УДК 004.8

Т.В. Афанасьева, А.А. Сапунков, Д.В. Заварзин

ПРИМЕНЕНИЕ АЛГОРИТМА КЛАСТЕРИЗАЦИИ K-MEANS ДЛЯ УЛУЧШЕНИЯ ТЕМПОРАЛЬНОЙ СТАТИСТИКИ ПРОСМОТРА КОММЕРЧЕСКИХ ПРЕДЛОЖЕНИЙ[1]

Афанасьева Татьяна Васильевна, *доктор технических наук, доцент, заместитель заведующего кафедрой «Информационные системы» Ульяновского государственного технического университета. Окончила радиотехнический факультет УлГТУ. Имеет статьи и монографии в области интеллектуального анализа временных рядов. [e-mail: tv.afanaseva@ulstu.ru].*

Сапунков Алексей Андреевич, *аспирант кафедры «Информационные системы» УлГТУ, окончил факультет информационных систем и технологий УлГТУ. Имеет работы в области интеллектуального анализа временных рядов. [e-mail: sapalks@gmail.com].*

Заварзин Денис Валерьевич, *аспирант кафедры «Информационные системы» УлГТУ, окончил факультет информационных систем и технологий УлГТУ. Имеет работы в области интеллектуального анализа временных рядов. [e-mail: dzavarzin91@gmail.com].*

Аннотация

Аномалии рассматриваются как нетипичные и редко встречающиеся значения, значительно искажающие данные. Обычно такие значения приводят к неточным результатам в процессе анализа данных, поэтому они должны быть удалены. В статье предлагается применение метода кластеризации k-means для решения практической задачи по обработке данных для отображения темпоральной статистики в секторе b2b. Предметной областью и источником данных является сервис отправки и трекинга коммерческих предложений B2BFamily. В статье предлагается удалять аномалии и отображать более адекватную темпоральную статистику о среднем времени просмотра слайда коммерческого предложения. Это поможет менеджеру по продажам корректировать стратегию общения с клиентами. В заключении обсуждаются полученные результаты и дальнейшие тенденции развития данного исследования.

Ключевые слова: кластеризация, аномалии, B2BFamily, алгоритм кластеризации k-means, обнаружение и удаление аномалий.

USING THE K-MEANS CLUSTERING ALGORITHM FOR IMPROVING THE TEMPORAL STATITICS OF COMMERCIAL PROPOSALS VIEWS

Tatiana Vasilevna Afanaseva, *Doctor of Engineering; Associate Professor, Deputy Head of Information Systems Department at Ulyanovsk State Technical University; graduated from the Faculty of Radioengineering of Ulyanovsk State Technical University; an author of articles and monographs in the field of the intellectual analysis of time series. e-mail: tv.afanasjeva@gmail.com.*

Aleksei Andreevich Sapunkov, *Postgraduate Student at the Information Systems Department of Ulyanovsk State Technical University; graduated from the Faculty of Information Systems and Technologies of Ulyanovsk State Technical University; an author of articles in the field of the intellectual analysis of time series. e-mail: sapalks@gmail.com.*

Denis Valerevich Zavarzin, *Postgraduate Student at the Information Systems Department of Ulyanovsk State Technical University; graduated from the Faculty of Information Systems and Technologies of Ulyanovsk State Technical University; an author of articles in the field of the intellectual analysis of time series. e-mail: dzavarzin91@gmail.com.*

Abstract

Anomalies are considered as not typical and rare values, which decrease accuracy in date significant. Such values would generally cause inaccuracy in date analysis results, so they must be deleted. The article proposes to use the k-means clustering method in order to solve practical problems of data processing for displaying the temporal statistics in the B2B sector. The

1 Работа поддержана грантом РФФИ, проект №16-07-00535.

B2BFamily service for sending and tracking commercial offer represents the subject area and the data source. The article also proposes to remove anomalies and display more adequate temporal view statistics about the average time of the commercial offer slide review. That will help the sales manager to adjust the strategy of communication with customers. Finally, the authors discuss the results and trends of the study further development.

Key words: clustering, anomaly, B2BFamily, k-means clustering algorithm, detection and removal of anomalies.

Введение

Задача обнаружения аномалий является актуальной и, как правило, возникает в связи с пропуском значений, появлением значений, выходящих за пределы диапазона допустимого интервала, а также появлением значений, не соответствующих характеру протекания данного процесса.

Это наблюдается в системах, где скорость и объем получаемых данных имеет тенденцию к непрерывному росту. В этих условиях требуется периодический анализ поступающей информации.

