27. Kurbatskiy V.G., Tomin N.V. Praktika ispol'zovaniya novykh informatsionnykh tekhnologiy dlya prognozirovaniya i analiza otdel'nykh kharakteristik setevykh energopredpriyatiy [The Practice of using new information technologies for forecasting and analysis of individual characteristics of the network utilities], *Problemy energetiki* [Problems of power engineering], 2006, No. 3-4.

Статью рекомендовал к опубликованию д.т.н., профессор В.И. Финаев.

Полуянович Николай Константинович – Южный федеральный университет; e-mail: nik1-58@mail.ru; 347928, г. Таганрог, пер. Некрасовский, 44; тел.: 89185693365; кафедра электротехники и мехатроники.

Дубяго Марина Николаевна – e-mail: w_m88@mail.ru; тел.: 89281758225; кафедра электротехники и мехатроники; аспирант.

Poluyanovich Nikolay Konstantinovich – Southern Federal University; e-mail: nik1-58@mail.ru; 44, Nekrasovsky, Taganrog, 347928, Russia; phone: +79185693365; the department of electric technics and mechatronics.

Dubyago Marina Nikolaevna – e-mail: w_m88@mail.ru; phone: +79281758225; the department of electrical engineering and mechatronics; graduate student.

УДК 004.853 DOI 10.18522/2311-3103-2020-2-66-78

В.В. Бова, Э.В. Кулиев, С.Н. Щеглов

ОЦЕНКА ЭФФЕКТИВНОСТИ МЕТОДА ПОИСКА АССОЦИАТИВНЫХ ПРАВИЛ ДЛЯ ЗАДАЧ ОБРАБОТКИ БОЛЬШИХ ДАННЫХ*

Объемы современных баз данных имеют значительные объемы и содержат большие массивы информации. Одним из популярных методов обнаружения знаний для задач обработки и анализа больших данных стали алгоритмы поиска ассоциативных правил. В статье решается задача построения баз ассоциативных правил для анализа представленной большими массивами неструктурированных данных на основе поиска в них различных закономерностей с учетом значимости их признаков. Предложен метод синтеза баз ассоциативных правил, в котором выполняется построение транзакционной базы данных на основе вычисления пороговых значений поддержки и применением критериев оценивания косвенных ассоциаций, что позволяет извлекать как частые, так и неявные наборы ассоциативных правил. С целью повышения вычислительной эффективности извлечения ассоциативных правил, применяется генетический алгоритм оптимизации входных параметров признакового пространства поиска. Метод позволяет улучшить время извлечения правил, сократить число сгенерированных обобщенных правил, избежать затратной процедуры предобработки синтезированной базы правил. Разработан программно-алгоритмический модуль, с помощью которого проведены экспериментальные исследования метода синтеза ассоциативных правил на основе фильтрации входных параметров модели поиска для решения задач обработки неструктурированных данных. Проведенные серии экспериментов на тестовых транзакционных базах данных позволили уточнить теоретические оценки временной сложности метода, в котором для вычисления взвешенной поддержки наборов правил с учетом оценки априорной информативности признаков, входящих в данный набор применяется генетический алгоритм. Временная сложность разработанного метода составляет $\approx O(I^2)$*. Сравнительной анализ проводился на тестовых данных корпуса Retail Data с алгоритмами Apriori и Frequent Pattern-Growth. Результаты исследований подтвердили эффективность метода поиска на больших наборах транзакций, позволяющего более чем на 40 % уменьшить мощность неизбыточного множества извлеченных ассоциативных правил по сравнению с известными алгоритмами и показали перспективность его применения для задачи обнаружения знаний при обработке данных большого объема.*

Извлечение ассоциативных правил; неструктурированные данные; генетический алгоритм; база ассоциативных правил; большие данные.

* Работа выполнена при поддержке грантов РФФИ № 19-07-00099 и № 18-07-00055.

V.V. Bova, E.V. Kuliev, S.N. Shcheglov

ESTIMATING THE EFFECTIVENESS OF THE METHOD FOR SEARCHING THE ASSOCIATIVE RULES FOR THE TASKS OF PROCESSING BIG DATA

The modern databases have significant volume and consist of large masses of information. One of the popular methods of knowledge identification in terms of tasks of analysis and processing of large data volumes is composed of the algorithms for searching the associative rules. The paper solves the problem of building the bases of associative rules for the analysis of the unstructured large data volumes on the basis of searching different regularities considering the importance of their characteristics. The authors propose the method for synthesizing the bases and building the transaction database to calculate the threshold values of support and application of criteria of estimating implicit associations. This allows us to extract repeated and implicit associative rules. To improve the computational effectiveness of extracting the associative rules, the paper applies the genetic algorithm for optimization of input parameters of the characteristic searching space. The developed method shortens the time of rules extraction, reduces the number of generated common rules, and avoid the resource-consuming procedure of pre-processing the synthesized rule base. The authors developed the program and algorithmic module to carry out the experimental research of the proposed method for synthesizing the associative rules on the basis of filtering the input parameters of the search model for solving the tasks of processing the unstructured data. The experiments conducted on the test transaction bases allow us to clarify the theoretical estimations of time complexity of the proposed method that used the genetic algorithm to calculate the weighed support of the set of rules considering the assessment of a priori informative content of the characteristics included in the dataset. The time complexity of the developed method is estimated as $\approx O(I^2)$*. The comparative analysis is performed using the test data of the Retail Data with the algorithms Apriori and Frequent Pattern-Growth. The results have proven the effectiveness of the search method on big sets of transactions. The method allows us to reduce the cardinal of an irredundant set of extracted associative rules in more than 40 % in comparison with the popular algorithms. The experiments have shown that the method can be effective for the tasks of knowledge discovery in terms of processing large volumes of data.*

Associative rule extraction; unstructured data; genetic algorithm; associative rule base; big data.

