исследуемых зависимостях в наборах транзакционных данных об активности пользователей возникают проблемы нахождения всех наборов элементов в том числе и неявных, позволяющих выявлять косвенные ассоциации в последовательности действий. Главная проблема при их поиске – большое число частых последовательностей, возникающих при исследовании больших бинарных контекстов длинных транзакций, масштабируемость во времени и пространстве для которых становится критичной. Это существенно усложняет экспертный анализ выявленных последовательных паттернов. Для решения этой проблемы используются различные меры значимости, такие как поддержка (support) и достоверность (confidence) [5, 7]. С их помощью найденные ассоциативные связи фильтруются, и для анализа предъявляются только те, для которых значения мер значимости превышают заданные пороговые значения. Подобная фильтрация способствует снижению количества сгенерированных паттернов, но не решает проблему размерности полностью. Это приводит к значительному возрастанию размерности решаемой задачи и повышает требования к вычислительным ресурсам.

Другой проблемой при поиске ассоциативных зависимостей в данных о сетевой активности пользователей, является распознавание событий, связанных с возможными уровнями уязвимости при выполнении задач информационного поиска. Поэтому требуется понимание семантики паттернов конкретных действий поисковой сессии [11, 14]. Для решения рассмотренных проблем авторами предлагается метод поиска последовательных паттернов и математическая модель поведения пользователей, для обнаружения скрытых закономерностей, указывающих на возникновение рисков поисково-познавательной деятельности при взаимодействии пользователей с Интернет-пространством.

Постановка задачи. Введем несколько основных понятий, необходимых для формализации задачи поиска последовательных паттернов (SPM) поведения пользователей при организации поисковой деятельности, определяемой следующими показателями [15]:

- эффективность – точность и полнота, с которой пользователи достигают поставленных целей (процент достигнутых целей и процент пользователей, успешно завершивших задачу);
- продуктивность – отношение израсходованных ресурсов к точности и полноте, с которой пользователи достигают поставленных целей (время необходимое на завершение задачи и задачи, выполненные в единицу времени);
- удовлетворенность – комфорт и приемлемость использования. Оценивается как восприятие пользователем таких показателей, как достоверность, полезность или легкость в нахождении необходимой информации.

Событие – факт, зафиксированный в определённый момент времени при взаимодействии определенного пользователя на определенном устройстве с информационной поисковой системой [8]. Событие обладает непустым уникальным набором атрибутов: пользователь, устройство, время, тип события (например, событие действия, командное событие) и специальные атрибуты, зависящие от типа события. Обозначим E множество всех зафиксированных событий: $E = \{e_1, ..., e_n\}$, где $\{e_i\}$, $i = \overline{1,n}$ – отдельные события; n – мощность множества E.

Сессия – зафиксированный временной промежуток, в течение которого пользователь взаимодействовал с поисковой системой [9]. В рамках сессии все события накапливаются и хранятся в хронологическом порядке. Важно отметить, что каждое событие является уникальным и может быть включено только в одну сессию. Обозначим S множество всех зафиксированных сессий: $S = \{s_1, ..., s_m\}$, где $\{s_i\}$, $i = \overline{1,m}$, – отдельные сессии; m – мощность множества S.

Сессия представляет собой размещение элементов множества *E* без повторений: $s_i = \langle e_{i1}, \dots, e_{ij} \rangle$, где $\{e_{ij}\}, i = \overline{1,m}$, $j = \overline{1,l_i}$ – отдельное событие *i*-й сессии; l_i – мощность размещения s_i.

Множество *E* формируется в результате объединения всех множеств сессий, полученных из данных активности пользователей. При этом любое событие принадлежит хотя бы одной сессии: $\forall e \in E, \exists_s \in S, e \in s$.

Размещение – упорядоченный набор элементов множества либо с повторениями, либо без повторений в соответствии с критериями ассоциативных правил классификации [4]. Заметим, что в размещении без повторений, которым изначально является сессия, невозможно существование часто повторяющихся паттернов как упорядоченных подмножеств. Поэтому требуется предварительная классификация всех событий [7].

Классом событий назовем произвольную совокупность событий, обладающих определенным свойством или признаком. Обозначим C_E множество всех определенных классов событий: $C_E = \{c_1, ..., c_k\}$, $\{c_i\}$, где $i = \overline{1,k}$, – отдельный класс; k – мощность множества C_E.

Паттерн представляет собой размещение элементов множества C_E с повторениями: $p_i = \langle c_{i1}, \dots, c_{ij} \rangle$, где $\{c_{ij}\}$, $i = \overline{1,r}$, $j = \overline{1,q_i}$, $c_{ij} \in C_E$ – отдельное событие *i*-го паттерна; r – мощность множества *P*; q_i – мощность размещения p_i, то есть совокупность событий в паттерне.

Каждое событие описывается непустым множеством признаков *M* (набором атрибутов), присущим объекту множества *E*. Предполагаем, что все объекты в *E* и признаки в *M* различны. Пусть задано отношение $I \mid E \times M$ инцидентности между множествами *E* и *M*. Существование в *I* пары *(e, m)*, $e|E$ и $m|M$, означает, что объект *e* имеет признак *m* и, наоборот, признак *m* характерен для объекта *e*. Тройку *K=(E,M,I)* принято называть контекстом поиска ассоциативных правил [5, 7].

Ассоциативное правило *AR=(X,Y)* на множестве признаков контекста *K=(E,M,I)* имеет вид $X \Rightarrow Y$ и количественно характеризуется с помощью двух числовых функций: *sup(X⇒Y)* – поддержка, *conf(X⇒Y)* – достоверность. Поддержка и достоверность ассоциативного правила определяются через понятие поддержки множества признаков *M*.