Примером такой системы является сервис отправки и трекинга коммерческих предложений B2BFamily (далее сервис) [1]. Ежедневно сервис обеспечивает мониторинг процесса просмотра коммерческих предложений (КП), представленных в виде презентации. На данный момент система отслеживает более 700 КП в день, и это значение растёт в среднем на 10%. Основная задача сервиса – сократить время отклика менеджера и предоставить ему объективную информацию о степени заинтересованности клиента по отправленному ему КП. Весь процесс просмотра КП отслеживается и сохраняется в базе данных (БД) сервиса. На основе этих данных менеджерам предоставляется темпоральная статистика: среднее время просмотра всей презентации и среднее время просмотра отдельного слайда, агрегированное по множеству клиентов (в дальнейшем будем называть это средним по просмотрам). В данном процессе часто встречается аномальная ситуация, когда клиент открыл КП, отошёл от компьютера, отвлёкся или по каким-либо другим причинам перестал смотреть КП, но не закрыл страницу в браузере. В этом случае время просмотра КП будет содержать неадекватно большие числа по сравнению с нормальным поведением. К примеру, клиент потратил на просмотр большинства слайдов в среднем около 20 секунд, а на одном слайде время просмотра составило 2 часа. Такие ситуации могут быть интерпретированы как аномальные, причинами которых может быть временный «уход» клиента. Такие аномально большие значения времени просмотра слайдов КП могут превышать темпоральную статистику в сотни раз и сильно искажать данные о заинтересованности клиента в КП. Поэтому для улучшения темпоральной статистики, характеризующей информацию о заинтересованности клиента, необходимо решить задачу обнаружения и удаления таких аномалий в данных о времени просмотра КП.

Обзор алгоритмов обнаружения аномалий

В настоящее время существует ряд методов для поиска и обнаружения аномалий, которые могут быть успешно применены для анализа данных о времени просмотра слайдов КП, среди которых методы классификации широко используются [2]. Выделяют два основных вида классификации: с учителем и без учителя (кластеризация).

К первому классу относятся методы и алгоритмы, которые обнаруживают аномалии в наборах данных. При этом необходимо иметь достаточное количество примеров в обучающих выборках. Такие алгоритмы получили название алгоритмов классификации с учителем. Основная решаемая задача алгоритмов классификации с учителем – по множеству примеров обучить алгоритм распознавать, содержат данные аномалии или нет [3].

Для обнаружения аномалий в наборе данных на основе методов классификации с учителем применяют:

1. Дерево решения [4];
2. Байесовский классификатор [5];
3. Метод опорных векторов [6].

Ко второму виду алгоритмов классификации данных относят алгоритмы, реализующие обнаружение аномалий в условиях отсутствия обучающей выборки. Данный класс методов и алгоритмов в процессе анализа не учитывает контекста протекания процесса, однако обращает внимание на нетипичные значения характеристик процесса. Наиболее известны следующие методы такого вида:

1. Метод скользящего окна и его модификации [2];
2. Кластерный анализ [7];
3. Нечёткие методы [8].

Каждый из этих методов имеет свои сильные и слабые стороны. Например, точность и адекватность поиска при помощи скользящего окна имеют свои ограничения в зависимости от размера окна и размера его шага в процессе анализа. Методы кластеризации способны неплохо обнаруживать одиночные выбросы, однако их точность снижается с усложнением характера исследуемого процесса. Нечеткие методы подразумевают участие эксперта в процессе выработки правил анализа, кроме того для сложных вычислений данная группа методов требует выделения значительных вычислительных ресурсов. Положительный аспект применения этой группы методов заключается в том, что в случае, когда данных об анализируемом процессе оказывается недостаточно, статистические методы становятся неэффективны.

Большинство алгоритмов ориентированы на обнаружение аномалий типа «выброс». При этом важно идентифицировать не только наличие аномалий, но и связанную с этой аномалией дополнительную информацию. Такой информацией может быть момент времени или пространственное положение аномальных значений.

Так как в нашем случае заранее неизвестно, какие объекты (просмотры) будут аномальны, для анализа был выбран метод кластеризации k-means, как простой и быстрый итеративный алгоритм. Данный алгоритм группирует данные в кластеры в два этапа:

1. Кластеризация всех точек данных в зависимости от расстояния между точкой и ее ближайшим представителем кластера;

2. Переоценка представителей кластера.

К ограничениям алгоритма k-means относят чувствительность k-means к инициализации и определению значения количества кластеров k [2]. Однако, данный алгоритм прост в реализации, и его ограничения нивелируются условиями поставленной задачи.

Алгоритм обнаружения и удаления аномалий

Представим информацию о времени просмотра слайдов презентаций по одному КП в виде матрицы X.

Пусть $X=\{x_{ij}, \forall x_{ij} \in \mathbb{R} \mid i=1,2,\ldots,n,\ j=1,2,\ldots,m\}$, где X – набор данных о времени просмотров по определённому КП, x_{ij} – время i-го просмотра j-го слайда, n – количество просмотров КП, а m – количество слайдов в презентации. Требуется найти темпоральную статистику каждого слайда в виде его среднего времени без аномальных значений $Y=\{y_j, \forall y_{ij} \in \mathbb{R} \mid j=1,2,\ldots,m\}$, где y_j – среднее время просмотра слайда. Темпоральная статистика времени просмотра j-го слайда рассчитывается по формуле:

$$y_j = \frac{\sum_{i=1}^{LENGTH\left(X_j'\right)} X_{ij}'}{LENGTH\left(X_j'\right)}, \quad (1)$$

где набор данных $X_j' = noAnomaly\left(X_j\right)$, $X_j=\{x_{ij}, \forall x_{ij} \in \mathbb{R} \mid i=1,2,\ldots,n\}$, j – номер слайда, статистика которого подсчитывается, $\mathrm{noAnomaly}()$ – алгоритм обнаружения и удаления аномалий, подробно описанный в следующем разделе, $LENGTH$ – количество элементов в X_j'.