Введение. В настоящее время для поиска закономерностей при решении задач, связанных с необходимостью извлечения новых знаний при обработке больших массивов неструктурированных данных широкое распространение получили ассоциативные правила [1–5]. Задача поиска закономерностей в больших данных связана с необходимостью анализа не только качественной, но и количественной информации [1–3, 5]. Исследуемые объекты данных описываются числовыми признаками, поэтому для поиска ассоциативных зависимостей и построения на их основе баз правил необходимо осуществлять предварительную обработку таких атрибутов с учетом значимости признаков [6–8].

Предобработка массивов данных производится на основе разбиения часто встречающихся наборов элементов (признаков) на непересекающиеся множества, каждый из которых представляется новым атрибутом признакового пространства транзакций. В теории ассоциативных правил понятие транзакции трактуется как набор событий или объектов, появляющихся одновременно в некотором наблюдении [6–10]. При решении задачи извлечения новых знаний об исследуемых зависимостях в данных возникают проблемы нахождения всех наборов элементов в том числе и неявных, позволяющих выявлять косвенные ассоциации. Это приводит к значительному возрастанию размерности решаемой задачи и повышает требования к вычислительным ресурсам [8].

Другой проблемой при поиске ассоциативных зависимостей, часто возникающей при анализе больших данных остается значительное число извлеченных правил, большинство из которых являются избыточными. Для устранения избыточности в представлении ассоциативных правил разрабатываются модификации известных алгоритмов поиска, основанные на свойствах частых замкнутых множеств [7–9].

Для оценки значимости выявленных ассоциативных правил производится фильтрация с варьированием параметров принадлежности поддержки (support) и достоверности (confidence) для различных наборов элементов. Анализу подвергаются только те правила, для которых значения мер близости могут быть больше порога минимальной поддержки [7]. Фильтрация по принадлежности элементов транзакций к той или иной группе позволяет снизить количество вычисленных ассоциативных правил, но полностью не решает проблему поиска в признаковом пространстве большой размерности [1–4].

Поэтому актуальной задачей является разработка и исследование эффективности метода синтеза баз ассоциативных правил, свободного от указанных недостатков и позволяющего извлекать не только часто встречающиеся наборы, но и неявные ассоциативные зависимости. *Предлагаемый метод решает задачу* отбора информативных признаков с использованием генетического алгоритма, что *позволяет* эффективно работать с категорийной входной информацией в процессе извлечения ассоциативных правил.

Метод построения базы обобщенных ассоциативных правил. Методы поиска обобщенных правил при вычислении используют информацию о группировке элементов (таксономию), что позволяет значительно расширить круг задач, решаемых алгоритмами поиска ассоциативных правил [6, 10]. При поиске обобщенных правил важной характеристикой, используемой в процессе их извлечения, является поддержка наборов элементов, а также ее пороговое значение, задаваемое экспертом в качестве параметра метода [8, 10–12]. Помимо априорных параметров для извлечения обобщенных правил: поддержки и достоверности представим такой параметр, как уровень интереса, позволяющий учитывать информацию об иерархической группировке элементов для отсечения «неинтересных» правил [6–8]. Сформулируем постановку задачи генерации баз обобщенных ассоциативных правил.

Пусть задан набор (база) транзакций $D = \left\{T_1, T_2, ..., T_{N_D}\right\}$, в котором каждый элемент T_j, $j = 1, 2, ..., N_D$ содержит информацию о последовательности взаимосвязанных событий; $N_D = |D|$ – количество элементов в *D*; $T_j = \left\{t_{1j}, t_{2j}, ..., t_{N_{item\,j}\,j}\right\} \subseteq I$ – *j*-я транзакция базы *D*, определяющая последовательность элементов t_{ij} с конкретным значением числовых атрибутов; t_{ij} – *i*-й элемент *j*-й транзакции T_j, $i = 1, 2, ..., N_{|T_j|}$; $N_{|T_j|}$ – количество элементов в *j*-й транзакции T_j; $I = \left\{\tau_1, \tau_2, ..., \tau_{N_1}\right\}$ – множество возможных переменных (признаков), которые могут входить в набор элементов каждой транзакции T_j, $j = 1, 2, ..., N_D$ набора данных D; а τ– *a*-й элемент множества *I*, $a = 1, 2, ..., N_I$; $N_I = |I|$ – количество элементов множества *I*. Представим задачу синтеза базы обобщенных ассоциативных правил (БП) на основе *D* как генерацию обобщенных ассоциативных правил в виде $\langle X, v(X)\rangle \to \langle Y, v(Y)\rangle$. Обобщенным ассоциативным правилом называется импликация $\langle X, v(X)\rangle \to \langle Y, v(Y)\rangle : X \subset I,\ Y \subset I,\ X \cap Y = \varnothing$, где $v(X)$ и $v(Y)$ – множества значений признаков, принадлежащих непересекающимся множествам *X* и *Y* [7–9].

Для всех транзакций T_j параметр поддержки рассчитывается как пересечение функций принадлежности наборов элементов (признаков) входящих в T_j:

$$\text{supp}(T_j) = \bigcap_{\tau_a \in T_j} \mu_a(T_j), \quad (1)$$

где $\mu_a(T_j)$ – значение функции принадлежности a-го признака, вычисленное для транзакции T_j. Согласно формуле (1) порог поддержки набора X вычисляется суммированием поддержек всех транзакций, содержащих это множество признаков:

$$\text{supp(X)} = \sum_{X \subseteq T_j} \text{supp}(T_j) = \sum_{X \subseteq T_j} \bigcap_{\tau_a \in T_j} \mu_a(T_j). \quad (2)$$