Задача поиска последовательных паттернов заключается в обнаружении максимальных последовательностей, имеющих поддержку выше заданного порога. Требуется найти для заданного контекста K_p*=(E,M,I)* множество *AR* всех ассоциативных зависимостей последовательных шаблонов *P*. Заметим, что искомый набор правил *AR* параметризирован относительно пороговых значений *sup* и *conf*. Для классификации событий вычисляется целевая функция (ЦФ) достоверности правила согласно:

$$conf(X \Rightarrow Y) = max\, f\,(E \rightarrow C_E). \quad (1)$$

2. Метод поиска закономерностей в последовательностях событий. Секвенциальный анализ (sequential pattern mining, поиск/добыча последовательных паттернов) – это метод интеллектуального анализа данных (data mining), объектом которого является база последовательностей – кортежа из наборов элементов (itemsets) – непустых множеств одновременно встречающихся элементов [2, 10]. Применительно к задаче поисковой деятельности пользователей, набор элементов будет соответствовать содержимому одного запроса, а вся последовательность будет представлять собой совокупность Интернет-ресурсов, просмотренных пользователем за всё время наблюдения. Пример последовательности с указанием её компонентного состава показан на рис. 1. Целью секвенциального анализа являет-

ся получение часто встречающихся подпоследовательностей классов событий C_E в заданной сессии S, удовлетворяющие ограничению минимальной поддержки, которые называются последовательными паттернами P [2, 16].

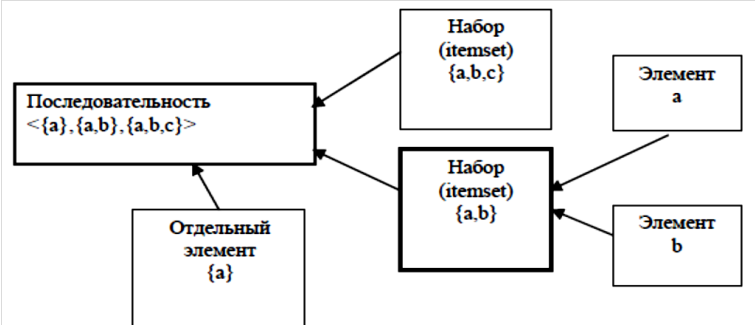

Рис. 1. Структура последовательного паттерна

Рассмотрим формализацию задачи поиска последовательных паттернов на основе модели поведения пользователя в поисковой сессии [13].

Пусть C_E множество всех определенных классов событий: C_E = {c_1, ..., c_k}, {c_i}, где $i = \overline{1,k}$, – отдельный класс; k – мощность множества C_E.

Обозначим через S' множество всех зафиксированных сессий, после классификации и фильтрации S' = {s_1', ..., s'_m}, где {s_i'}, $i = \overline{1,m}$, – отдельная сессия; m – мощность множества S'. Сессия после классификации представляет собой размещение элементов множества C_E с повторениями: $s'_i = \langle c_{i1}, \dots, c_{ij} \rangle$, где {$c_{ij}$}, $i = \overline{1,m}, j = \overline{1,l'_i}$ – отдельное событие i-й сессии; l'_i – мощность размещения s'_i.

Введем n'суммарное количество событий в сессиях после классификации и фильтрации: $n' = \sum_{i=1}^{m} l'_i$, где m – мощность множества S'; l'_i – мощность размещения s'_i.

Обозначим P множество кандидатов последовательных паттернов: P = {p_1, ..., p_r}, где {p_i}, $i = \overline{1,r}$ – отдельный паттерн; r – мощность множества P. Будем считать длиной последовательного паттерна или сессии количество событий в указанном размещении, а также то, что паттерн p входит в сессию s', если все элементы p содержатся в s', при этом порядок элементов в подмножестве из s' соответствует порядку элементов p.

Примем за $\mu_p^{s'}$ значение функции принадлежности i-го ассоциативного признака паттерна, вычисленное как количество непересекающихся упорядоченных вхождений p в сессию s'. Тогда паттерн p называется поддерживаемым сессией s', и как следствие, поддерживаемым пользователем, если количество вхождений $\mu_p^{s'}$>0.

Выразим sup(X⇒Y) через $\lambda_p^{s'}$ поддержку p сессией s', рассчитываемую как:

$$\lambda_p^{s'} = \frac{\mu_p^{s'} * q}{l}, \; 0 \le \lambda_p^{s'} \in \mathbb{R} \le 1, \qquad (2)$$

где $\lambda_p^{s'}$ – количество вхождений p в сессию s'; q – длина паттерна p; l – длина сессии s'.

Таким образом, для одной сессии можно описать значение поддержки как долю содержания паттерна в сессии. Данное условие необходимо для сравнения разных паттернов по степени влияния на процесс генерации максимальных паттернов взаимодействия пользователя с Интернет-ресурсами [17].

Число сессий пользователя может быть различным, поэтому необходимо агрегировать значение поддержки, сохранив их семантику. Рассчитаем общую поддержку паттерна как взвешенную среднюю арифметическую согласно формуле:

$$\lambda_p^{s'} = \sum_{i=1}^{m} \left(\frac{\mu_p^{s_i'} * q}{l_i} * \frac{l_i}{n'} \right) \leq \lambda_p^{s'} \in \mathbb{R} \leq 1, \quad (3)$$

где S' – множество сессий после классификации и фильтрации; m – мощность множества S'; $\mu_p^{s_i'}$ – количество вхождений паттерна p в сессию s_i'; q – длина паттерна p; l' – длина сессии s_i'; n' – суммарное количество событий в сессиях после классификации и фильтрации.

Формулы (2) и средняя ЦФ (3) позволяют рассчитывать значение поддержки различных последовательных паттернов активности пользователей. Сравнение числовых значений поддержки позволяет ранжировать паттерны по степени приоритета для дальнейшего расчета ЦФ достоверности.

Для оценки эффективности поисковой деятельности пользователя необходимо рассчитать затраченные ресурсы пользователя относительно выбранных паттернов, которые содержат не исходные события, а классы, не имеющие информации о реальной длительности временных интервалов между событиями.

Для преобразования классов в затрачиваемое время вводится скалярная функция t, которая может быть записана следующим образом: $t : C_E \rightarrow \mathbb{R}_+$, где C_E – множество всех определенных классов событий; $\mathbb{R}_+$ – множество положительных вещественных чисел. На данный момент авторы предлагают, чтобы функцию t определял эксперт на основе знаний о возможном времени исполнения поисковой сессии.