Для вычисления взвешенной поддержки набора X предлагается использовать генетический алгоритм поиска наилучшей оценки индивидуальной информативности признаков, предложенный авторами в работах [13–15]. Согласно формулам (1) и (2) оператор вычисления поддержки, определим следующим образом:

$$\text{wsupp(X)} = \text{supp(X)} \sum_{\tau_a \in X} w_a, \quad (3)$$

где величина $\sum_{\tau_a \in X} w_a$ определяет оценку информативности набора признаков X. Взвешенная поддержка обобщенного правила $X \to Y$ определяется по формуле:

$$\text{wsupp}(X \to Y) = \text{supp}(X \bigcup Y) \sum_{\tau_a \in X \bigcup Y} w_a. \quad (4)$$

Рассмотрим случай, когда X – часто встречающийся взвешенный набор признаков, для которого справедливо условие: $\text{wsupp(X)} \geq \text{wminsupport}$, где *wminsupport* – минимальный порог взвешенной поддержки [7]. Для извлечения частых наборов X установим значение минимальной поддержки $\text{minsupport} = \alpha \cdot \text{wminsupport}$ с учетом α – коэффициента оценки информативности самой длинной иерархической последовательности признаков в БД D для транзакции $T_j\left(\left|T_j\right| = max_{T_J \in D} \left|T_J\right|\right)$ в D:

$$\alpha = \frac{1}{\sum_{T_j\left(\left|T_j\right| = max_{T_J \in D} \left|T_J\right|\right)} w_a}. \quad (5)$$

Особым случаем для извлечения новых знаний об исследуемых объектах в массивах данных большого объема являются косвенные (непрямые) наборы признаков, позволяющие выявлять интересные ассоциативные зависимости [9, 14]. Для их поиска рассмотрим зависимость элементов множеств X и Y косвенно связанных с третьим набором Z, Z: $X \xrightarrow{Z} Y$. Наличие косвенной ассоциативной связи подтверждается, если подтверждается истинность условий [7, 15]:

1) значение взвешенной поддержки набора $X \bigcup Y$ меньше: $wsupp(X \bigcup Y) < \beta_{wsupp(X \bigcup Y)}$, где $\beta_{wsupp(X \bigcup Y)}$ – минимальный порог значения взвешенной поддержки для нечасто встречающихся множеств X и Y, т.е. $\beta_{wsupp(X \bigcup Y)} = wminsupport$;

2) для непустой набора $Z(\exists Z \neq \varnothing)$ справедливы условия:

$$\begin{cases} \text{wsupp}(X \bigcup Z) \geq \beta_{\text{wsupp(Z)}}; \\ \text{wsupp}(Y \bigcup Z) \geq \beta_{\text{wsupp(Z)}}; \end{cases} \text{ и } \begin{cases} w(X,Z) \geq w_{\min}; \\ w(Y,Z) \geq w_{\min}, \end{cases} \quad (6)$$

где $\beta_{\text{wsupp(Z)}}$ – пороговое значение взвешенной поддержки между некоторым набором X и Y и набором Z. Параметр $\beta_{\text{wsupp(Z)}}$ задается по правилу: $\beta_{wsupp(Z)} \geq \beta_{wsupp(X \bigcup Y)}$; $w(X,Z)$ и $w(Y,Z)$ – меры близости взаимосвязи между множествами

X, Y и Z соответственно; w_{min} – минимально допустимое значение показателя значимости взаимосвязи между множествами элементов базы D. Меру оценки близости $w(X,Z)$ рассчитаем по следующей формуле [15]:

$$w(X,Z)=\frac{p(X\bigcap Z)}{\sqrt{p(X)p(Z)}}, \quad (7)$$

где $p(X)$, $p(Z)$, $p(X\bigcap Z)$ – вероятность появления наборов X, Z и $X\bigcap Z$ в базе данных D. Согласно формуле (7), использование предложенных выше критериев и их пороговых значений позволит извлекать не только часто встречающиеся, но и неявные наборы, являющиеся интересными и позволяющие выявлять новые знания об исследуемых транзакционных и категорийных данных.

После выполнения механизма фильтрации, связанной с предобработкой транзакционной БД и нахождением пороговых значений *wminsupport* и *wminconfidence*, выполняется извлечение ассоциативных правил и построение на их основе БП [9].

Для построения базы ассоциативных правил задается транзакционная база данных D, содержащая транзакции T_j с числовыми значениями атрибутов τ_a, определяется набор функций принадлежности μ, используемых для разбиения диапазонов Δ_{ak} численных признаков на множества, определяется минимальная взвешенная поддержка *wminsupport* и взвешенная достоверность *wminconfidence*, задаются пороговые значения $(\beta_{wsupp(X\cup Y)}, \beta_{wsupp(Z)}, w_{min})$, необходимые для работы алгоритма генетического поиска [15].

На следующем этапе производится вычисление мощности k-го диапазона разбиения a-го признака: $C_{\Delta_{ak}}=\sum_{j=1}^{N_D}\mu_{ak}(\tau_a\in T_j)$ и находится его максимальная величина: $max\ C_{\Delta_a}=max_{k=1,2,\dots,N_{\text{инт.а}}}C_{\Delta_{ak}}, a=1,2,\dots,|I|$. Для всех наборов разбиения $max\Delta_a$, $a=1,2,...,|I|$ вычисляется взвешенная поддержка $wsupp(max\Delta_a)$ по формулам (3)-(5). В массив часто встречающихся наборов: $FI_1=\{max_{\Delta_a}|wsupp(max_{\Delta_a})\geq wminsupport\}$ заносятся все интервалы $max\Delta_a$, с допустимым порогом поддержки *wminsupport*. А в массив RI_1:$\{max_{\Delta_a}|wsupp(max_{\Delta_a})\geq wminsupport\}$ – нечасто встречающиеся последовательности элементов, выявленные на множествах с низкими значениями взвешенных поддержек $wsupp(max\Delta_a)$.