Представим математическую модель поискового поведения пользователей с учетом затраченного времени как:

$$M_{User} = \{E, C_E, f, S, S', P, \Lambda, t\}, \quad (4)$$

где $E=\{e_i\}$, $i = \overline{1,n}$ – множество событий с атрибутами; $C_E = \{c_i\}$, $i = \overline{1,k}$ – множество классов событий; f: $E \rightarrow C_E$ – функция классификации событий; $S = \{s_i\}$, $i = \overline{1,m}$ – множество сессий до классификации, $s_i = \langle e_{i1}, \dots, e_{il} \rangle$, $i = \overline{1,m}$ – сессия до классификации, l – длина i-й сессии; $S' = \{s_i'\}$, $i = \overline{1,m}$ – множество сессий после классификации и фильтрации, $s_j' = \langle c_{i1}, \dots, c_{il} \rangle$, $i = \overline{1,m}$ – сессия после классификации и фильтрации, l' – длина i-й сессии после классификации и фильтрации; $P = \{p_i\}$, $i = \overline{1,r}$ – множество последовательных паттернов, $p_i = \langle c_{i1}, \dots, c_{iq} \rangle$, $i = \overline{1,r}$ – последовательный паттерн; $\Lambda=\{\lambda_{p_i}^{s'} \in \mathbb{R} | 0 \leq \lambda_p^{s'} \leq 1, p_i \in P, i = \overline{1,r}\}$ – множество значений поддержки последовательных шаблонов; t: $C_E \rightarrow \mathbb{R}_+$ – функция преобразования класса событий в затрачиваемое время.

Алгоритм поиска максимальных последовательных паттернов на основе генетической оптимизации. Согласно постановке задачи, база данных (БД) о транзакциях пользователей имеет следующие поля: идентификатор пользователя, время транзакции события поиска и посещенные Интернет-ресурсы. Каждая транзакция соответствует набору элементов, упорядоченных по времени транзакции для формирования последовательности поведения пользователей. Поддержка последовательности – это количество клиентских последовательностей, которые ее содержат. Если поддержка последовательности больше, чем минимальная поддержка, указанная пользователем, мы называем это частой последовательностью. Алгоритм анализа последовательного паттерна обнаружит частые последователь-

ности, называемые последовательными паттернами, среди всех транзакций. Таким образом, цель последовательного исследования паттернов - обнаружить все частые последовательности как последовательные паттерны, которые отражают потенциальные связи внутри элементов из базы данных последовательностей при заданной минимальной поддержке.

В последние десятилетия было предложено много алгоритмов [14, 18–20] для широкого спектра реальных приложений, таких как анализ поведения клиентов [9, 16] и извлечение информации [21] и др. Для повышения вычислительной эффективности метода разработан двухэтапный алгоритм генерации последовательных паттернов, на первом этапе которого выполняется AprioriAll [20], формирующий частые последовательности-кандидаты всех возможных длин, а на втором - генетический алгоритм (ГА) [22] оптимизации входных параметров признакового пространства сгенерированного множества для поиска максимальных паттернов.

В отличие от почти всех существующих алгоритмов SPM, которые итеративно чередуют расширение набора элементов и расширение последовательности, предложенный алгоритм позволяет проводить поиск лучших решений в глубину на основе:

1) анализа закрытого набора элементов на основе алгоритма AprioriAll. Все частые наборы элементов извлекаются, чтобы получить начальный набор последовательностей размера l;

2) генерации максимальных последовательных паттернов на основе генетического поиска. Новые последовательности формируются путем анализа значения достоверности (приспособленности) сформированных на 1 этапе последовательностей с иерархической связью без поиска дополнительных часто встречающихся наборов элементов.

1 этап: работа алгоритма AprioriAll основана на подходе генерации кандидатов, который сканирует базу данных для создания и подсчета последовательных паттернов кандидатов и удаляет те, которые редко встречаются. Обобщенная схема алгоритма AprioriAll с введенными ранее обозначениями, представлена на рис. 2, где S_l – множество всех частых последовательностей, l – длина последовательности, P_l – множество последовательностей кандидатов длины l, p_i – последовательности-кандидаты, входящие в P_l, *Sup* - оператор вычисления поддержки.

Рассмотрим подробнее работу алгоритма. В качестве входных данных определены: база данных последовательностей (SDB), определяемый пользователем порог минимальной поддержки *minsupp* (значение в диапазоне [0, 1], представленное в процентах), максимальная длина последовательности l. Сначала сканируется входная база данных последовательностей один раз, чтобы построить вертикальное представление данных и множества частых последовательностей S_l. Каждая итерация алгоритма предусматривает «проход» по исходному набору данных, в котором каждый частый элемент $p_i \in S_l$, вызывает процедуру поиска [17].

Во время первой итерации вычисляется поддержка для каждого объекта (одноэлементной последовательности P_1) и выполняется фильтрация. В результате исходные данные для следующей итерации алгоритма формируются из последовательностей, чья поддержка равна, либо превышает пороговое значение *minsupp*. Последовательности, не имеющие достаточного уровня поддержи, исключаются. Далее генерируются более длинные последовательности-кандидаты, снова подсчитывается их поддержка и производится фильтрация множества кандидатов последовательны паттернов и их размещение во множестве частых последовательностей S_l зафиксированной сессии, результаты которой послужат исходными данными для работы генетического алгоритма. Число элементов последовательностей–кандидатов на каждой итерации алгоритма одинаково.

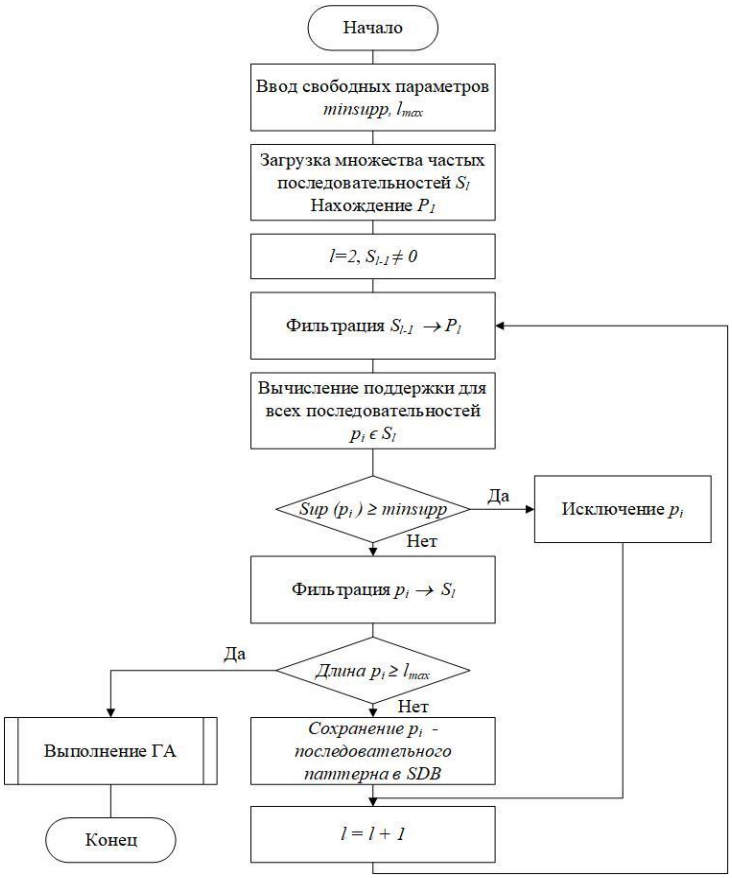

Рис. 2. Обобщенная схема алгоритма на основе AprioriAll

Работа 1 этапа алгоритма завершается тогда, когда не найдено ни одной новой последовательности с достаточным уровнем поддержки в конце очередного шага или, когда невозможно сформировать новых кандидатов.

На рис. 3 представлен фрагмент возможной структуры множества событий S_i-сессии, где красным цветом выделены элементы сформированного последовательного паттерна на 1 этапе алгоритма.

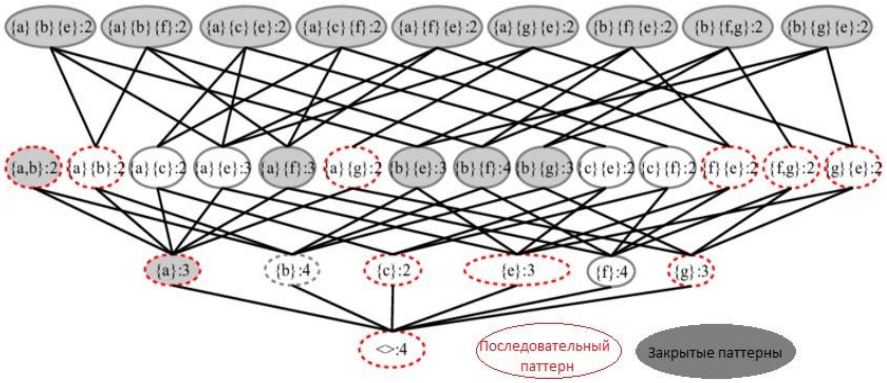

Рис. 3. Фрагмент структуры сформированного паттерна

2 этап: выполнение ГА. Для эффективной настройки параметров решаемой задачи генерации максимальных паттернов, авторы предлагают модифицированный алгоритм генетической оптимизации, структурная схема которого приведена на рис. 4 [22]. Рассмотрим данную структурную схему и опишем назначение каждого ее блока более подробно.

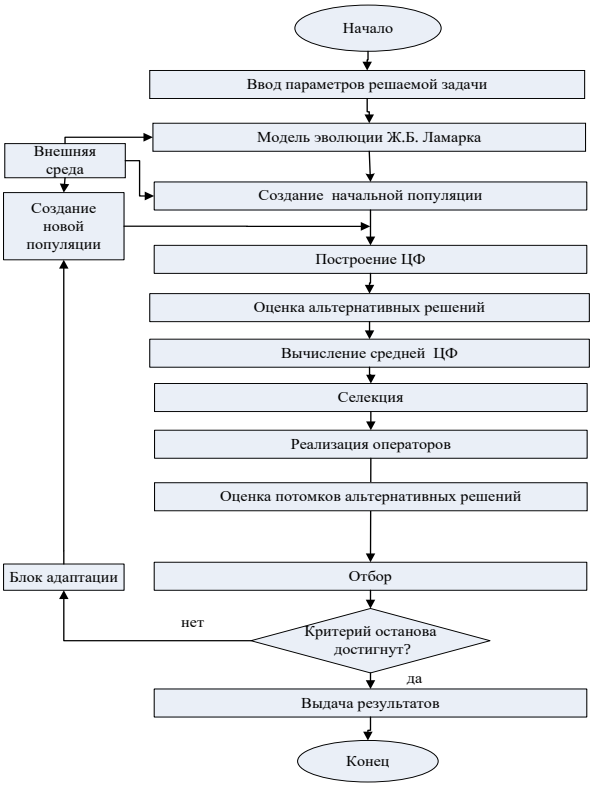

Рис. 4. Структурная схема алгоритма генетической оптимизации

Сначала на предварительном этапе производится ввод свободных параметров решаемой задачи генерации максимальных паттернов: количество сформированных последовательностей на 1 этапе алгоритма в базе данных транзакций – мощность сессии *S*, *M*; длина максимального паттерна, *q*; количество итераций, *N*. Затем происходит ввод таких параметров ГА, как: применяемая модель эволюции и селекции; виды и вероятности операторов поиска; выбор критерия останова алгоритма. В качестве модели эволюции по которому будет реализовываться генетическая оптимизация, авторы предлагают использовать модель эволюции Ж.Б. Ламарка – «наследование благоприобретенных признаков» [23].

Далее производится построение целевой функции и создание начальной популяции решаемой задачи на основе известного принципа «дробовика» [22]. Затем на основе построенной согласно формуле (1) ЦФ производится оценка приспособленности альтернативных решений и вычисляется среднее значение ЦФ популяции всех сгенерированных альтернативных решений по формуле (3).

В следующем блоке формируются родительские пары на основе выполнения турнирной селекции для реализации с заданной вероятностью различных операторов генетической оптимизации [24]. После получения набора потомков производится их оценка.

Далее на основе используемой модели эволюции производится отбор решений для последующей итерации поиска. Затем производится проверка достижения критерия останова, и если он не достигнут, то решения передаются в блок эволюционной адаптации. В этом блоке происходит выработка решений о перестройки текущей популяции и возможности изменения ее размера на каждой итерации поиска, что позволяет алгоритму избегать локальные оптимумы.

Затем решения передаются в блок создания новой популяции альтернативных решений и поиск продолжается итерационно до получения набора квазиоптимальных решений решаемой задачи. При достижении критерия останова – количества итераций алгоритма, производится «выдача результата».

Для управления всем процессом генетической оптимизации авторами введен блок внешней среды. Отметим, что предложенный алгоритм генетической оптимизации позволяет быстро получать более эффективные решения за счет концентрации поиска, т.е. отбирать те наследуемые признаки, которые в последующем будут переданы потомкам.