На следующем шаге производится формирование множества C_{d+1} $(d+1)$-элементных кандидатов в часто встречающиеся наборы на основе текущего множества FI_d. На этапе генерации множества кандидатов C_{d+1} отсекаются (не извлекаются и не сохраняются в C_{d+1}) те наборы, которые не относятся к часто встречающимся, на основе значений *wsupp*, рассчитанными по формуле (3) и хранящимися во множестве *RI*. По формуле (6), при вычислении нового множества C_{d+1} кандидатов, кандидат X, содержащий подмножество $Y\subset X$, отброшенный на предыдущих этапах как нечасто встречающееся правило $(Y\in RI)$, не отбирается в следующее множество C_{d+1} кандидатов в частые наборы.

После формирования множества C_{d+1} для всех $X=\{\tau_1,\tau_2,...,\tau_{d+1}\}\in C_{d+1}$ $(|X|=d+1)$ вычисляется его степень принадлежности T_j:

$$\mu_X(T_j)=\bigcap_{a:\tau_a\in X}\mu_a(\tau_a\in T_j), \quad (8)$$

далее с учетом формулы (8) определим взвешенную поддержку набора X как:

$$\text{wsupp(X)}=\sum_{X\in T_j,T_j\in D}\mu_X(T_j)\sum_{\tau_a\in X}w_a. \quad (9)$$

Если значение *wsupp*, найденное по формуле (9) не менее минимально допустимого порога *wminsupport*, множество X заносится в массив FI_{d+1}, в противном случае – в массив редко встречающихся наборов RI_{d+1}. В случае, если $FI_{d+1} \neq \varnothing$, выполняются действия, аналогичные описанным выше. Обобщенные правила генерируются исходя из формул (3)-(5) с учетом достоверности *wminconfidence*:

$$\text{wconf}(X \to Y) = \frac{\text{wsupp}(X \to Y)}{\text{wsupp(X)}} \geq \text{wminconfidence}, \; X \bigcap Y = \varnothing.$$

Рассмотренный метод построения базы обобщенных ассоциативных правил предполагает использование критериев для оценивания косвенных ассоциаций, что уменьшает вероятность извлечения правил, некорректно описывающих исследуемые объекты неструктурированной информации, а также позволяет находить не только часто встречающиеся наборы, но и интересные ассоциативные правила.

2. Оценка эффективности метода поиска и синтеза ассоциативных правил. Для исследования эффективности предложенного метода оценим его вычислительную сложность *O*. Извлечение обобщенных правил связано с построением множества часто встречающихся наборов *FI*, что в свою очередь требует вычисления ГА пороговых значений поддержек каждого из кандидатов [15], максимальное количество которых не превышает $|I|^2$. Сложность этого процесса составляет O_{FI} *(I)* операций. Процесс извлечения правил из множества *FI* предполагает обработку каждого подмножества $X \in FI$, на что потребуется $O_{извл.}$ $(|I|^2)$ операций. Поэтому вычислительная сложность предложенного метода составит $O = O_{FI}(|I|^2) + O_{извл.}(|I|^2) = O(|I|^2)$. Поскольку зависимость элементарных операций от размера входных данных является полиномиальной (квадратичной), можно сделать вывод о том, что предложенный метод является вычислительно эффективным.

Для исследования свойств и характеристик предложенного метода построения баз ассоциативных правил была выполнена его программная реализация на языке программирования C# [16]. Программно-алгоритмический модуль позволяет настроить параметры алгоритма поиска обобщенных правил: предельно допустимые значения коэффициента поддержки, минимальную и максимальную достоверность процедуры поиска правил и максимальную мощность множества (рис. 1).

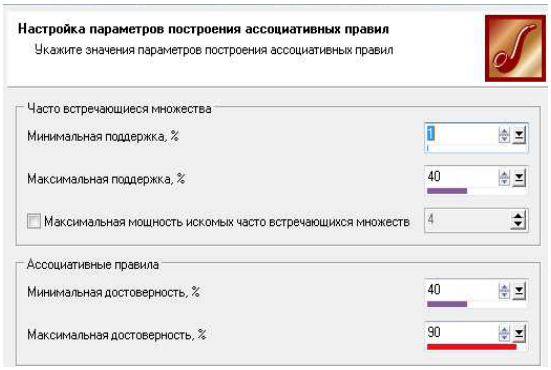

Рис. 1. Интерфейс настройки параметров

После запуска алгоритма поиска вычисления правил, на экране отображается информация о количестве множеств, количестве найденных правил, а также диаграмма распределения найденных часто встречающихся множеств по мощности (рис. 2).

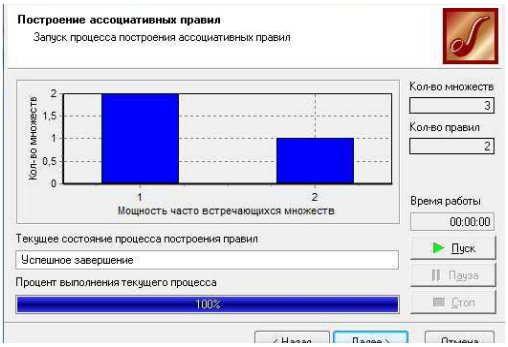

Рис. 2. Процесс выявления ассоциативных правил

Экспериментальное исследование разработанного метода выполнялось с помощью тестовых транзакционных баз данных. Результаты исследований приведены в табл. 1, в которой используются такие обозначения: N_D – количество транзакций T_j в базе D; I – количество элементов (признаков), из которых могли формироваться транзакции, $|T_j|_{ср}$ – среднее количество признаков в транзакциях базы D; $|БП|$ – количество извлеченных обобщенных ассоциативных правил в синтезированной базе правил $БП$; t – время работы метода.

В результате анализа данных, представленных в табл. 1, можно сделать вывод о том, что время работы предложенного метода существенно зависит от количества I элементов в базе D, что подтверждает оценку вычислительной сложности $O(I^2)$ метода.