Экспериментальные исследования. Продемонстрируем результаты экспериментов по секвенциальному анализу поведения пользователей при взаимодействии с информационно-поисковым пространством Интернет с последующим построением максимальных паттернов. Поисковая деятельность демонстрируют типовые последовательности действий в поведении пользователя в процессе поисковой сессии, которую можно представить в виде транзакционной БД, состоящей из последовательностей событий при просмотре Интернет-ресурсов, сохраненных в лог-файлах. Анализ последовательности действий, которую можно обнаружить в логах будем проводить с помощью поиска последовательных паттернов (рис. 5). Каждый лог-файл будем считать единичным набором элементов в последовательностях-кандидатах потоков событий поисковой сессии.

Для экспериментов использовалась коллекция наборов данных MSNBC, библиотеки интеллектуального анализа данных с открытым исходным кодом SPMF [25], содержащая 989 818 последовательностей о потоках событий с веб-сайта MSNBC, преобразованных из исходных данных из репозитория UCI, со средним значением длины последовательности – 14 элементов.

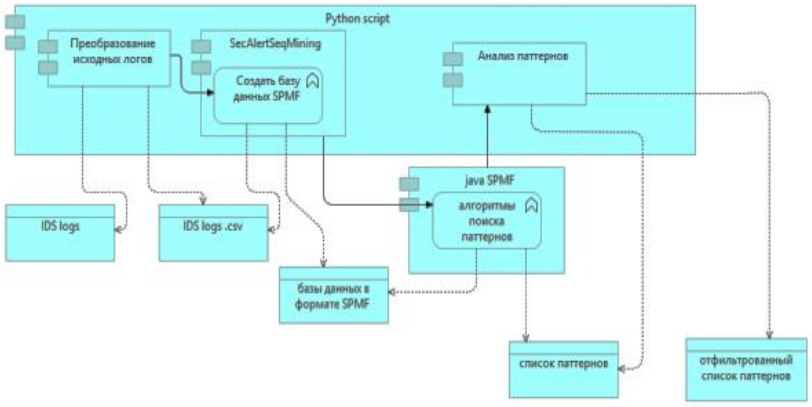

Рис. 5. Схема получения и анализа паттернов

Для анализа были выбраны алгоритмы поиска максимальных последовательных паттернов Generalized Sequential Pattern (GSP) и Vertical Mining of Maximal Sequential Patterns (VMSP). В алгоритме GSP используется горизонтальное внутреннее представление исходных данных и он обладает приемлемой скоростью работы [17]. VMSP – это быстрый алгоритм для обнаружения максимальных после-

довательных паттернов в базах данных последовательностей с вертикальным представлением данных, извлекающий максимальные последовательные паттерны с однократным сканированием базы данных [14].

Для тестирования эффективности алгоритмы сравнивались по числу сгенерированных максимальных паттернов – *N* и времени работы – *t*. В экспериментах оценивались максимальные наборы, для которых вычисляется поддержка $sup(\{p_1, p_2,.. p_n,\}) \geq minsupp$ и для $\forall C_E$ осуществлялась классификация элементов по правилу $sup(\{p_1, p_2,.. p_n, C_E\}) > minsupp$. Для рассматриваемых алгоритмов тестовая серия производилась на разных значениях коэффициента уровня поддержки от 0,01 до 1. Численные результаты экспериментов для алгоритмов по критериям сравнения приведены в таблице, а также на рис. 6 и 7.

Таблица

Результаты сравнения алгоритмов для задачи поиска последовательных паттернов

Значения параметра поддержки	Методы					
	Apriori + ГА		GSP		VMSP	
minsupp, %	t	N	t	N	t	N
1	2	90	5	69	2	80
0,5	2,5	166	6	106	3	154
0,3	3	172	8,5	125	4	163
0,2	4	285	9	130	4	186
0,1	5	297	12	162	6	180
0,05	6	305	17	170	8	232
0,03	8	380	25	292	12	304
0,02	9	580	32	518	14	419

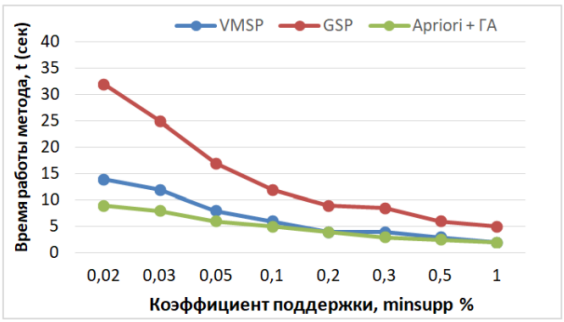

Рис. 6. График зависимости времени извлечения максимальных паттернов от значений уровня поддержки

Методика предобработки данных алгоритмом AprioriAll и ГА позволяет добиться лучших результатов извлечения максимальных паттернов в среднем на 20–40 % по сравнению с алгоритмами GSP и VMSP соответственно. На сравнительных графиках зависимости найденных паттернов и времени работы метода при различных значениях *minsupp*, представленных на рис. 6 и 7 можно оценить, что наилучшие результаты работы метода достигаются при значениях *minsupp = 0,02*. Предложенный метод с оптимизацией на основе ГА по времени работы сопоставим с алгоритмом VMSP и позволяет эффективно идентифицировать максимальные паттерны из большого набора данных при минимальном значении поддержки.

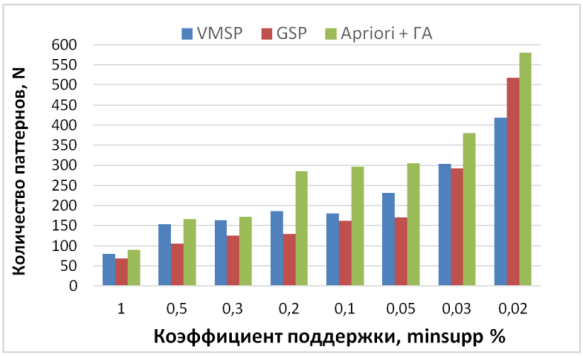

Рис. 7. График зависимости количества сгенерированных максимальных паттернов от значений уровня поддержки

Для определения эффективности разработанного алгоритма были проведены исследования количества выявленных последовательных паттернов на наборах тестовых примеров, различающихся количеством элементов выборки из транзакционной БД при варьировании параметра поддержки *sup*. На рис. 8 показано распределение паттернов, обнаруженных на DataSet1 и DataSet2, содержащих 29248 и 40100 транзакцию событий соответственно. В ходе вычислительного эксперимента было определено лучшее значение длины сформированных последовательных паттернов $l=5$, при котором число паттернов достигает максимального значения при различных варьируемых показателях уровня поддержки. Снижение данного показателя при $l>5$, обусловлено большим количеством различных элементов в множестве часто встречающихся последовательностей. Расчетные данные результатов экспериментов показали, что количество сгенерированных закономерностей (рис. 8) увеличивается с при минимальных значениях *sup*. Так, например, в DS1 для $sup=0{,}0125$ количество $P=550$. Это объясняется тем, что по мере уменьшения уровня поддержки общее число связей увеличивается и количество несвязанных элементов в транзакциях уменьшается.

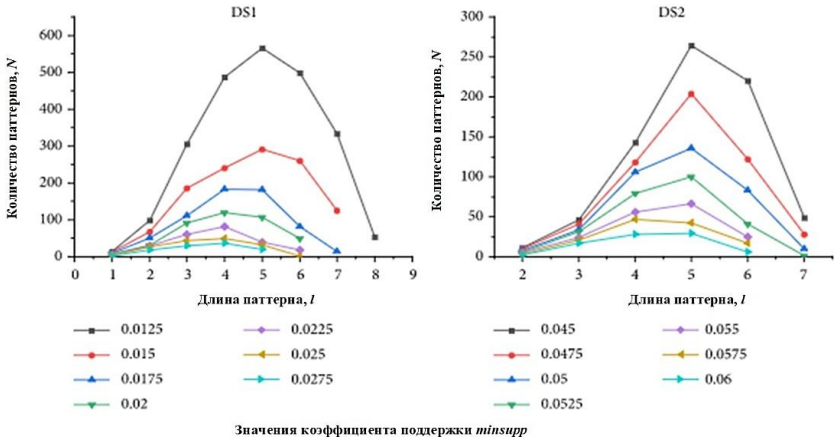

Рис. 8. Графики распределения количества сгенерированных максимальных паттернов от значений уровня поддержки на различных наборах данных

Результаты, полученные в ходе экспериментов доказывают, что стратегия на основе генетической оптимизации снижает время выполнения, так как она значительно сокращает пространство для поиска в тестовых наборах данных. Кроме того, результаты также показывают, что предложенный алгоритм позволяет эффективно извлечь последовательности при низком значении *sup*.

Заключение. Предложен метод получения максимальных последовательных паттернов, использующий методы секвенциального анализа и генетической оптимизации. Метод позволяет эффективно работать с категорийной входной информацией в процессе извлечения ассоциативных зависимостей при меньших временных затратах и избежать затратной процедуры предобработки транзакционной базы данных событий поисковой активности пользователей. Результаты проведенных экспериментальных исследований метода показали, что для увеличения стабильности и точности работы размер выборки, полученной в результате работы ГА, позволит сократить необходимое число сканирований базы данных паттернов, обеспечивая приемлемые вычислительные затраты, сопоставимые с алгоритмом VMSP и превосходящий по времени поиска максимальных паттернов алгоритм GSP более чем на 150 %. Так же результаты экспериментов показали, что ГА может эффективно идентифицировать максимальные паттерны из большого набора данных с низкими значениями показателя поддержки и позволяет извлекать на 20–40 % больше максимальных последовательностей чем алгоритмы VMSP и GSP соответственно. Предложенный метод можно использовать для исследования и анализа данных о поисковом поведении пользователей, выявления аномалий и угроз в информационных событиях их сетевой активности и распознавания возможных рисков информационно-образовательной деятельности в интернет-пространстве.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Tingting Z., Chen L.Y., Liang-Hsien T.* Understanding user motivation for evaluating online content: a self-determination theory perspective // J. Behaviour and Information Technology. – 2015. – Vol. 34. – P. 479-491.
2. *Gupta M., Han J.* Approaches for pattern discovery using sequential data mining. Pattern Discovery Using Sequence Data Mining: Applications and Studies // IGI Global. – 2012. – P. 137-154.
3. *Jalalirad A., Tjalkens T.* Using feature-based models with complexity penalization for selecting features // J. Signal Processing Systems. – 2018. – Vol. 90, Isssue 2. – P. 201-210.
4. *Зайко Т.А., Олейник А.А., Субботин С.А.* Извлечение численных ассоциативных правил с учетом значимости признаков // Восточно-Европейский журнал передовых технологий. – 2013. – Т. 5, № 4 (65). – С. 28-34.
5. *Бова В.В., Щеглов С.Н., Лещанов Д.В.* Modified Approach to Problems of Associative Rules Processing based on Genetic Search // International Russian Automation Conference (RusAutoCon). – 2019. – № 8867675.
6. *Bova V., Kravchenko Yu., Rodzin S., Kuliev E.* Hybrid method for prediction of users' information behavior in the Internet based on bioinspired search // J. of Physics: Conference Series. – 2019. – DOI: 10.1088/1742-6596/1333/3/032008.
7. *Wedyan S.* Review and Comparison of Associative Classification Data Mining Approaches // International Journal of Computer, Information, Systems and Control Engineering. – 2014. – Vol. 8. – P. 34-45.
8. *Оболонный В.И.* Обнаружение последовательностных паттернов в событиях безопасности системы детекции вторжений // Молодой ученый. – 2018. – № 23 (209). – С. 181-187.
9. *Jingjun Zhu GG, Wu Haiyan.* An efficient method of web sequential pattern mining based on session filter and transaction identification // J. Netw. – 2010. – № 5 (9). – P. 1017-1024.
10. *Wang J.-Z., Huang J.-L.* On efficiently mining high utility sequential patterns // Knowledge and Information Systems. – 2016. – Vol. 49, No. 2. – P. 597-627.
11. *Кравченко Ю.А.* Модель фильтра знаний для задач семантической идентификации // Известия ЮФУ. Технические науки. – 2018. – № 4 (165). – С. 197-211.