Расчетные данные результатов экспериментов показали, что количество сгенерированных правил $|БП|$ незначительно увеличивается с ростом параметров N_D, I и T_j базы транзакций D.

Таблица 1

Результаты экспериментальных исследований

№	Характеристики базы транзакций D			Результаты синтеза баз ассоциативных правил	
	N_D	$\|I\|$	$\|T_j\|$	$\|БП\|$	t
1	10000	100	10	212	0,22
2	10000	500	20	257	0,41
3	10000	1000	30	290	0,92
4	50000	100	10	491	0,53
5	50000	500	20	525	2,27
6	50000	1000	30	546	5,82
7	100000	1000	10	612	11,13
8	100000	5000	20	560	14,15
9	100000	10000	30	486	17,68

Так при увеличении среднего числа признаков от 100 до 10000 для постоянных значений элементов транзакций размерности T_j = {10,20,30} количество сформированных взаимосвязей (правил) находится в диапазоне от 20 % до 6 %. Снижение данного показателя обусловлено большим количеством различных элементов в множестве часто встречающихся наборов FI, что позволило сгенерировать оптимальное множество ассоциативных правил. На сравнительных графиках зависимости найден-

ных обобщенных правил и времени работы метода от $|T_j|$ при различных значениях числа транзакций N_D, представленных на рис. 3 и 4 можно оценить, что наилучшие результаты работы метода достигаются при больших значениях N=100000 и увеличении усредненного показателя размерности признакового пространства, при которых количество извлеченных правил уменьшается в среднем на 14 %.

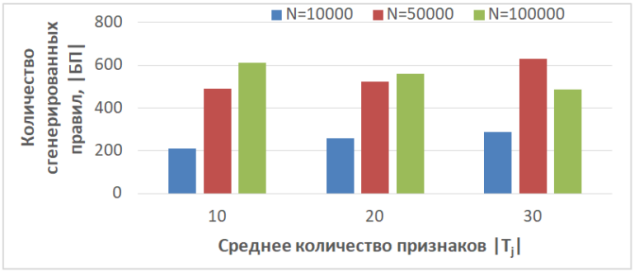

Рис. 3. Гистограмма зависимости количества сформированных ассоциативных правил от числа признаков наборов

Сравнительные исследования показали, что с увеличением числа транзакций и их признакового пространства в БД временные затраты на поиск частых предметных наборов растут пропорционально (рис. 4).

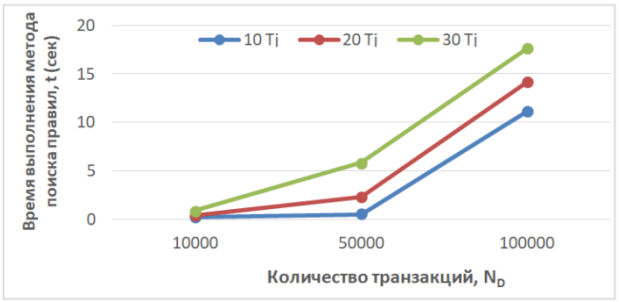

Рис. 4. График зависимости времени генерации ассоциативных правил от числа транзакций в БД

Выявленные факторы и зависимости позволяют сформулировать вывод о том, что метод является эффективным для задач извлечения ассоциативных зависимостей при обработке данных большого объема.

3. Экспериментальные исследования. С целью исследования эффективности метода для задачи поиска ассоциативных правил используется тестовый набор данных, на котором строится методика сравнительной оценки качества с алгоритмами Apriori и Frequent Pattern-Growth (FRG) [17–22]. Алгоритм Apriori предназначен для поиска всех частых множеств признаков. Он является поуровневым, использует стратегию поиска в ширину и осуществляет его снизу-вверх [17]. Apriori использует свойство анти-монотонности нахождения многоэлементных наборов. Благодаря этому свойству перебор не является «жадным» и позволяет снижать размерность пространства поиска и обрабатывать большие массивы информации за приемлемое время [20]. В основе метода FPG лежит предобработка базы транзакций, в процессе которой эта БД преобразуется в компактную структуру, называемую Frequent-Pattern Tree – дерево популярных предметных наборов [19]. Алгоритм FPG позволяет избежать затратной процедуры генерации кандидатов и сократить число проходов БД.

Алгоритмы Apriori, FPG и разработанный метод сравнивались по числу сгенерированных обобщенных правил и времени работы. В экспериментах оценивались максимальные наборы, для которых вычисляется поддержка $supp(\{t_1, t_2,.. t_n,\}) \geq minsupp$ и для $\forall b$ осуществляется фильтрация элементов по правилу $supp(\{t_1, t_2,.. t_n, b\})<minsupp$.

Для определения эффективности разработанного метода были проведены исследования времени и качества извлечения ассоциативных правил для разного набора тестовых примеров, различающихся количеством элементов транзакций T_j корпусов данных. Для тестирования эффективности предложенной методики фильтрации входных данных использовался входной набор корпуса данных Retail Data [23], содержащий 89238 транзакцию, а число различных элементов в ней равно 17562. Для рассматриваемых алгоритмов тестовая серия производилась на разных значениях коэффициента уровня поддержки от 0,2 до 0,001. В ходе вычислительного эксперимента было определено лучшее его значение 0,012. Это объясняется тем, что по мере уменьшения уровня поддержки общее число связей увеличивается и количество несвязанных элементов в транзакциях и сгенерированных наборов уменьшается [24, 25], а алгоритм FPG не может быть завершен за приемлемое время. Численные результаты экспериментов для алгоритмов по критериям сравнения приведены в табл. 2, а также на рис. 5 и 6.