12. *Bova V.V., Nuzhnov E.V., Kureichik V.V.* The combined method of semantic similarity estimation of problem oriented knowledge on the basis of evolutionary procedures // Advances in Intelligent Systems and Computing. – 2017. – Vol. 573. – P. 74-83.
13. *Бова В.В., Кравченко Ю.А.* Биоинспирированный подход к решению задачи классификации профилей поведения пользователей в интеллектуальных интернет-сервисах // Известия ЮФУ. – 2019. – № 4 (206). – С. 89-102.
14. *Fournier Viger P., Cheng-W., Gomariz A., Tseng V.* VMSP: Efficient Vertical Mining of Maximal Sequential Patterns. – 2014. – DOI: 10.1007/978-3-319-06483-3_8.
15. *Лызь Н.А., Истратова О.Н.* Информационно-образовательная деятельность в интернет-пространстве: виды, факторы, риски // Педагогика. – 2019. – № 4. – С. 16-26.
16. *Truong-Chi T. Fournier-Viger P.* A survey of high utility sequential pattern mining // High-Utility Pattern Mining: Theory, Algorithms and Applications. – Springer, 2019. – P. 97-129.
17. *Fournier-Viger P., Wu C.-W., Tseng V.-S.* Mining Maximal Sequential Patterns without Candidate Maintenance // Proc. 9th Intern. Conference on Advanced Data Mining and Applications. Springer. LNAI 8346. – 2013. – P. 169-180.
18. *Gan W., Lin J. C.-W., Fournier-Viger P., Chao H.-C., Hong T.-P.*A survey of incremental high-utility itemset mining // Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. – 2018. – Vol. 8, No.2. Art. e1242.
19. *Gomariz A., Campos M., Marin R., Goethals B.* ClaSP: An Efficient Algorithm for Mining Frequent Closed Sequences // Proc. 17th Pacific-Asia Conf. Knowledge Discovery and Data Mining. – Springer, 2013. – P. 50-61.
20. *Singh J., Ram H.* Improving Efficiency of Apriori Algorithm Using // Journal of Scientific and Research Publications. – 2013. – Vol. 3. – P. 1-4.
21. *Лежебоков А.А., Кулиев Э.В.* Технологии визуализации для прикладных задач интеллектуального анализа данных // Известия Кабардино-Балкарского научного центра РАН. – 2019. – № 4 (90). – С. 14-23.
22. *Гладков Л.А., Курейчик В.В., Курейчик В.М.* Генетические алгоритмы: учебник. – М.: Физматлит. 2010. – 368 с.
23. *Курейчик В.В., Курейчик В.М., Сороколетов П.В.* Анализ и обзор моделей эволюции // Известия РАН. Теория и системы управления. – 2007. – № 5. – С. 114-126.
24. *Kureichik V., Zaporozhets D., Zaruba D.* Generation of bioinspired search procedures for optimization problems // Application of Information and Communication Technologies, AICT 2016 - 10. – 2016. – P. 7991822.
25. SPMF: an open-source data mining mining library. – https://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php.

REFERENCES

1. *Tingting Z., Chen L.Y., Liang-Hsien T.* Understanding user motivation for evaluating online content: a self-determination theory perspective, *J. Behaviour and Information Technology*, 2015, Vol. 34, pp. 479-491.
2. *Gupta M., Han J.* Approaches for pattern discovery using sequential data mining. Pattern Discovery Using Sequence Data Mining: Applications and Studies, *IGI Global*, 2012, pp. 137-154.
3. *Jalalirad A., Tjalkens T.* Using feature-based models with complexity penalization for selecting features, *J. Signal Processing Systems*, 2018, Vol. 90, Isssue 2, pp. 201-210.
4. *Zayko T.A., Oleynik A.A., Subbotin S.A.* Izvlechenie chislennykh associativnykh pravil s uchetom znachimosti priznakov [Extracting numeric Association rules taking into account the importance of the signs], *Vostochno-Evropeyskiy zhurnal peredovykh tekhnologiy* [East European journal of advanced technologies], 2013, Vol. 5, No. 4 (65), pp. 28-34.
5. *Bova V.V., Shcheglov S.N., Leshchanov D.V.* Modified Approach to Problems of Associative Rules Processing based on Genetic Search, *International Russian Automation Conference (RusAutoCon)*, 2019, No. 8867675.
6. *Bova V., Kravchenko Yu., Rodzin S., Kuliev E.* Hybrid method for prediction of users' information behavior in the Internet based on bioinspired search, *J. of Physics: Conference Series*, 2019. DOI: 10.1088/1742-6596/1333/3/032008.
7. *Wedyan S.* Review and Comparison of Associative Classification Data Mining Approaches, *International Journal of Computer, Information, Systems and Control Engineering*, 2014, Vol. 8, pp. 34-45.

8. *Obolonnyy V.I.* Obnaruzhenie posledovatel'nostnykh patternov v sobytiyakh bezopasnosti sistemy detekcii vtorzheniy [Detection of sequential patterns in security events of the intrusion detection system], *Molodoy uchenyy* [Young scientist], 2018, No. 23 (209), pp. 181-187.
9. *Jingjun Zhu GG, Wu Haiyan.* An efficient method of web sequential pattern mining based on session filter and transaction identification, *J. Netw*., 2010, No. 5 (9), pp. 1017-1024.
10. *Wang J.-Z., Huang J.-L.* On efficiently mining high utility sequential patterns, *Knowledge and Information Systems*, 2016, Vol. 49, No. 2, pp. 597-627.
11. *Kravchenko Yu.A.* Model' fil'tra znaniy dlya zadach semanticheskoy identifikacii [Knowledge filter model for semantic identification tasks], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2018, No. 4 (165), pp. 197-211.
12. *Bova V.V., Nuzhnov E.V., Kureichik V.V.* The combined method of semantic similarity estimation of problem oriented knowledge on the basis of evolutionary procedures, *Advances in Intelligent Systems and Computing*, 2017, Vol. 573, pp. 74-83.
13. *Bova V.V., Kravchenko Yu.A.* Bioinspirirovannyy podkhod k resheniyu zadachi klassifikacii profiley povedeniya pol'zovateley v intellektual'nykh internet-servisakh [Bioinspired approach to solving the problem of classifying user behavior profiles in intelligent Internet services], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2019, No. 4 (206), pp. 89-102.
14. *Fournier Viger P., Cheng-W., Gomariz A., Tseng V.* VMSP: Efficient Vertical Mining of Maximal Sequential Patterns, 2014. DOI: 10.1007/978-3-319-06483-3_8.
15. *Lyz' N.A., Istratova O.N.* Informacionno-obrazovatel'naya deyatel'nost' v internet-prostranstve: vidy, faktory, riski [Information and educational activities in the Internet space: types, factors, risks], *Pedagogika* [Pedagogy], 2019, No. 4, pp. 16-26.
16. *Truong-Chi T. Fournier-Viger P.* A survey of high utility sequential pattern mining, *High-Utility Pattern Mining: Theory, Algorithms and Applications*. Springer, 2019, pp. 97-129.
17. *Fournier-Viger P., Wu C.-W., Tseng V.-S.* Mining Maximal Sequential Patterns without Candidate Maintenance, *Proc. 9th Intern. Conference on Advanced Data Mining and Applications*. Springer. LNAI 8346, 2013, pp. 169-180.
18. *Gan W., Lin J. C.-W., Fournier-Viger P., Chao H.-C., Hong T.-P.*A survey of incremental high-utility itemset mining, *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018, Vol. 8, No.2. Art. e1242.
19. *Gomariz A., Campos M., Marin R., Goethals B.* ClaSP: An Efficient Algorithm for Mining Frequent Closed Sequences, *Proc. 17th Pacific-Asia Conf. Knowledge Discovery and Data Mining*. Springer, 2013, pp. 50-61.
20. *Singh J., Ram H.* Improving Efficiency of Apriori Algorithm Using, *Journal of Scientific and Research Publications*, 2013, Vol. 3, pp. 1-4.
21. *Lezhebokov A.A., Kuliev E.V.* Tekhnologii vizualizacii dlya prikladnykh zadach intellektual'nogo analiza dannykh [Visualization technologies for data mining applications], *Izvestiya Kabardino-Balkarskogo nauchnogo centra RAN* [Izvestiya Kabardino-Balkar scientific center of the Russian Academy of Sciences], 2019, No. 4 (90), pp. 14-23.
22. *Gladkov L.A., Kureychik V.V., Kureychik V.M.* Geneticheskie algoritmy: uchebnik [Genetic algorithms: textbook]. Moscow: Fizmatlit. 2010, 368 p.
23. *Kureychik V.V., Kureychik V.M., Sorokoletov P.V.* Analiz i obzor modeley evolyutsii [Analysis and review of evolution models], *Izvestiya RAN. Teoriya i sistemy upravleniya* [Izvestiya RAS. Theory and control systems], 2007, No. 5, pp. 114-126.
24. *Kureichik V., Zaporozhets D., Zaruba D.* Generation of bioinspired search procedures for optimization problems, *Application of Information and Communication Technologies, AICT 2016 – 10*, 2016, pp. 7991822.
25. SPMF: an open-source data mining mining library. Available at: https://www.philippe-fournier-viger.com/spmf/index.php?link=datasets.php.

Статью рекомендовал к опубликованию д.т.н., профессор Ю.А. Гатчин.

Курейчик Владимир Викторович – Южный федеральный университет; e-mail: vkur@sfedu.ru; 347928, г. Таганрог, пер. Некрасовский, 44; тел.: 88634371651; кафедра систем автоматизированного проектирования; зав. кафедрой; профессор.

Бова Виктория Викторовна – e-mail: vvbova@sfedu.ru; кафедра систем автоматизированного проектирования; доцент.

Кравченко Юрий Алексеевич – e-mail: yakravchenko@sfedu.ru; кафедра систем автоматизированного проектирования; доцент.

Kureychik Vladimir Victorovich – Southern Federal University; e-mail: vkur@sfedu.ru; 44, Nekrasovskiy lane, Taganrog, 347928, Russia; phone: +78634371651; the department of computer aided design; head of department; professor.

Bova Victoria Victorovna – e-mail: vvbova@sfedu.ru; the department of computer aided design; associate professor.

Kravchenko Yury Alekseevich – e-mail: yakravchenko@sfedu.ru; the department of computer aided design; associate professor.

УДК 519.178 DOI 10.18522/2311-3103-2020-4-21-31

Е.М. Герасименко

РЕШЕНИЕ ЗАДАЧИ МОДЕЛИРОВАНИЯ ЧАСТИЧНО РЕВЕРСИВНОГО ПОТОКА МИНИМАЛЬНОЙ СТОИМОСТИ В НЕЧЕТКИХ УСЛОВИЯХ*

Данная статья посвящена разработке алгоритма решения задачи моделирования частично реверсивного потока минимальной стоимости в нечеткой транспортной сети. Задача нахождения потока минимальной стоимости является центральной задачей при планировании перевозок и эвакуационном моделировании. Актуальность такого рода задач обусловлена необходимостью поиска оптимальных с точки зрения стоимости маршрутов перевозок и передачи по ним максимального потока. Данная статья посвящена решению данной задачи в нечетких условиях, так как аппарат теории нечетких множеств позволяет задавать параметры сети, такие как пропускные способности участков дорог, стоимости перевозок в нечётком виде. Такой способ представления удобен в ситуациях, когда имеет место нехватка данных о моделируемом объекте, их лингвистический характер, погрешности в измерениях и пр. В задачах эвакуационного моделирования, которые происходят спонтанно, также наблюдается нехватка точной информации о пропускных способностях и стоимостях перевозок. Концепция контрпотока, используемая в статье, используется для увеличения суммарной пропускной способности путем реверсирования движения. Техника реверсирования движения является современной методикой увеличения передаваемого потока путем увеличения выходной пропускной способности сети. Применение реверсирования движения позволяет освободить загруженные участки дороги и перераспределить движение в сторону пустых дорог, устраняя заторы и «пробки» на дорогах. Предложен метод оперирования нечеткими числами, не приводящий к «размытию» границ результирующего числа и позволяющий оперировать нечеткими границами на последних итерациях, в то время как на остальных предшествующих итерациях производятся вычислениями только с центрами нечетких чисел. Рассмотрен численный пример, который иллюстрирует работу предложенного алгоритма.

Частично реверсивный поток, нечеткий поток минимальной стоимости, транспортные сети.

E.M. Gerasimenko

SOLUTION OF THE PARTIALLY REVERSAL MODELLING TASK OF THE MINIMUM COST FLOW FINDING IN FUZZY CONDITIONS

This article is devoted to the development of an algorithm for solving the problem of modeling a partially reversal flow of minimum cost in a fuzzy transportation network. The minimum cost flow problem is a central problem in transportation planning and evacuation modelling. The relevance of these tasks is due to necessity to find optimal transportation routes in terms of cost and

* Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-01-00559 а.