Таблица 2

Результаты сравнения методов для задачи поиска ассоциативных правил

Характеристики транзакций корпуса данных		Методы					
		Apriori		FPG		Предложенный метод	
$\|T_j\|$	|I|	t	|БП|	t	|БП|	t	|БП|
1000	500	16,34	66	1,55	26	1,23	15
5000	500	17,54	72	2,62	35	2,76	17
10000	500	23,79	85	2,96	50	3,54	26
15000	500	25,90	97	3,88	62	4,71	30
20000	500	26,87	105	4,94	70	5,77	32
25000	500	27,12	127	4,98	84	6,85	33

В экспериментальном исследовании оценивались элементы наборов тестовых данных транзакции размерностью от 1000 до 25000 записей со средним количеством признаков |I| равным 500. В результате анализа табл. 2 и графиков зависимостей можно сделать, что алгоритмы FPG и разработанный метод демонстрируют высокую работоспособность в сравнении с алгоритмом Apriori по числу извлеченных строгих ассоциативных правил |*БП*| и по времени работы *t*.

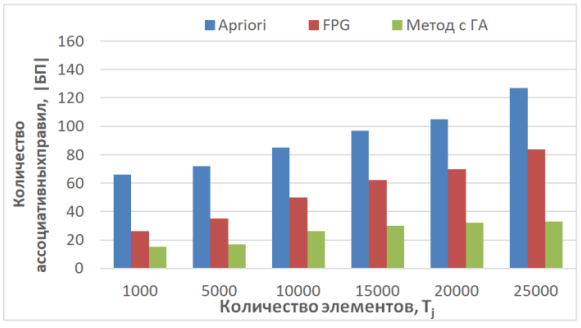

Рис. 5. Гисторгамма сравнения числа извлеченных ассоциативных правил от числа транзакций в БД

Методика предобработки данных позволяет добиться лучших результатов извлечения частых наборов алгоритмом FPG в среднем на 30 %, а с увеличением размерности пространства поиска до 25000 элементов предложенный метод дает улучшение качества сформированных правил более 40 %.

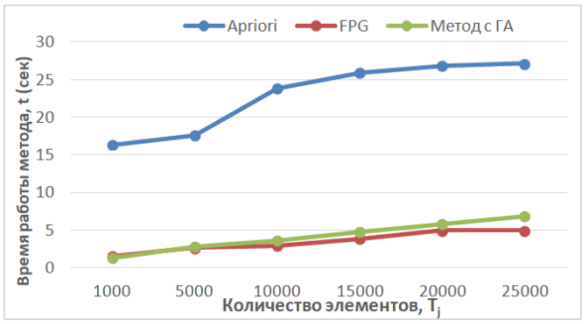

Рис. 6. График зависимости времени извлечения ассоциативных правил от числа элементов транзакций в БД

Предложенный метод с фильтрацией на основе ГА по времени работы сопоставим с алгоритмом FPG и позволяет более чем в два раза уменьшить мощность неизбыточного множества обобщенных ассоциативных правил, формируемого алгоритмом FPG.

Заключение. Предложен метод извлечения ассоциативных правил, основными этапами которого являются: фильтрация случайной выборки исходной базы данных транзакций, на основе ГА вычисления индивидуальной значимости признаков, вычисление пороговых значений поддержки и построение базы обобщенных ассоциативных правил. Отличительной особенностью метода является предобработка базы транзакций, в процессе которой БД преобразуется в компактную структуру, что обеспечивает эффективное и полное извлечение частых предметных наборов, позволяет избежать затратной процедуры генерации кандидатов, по сравнению с известными методами Apriori и FPG.

Разработан программно-алгоритмический модуль, позволяющий выполнять извлечение обобщенных ассоциативных правил. Теоретическая оценка эффективности метода показала, что для увеличения стабильности и точности работы метода размер выборки, полученной в результате работы ГА, позволит сократить необходимое число сканирований базы данных транзакций, обеспечивая приемлемые вычислительные затраты, сопоставимые с известным алгоритмом FPG и превосходящие по времени поиска ассоциативных правил алгоритм Apriori более чем в два раза. Временная сложность разработанного метода составляет $\approx O(I^2)$. Результаты проведенных экспериментов выявили, что предложенный метод более чем в два раза уменьшает мощность неизбыточного множества извлеченных ассоциативных правил, формируемого известными алгоритмами и при значительных показателях параметров транзакционной БД, показывает качество поиска на 4 % выше чем алгоритм FPG. Это подтверждает эффективность метода для решения задачи извлечения ассоциативных правил без потери информации о связях при обработке данных большого объема.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Газиев Г.З., Курдюкова Г.Н., Курдюков В.В.* Кластеризация Big Data для их анализа и обработки // Сб. конференции «Направления и механизмы развития науки нового времени: от теории до внедрения результатов». – 2017. – С. 150-162.

2. *Бова В.В., Щеглов С.Н., Лещанов Д.В.* Модифицированный алгоритм EM-кластеризации для задач интегрированной обработки больших данных // Известия ЮФУ. Технические науки. – 2018. – № 4 (165). – С. 197-211.
3. *Bova V.V., Kureichik V.V., Scheglov S.N., Kureichik L.V.* Multi-level ontological model of big data processing // Advances in Intelligent Systems and Computing. – 2019. – Vol. 874. – P. 171-181.
4. *Wu X., Zhu X., Wu G., Ding W.* Data mining with big data // IEEE Transaction on Knowledge and Data Engineering. – 2014. – Vol. 26. – P. 97-107.
5. *Kravchenko Y.A., Kuliev E.V., Kursitys I.O.* Information's semantic search, classification, structuring and integration objectives in the knowledge management context problems // 10th IEEE International Conference on «Application of Information and Communication Technologies, AICT 2016. – P. 136-141.
6. *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.
7. *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.
8. *Ibrahim S., Chandran K.R.* Compact Weighted Class Association Rule Mining using Information Gain // International Journal of Data Mining and Knowledge Management Process. – 2011. – Vol. 1. – P. 1-13.
9. *Muyeba M., Khan M. S., Coenen F.* Fuzzy weighted association rule mining with weighted support and confidence framework // New Frontiers in Applied Data Mining Lecture Notes in Computer Science. – 2009. – Vol. 5433. – P. 312-320.
10. *Зайко Т.А., Олейник А.А., Субботин С.А.* Ассоциативные правила в интеллектуальном анализе данных // Вестник Национального технического университета Харьковский политехнический институт. Серия: Информатика и моделирование. – 2013. – № 39 (1012). – С. 82-96.
11. *Кравченко Ю.А.* Модель фильтра знаний для задач семантической идентификации // Известия ЮФУ. Технические науки. – 2018. – № 4 (165). – С. 197-211.
12. *Субботин С.А., Олейник Ан.А., Гофман Е.А.* Интеллектуальные информационные технологии проектирования автоматизированных систем диагностирования и распознавания образов: монография / под ред. С.А. Субботина. – Харьков: ООО «Компания Смит», 2012. – 317 с.
13. *Бова В.В., Щеглов С.Н., Лещанов Д.В.* Применение методов генетического поиска для задач обработки ассоциативных правил // XXI Международная конференция по мягким вычислениям и измерениям (SCM-2018). – СПб.: СПбГЭТУ «ЛЭТИ», 2018. – Т. 1. – С. 761-769.
14. *Щеглов С.Н.* Модифицированный алгоритм обработки и анализа неструктурированной информации на основе поиска ассоциативных правил // Тр. Конгресса по интеллектуальным системам и информационным технологиям – «IS&IT'18». – Таганрог: Изд-во ЮФУ, 2018. – Т. 2. – С. 183-191.
15. *Бова В.В., Щеглов С.Н., Лещанов Д.В.* Modified Approach to Problems of Associative Rules Processing based on Genetic Search // 2019 International Russian Automation Conference (RusAutoCon). 10.1109/RUSAUTOCON. – 2019. – № 8867675.
16. *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.
17. *Guo Z., Chi D., Wu J., Zhang W.* A new wind speed forecasting strategy based on the chaotic time series modelling technique and the Apriori algorithm // Energy Conversion and Management. – 2014. – No. 84. – P. 140-151.
18. *Kumar B.S. Rukmani K.V.* Implementation of web usage mining using Apriori and FP Growth algorithms // International Journal of Advanced Networking and Applications. – 2010. – Vol. 400. – P. 400-404.
19. *Пальмов С.В., Франузова Е.Н.* Алгоритм поиска ассоциативных правил FP-GROWTH // Национальная ассоциация ученых. – М.. Изд-во: ООО «Евразийское Научное Содружество», 2016. – № 10-1 (26). – С. 27-32.

20. *Qureshi Z. Bansal S.* Improving Apriori Algorithm to get better performance with Cloud Computing // International Journal of Software and Hardware Research in Engineerin. – 2014. – Vol. 2. – P. 33-37.
21. *Singh J., Ram H.* Improving Efficiency of Apriori Algorithm Using // International Journal of Scientific and Research Publications. – 2013. – Vol. 3. – P. 1-4.
22. *Yahya O., Hegazy O, Ezat E.* An efficient implementation of Apriori algorithm based on Hadoop-Mapreduce model // International Journal of Reviews in Computing. – 2012. – Vol. 12. – P. 59-67.
23. Frequent Itemset Mining Implementations Repository. Retail. – URL: http://fimi.ua.ac.be/data/retail.dat/.
24. *Zhao Y., Zhang C., Cao L.* Post-mining of association rules: techniques for effective knowledge extraction. – New York: Information Science Reference. 2009. – 372 p.
25. *Gkoulalas-Divanis A., Verykios V.S.* Association Rule Hiding for Data Mining. – New York: Springer-Verlag. 2010. – 150 p.

REFERENCES

1. *Gaziev G.Z., Kurdyukova G.N., Kurdyukov V.V.* Klasterizatsiya Big Data dlya ikh analiza i obrabotki [Clusterization of Big Data for their analysis and processing], Sb. konferentsii «Napravleniya i mekhanizmy razvitiya nauki novogo vremeni: ot teorii do vnedreniya rezul'tatov» [Collection of the conference "Directions and mechanisms of modern science development: from theory to implementation of results"], 2017, pp. 150-162.
2. *Bova V.V., Shcheglov S.N., Leshchanov D.V.* Modifitsirovannyy algoritm EM-klasterizatsii dlya zadach integrirovannoy obrabotki bol'shikh dannykh [Modified EM clustering algorithm for integrated big data processing], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2018, No. 4 (165), pp. 197-211.
3. *Bova V.V., Kureichik V.V., Scheglov S.N., Kureichik L.V.* Multi-level ontological model of big data processing, *Advances in Intelligent Systems and Computing*, 2019, Vol. 874, pp. 171-181.
4. *Wu X., Zhu X., Wu G., Ding W.* Data mining with big data, *IEEE Transaction on Knowledge and Data Engineering*, 2014, Vol. 26, pp. 97-107.
5. *Kravchenko Y.A., Kuliev E.V., Kursitys I.O.* Information's semantic search, classification, structuring and integration objectives in the knowledge management context problems, *10th IEEE International Conference on «Application of Information and Communication Technologies, AICT 2016,* pp. 136-141.
6. *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.
7. *Zayko T.A., Oleynik A.A., Subbotin S.A.* Izvlechenie chislennykh assotsiativnykh 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.
8. *Ibrahim S., Chandran K.R.* Compact Weighted Class Association Rule Mining using Information Gain, *International Journal of Data Mining and Knowledge Management Process,* 2011, Vol. 1, pp. 1-13.
9. *Muyeba M., Khan M. S., Coenen F.* Fuzzy weighted association rule mining with weighted support and confidence framework, *New Frontiers in Applied Data Mining Lecture Notes in Computer Science*, 2009, Vol. 5433, pp. 312-320.
10. *Zayko T.A., Oleynik A.A., Subbotin S.A.* Assotsiativnye pravila v intellektual'nom analize dannykh [Associative rules in data mining], *Vestnik Natsional'nogo tekhnicheskogo universiteta Khar'kovskiy politekhnicheskiy institut. Seriya: Informatika i modelirovanie* [Bulletin of the national technical University Kharkiv Polytechnic Institute. Series: computer Science and modeling], 2013, No. 39 (1012), pp. 82-96.
11. *Kravchenko Yu.A.* Model' fil'tra znaniy dlya zadach semanticheskoy identifikatsii [Knowledge filter model for semantic identification tasks], *Izvestiya YuFU. Tekhnicheskie nauki* [Izvestiya SFedU. Engineering Sciences], 2018, No. 4 (165), pp. 197-211.
12. *Subbotin S.A., Oleynik An.A., Gofman E.A.* Intellektual'nye informatsionnye tekhnologii proektirovaniya avtomatizirovannykh sistem diagnostirovaniya i raspoznavaniya obrazov: monografiya [Intelligent information technologies of automated diagnosis and pattern recognition: monograph], ed. by S.A. Subbotina. Khar'kov: OOO «Kompaniya Smit», 2012, 317 p.

13. *Bova V.V., Shcheglov S.N., Leshchanov D.V.* Primenenie metodov geneticheskogo poiska dlya zadach obrabotki assotsiativnykh pravil [Application of genetic search methods for processing associative rules], *XXI Mezhdunarodnaya konferentsiya po myagkim vychisleniyam i izmereniyam (SCM-2018)* [XXI international conference on soft computing and measurement (SCM-2018)]. Saint Petersburg: SPbGETU «LETI», 2018, Vol. 1, pp. 761-769.
14. *Shcheglov S.N.* Modifitsirovannyy algoritm obrabotki i analiza nestrukturirovannoy informatsii na osnove poiska assotsiativnykh pravil [Modified algorithm for processing and analysis of unstructured information based on search for associative rules], *Tr. Kongressa po intellektual'nym sistemam i informatsionnym tekhnologiyam – «IS&IT'18»* [Proceedings of the Congress on intelligent systems and information technologies – "IS&IT'18"]. Taganrog: Izd-vo YuFU, 2018, Vol. 2, pp. 183-191.
15. *Bova V.V., Scheglov S.N., Lemanov D.V.* Modified Approach to Problems of Associative Rules Processing based on Genetic Search, *2019 International Russian Automation Conference (RusAutoCon). 10.1109/RUSAUTOCON*, 2019, No. 8867675.
16. *Lezhebokov A.A., Kuliev E.V.* Tekhnologii vizualizatsii dlya prikladnykh zadach intellektual'nogo analiza dannykh [Visualization technologies for data mining applications], *Izvestiya Kabardino-Balkarskogo nauchnogo tsentra RAN* [Izvestiya Kabardino-Balkar scientific center of the Russian Academy of Sciences], 2019, No. 4 (90), pp. 14-23.
17. *Guo Z., Chi D., Wu J., Zhang W.* A new wind speed forecasting strategy based on the chaotic time series modelling technique and the Apriori algorithm, *Energy Conversion and Management*, 2014, No. 84, pp. 140-151.
18. *Kumar B.S. Rukmani K.V.* Implementation of web usage mining using Apriori and FP Growth algorithms, *International Journal of Advanced Networking and Applications*, 2010, Vol. 400, pp. 400-404.
19. *Pal'mov S.V., Franuzova E.N.* Algoritm poiska assotsiativnykh pravil FP-GROWTH [Search algorithm for associative rules FP-GROWTH], *Natsional'naya assotsiatsiya uchenykh* [National Association of scientists]. Moscow. Izd-vo: OOO «Evraziyskoe Nauchnoe Sodruzhestvo», 2016, No. 10-1 (26), pp. 27-32.
20. *Qureshi Z. Bansal S.* Improving Apriori Algorithm to get better performance with Cloud Computing, *International Journal of Software and Hardware Research in Engineerin*, 2014, Vol. 2, pp. 33-37.
21. *Singh J., Ram H.* Improving Efficiency of Apriori Algorithm Using, *International Journal of Scientific and Research Publications*, 2013, Vol. 3, pp. 1-4.
22. *Yahya O., Hegazy O, Ezat E.* An efficient implementation of Apriori algorithm based on Hadoop-Mapreduce model, *International Journal of Reviews in Computing*, 2012, Vol. 12, pp. 59-67.
23. Frequent Itemset Mining Implementations Repository. Retail. Available at: http://fimi.ua.ac.be/data/retail.dat/.
24. *Zhao Y., Zhang C., Cao L.* Post-mining of association rules: techniques for effective knowledge extraction. New York: Information Science Reference. 2009, 372 p.
25. *Gkoulalas-Divanis A., Verykios V.S.* Association Rule Hiding for Data Mining. New York: Springer-Verlag. 2010, 150 p.

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

Бова Виктория Викторовна – Южный федеральный университет; e-mail: vvbova@sfedu.ru, 347928, г. Таганрог, Некрасовский, 44; тел.: 88634371651; доцент.

Кулиев Эльмар Валерьевич – e-mail: elmar_2005@mail.ru; кафедра систем автоматизированного проектирования; доцент.

Щеглов Сергей Николаевич – e-mail: srg_sch@mail.ru; доцент.

Bova Victoria Victorovna – Southern Federal University; e-mail: vvbova@sfedu.ru; 44, Nekrasovskiy, Taganrog, 347928, Russia; phone: +78634371651; associate professor.

Kuliev Elmar Valerevich – e-mail: elmar_2005@mail.ru; the department of computer aided design; associate professor.

Scheglov Sergey Nikolaevich – e-mail: srg_sch@mail.ru; associate professor.