Особенностью аномалий в данных о времени просмотров слайдов является их множественность и многоуровневость, причём аномальными значениями должны считаться только те значения, которые значительно отличаются от остальных. Под многоуровневостью в аномалиях подразумевается значительное различие в аномальных значениях. Поэтому после обнаружения и удаления первого уровня аномалий необходимо искать аномалии в оставшемся наборе данных.

В таблице 1 приведены обозначения, которые используется в рассмотренном ниже алгоритме.

Предлагается алгоритм обнаружения и удаления аномалий в данных о просмотрах слайдов презентации КП, основанный на методе кластеризации k-means.

Алгоритм принимает на вход набор данных X_j и возвращает набор данных X_j'. В качестве метрики при кластеризации использована формула:

$$V = \sum_{i=1}^{k} \sum_{X_j \in S_i} \left(X_j - \mu_i\right)^2, \quad (2)$$

где k – число кластеров, S_i – полученные кластеры, $i = 1, 2, \ldots, k$ и μ_i – центры масс векторов X_j принадлежат S_i.

Основная проблема использования данного метода – это необходимость указывать количество кластеров. В то же время при обнаружении аномалий достаточно задать количество кластеров, равное двум. При этом на основе экспертного мнения было выяснено, что количество неадекватных значений (аномалий) не может превышать 10% от общего количества значений, так как если аномалий больше этого порогового значения, то эти данные уже необходимо донести до менеджера, который работает со статистикой. Превышение порога зачастую означает, что при просмотре определенного слайда у клиентов возни-

Таблица 1

Обозначения алгоритма обнаружения аномалий

J	Номер слайда, данные о просмотре которого обрабатываются на предмет наличия аномалий $j = 1, 2, \ldots m$
M	Количество слайдов в презентации
X_j	Набор данных о длительности просмотра j-го слайда за все время наблюдения. $X_j = \left\{x_{ij}, \forall x_{ij} \in \mathbb{R} \mid i = 1,2,\ldots,n\right\}$
X'_j	Набор данных о длительности просмотра j-го слайда без аномалий
N	Количество кластеров
A^N	Набор кластеров, является результатом алгоритма кластеризации, $A^N=\{a_k \mid k = 1,2,\ldots,N\}$
L_k	Мощность кластера a_k
$LENGTH(X)$	Количество просмотров в наборе данных X

кают проблемы, затруднения, а возможно, они полностью теряют интерес к КП, что необходимо отобразить менеджеру в статистике для последующей доработки КП.

Ниже предложен алгоритм обнаружения и удаления аномалий в отдельном слайде презентаций в виде последовательности шагов.

1. Задаем количество кластеров N=2.

2. Применяем метод кластеризации – кластеризуем множество X_j на N кластеров, получаем набор кластеров A^N.

3.

$$\varphi = \begin{cases} true, \text{если } \exists L_k \geq 0{,}9 * LENGTH\left(X_j\right) \middle| k = 1,2,...N, \\ false, \text{во всех остальных случаях.} \end{cases}$$

Если $\varphi = true$, то это значит, что есть кластер мощностью равный или более 90% от общего размер набора данных, а в нем все адекватные значения и возможные аномальные значения более низкого уровня, и, следовательно, в остальных кластерах находятся аномалии более высокого уровня. В таком случае необходимо увеличить количество кластеров $N = N+1$ и вернуться к шагу 2. Если же $\varphi = false$, то получается, что мощность каждого из кластеров меньше 90% от общего размер набора данных. В таком случае переходим к следующему шагу.

4. Так как мощность каждого из кластеров меньше 90% от общего размера набора данных, то берем предыдущее значение N=N–1.

5. Находим кластер с наибольшей мощностью L_k, $X_j' = a_k$. Значения времени просмотра j-го слайда в этом кластере не содержат аномалии типа «выброс» и могут быть использованы для вычисления адекватной темпоральной статистики.

Эксперимент

Целью эксперимента является исследование возможности применения разработанного алгоритма для улучшения темпоральной статистики просмотра КП. Для эксперимента взято КП, которое является презентацией из $m = 33$ слайдов и имеет $n = 852$ просмотров. Этот набор данных был получен из системы анализа КП B2BFamily. График среднего времени просмотра каждого слайда презентации без удаления аномалий представлен на рисунке 1.

Как видно из графика, первый слайд сильно превышает время просмотра всех остальных слайдов и составляет 1195 секунд. Предложенный в этой статье алгоритм разбил набор данных о просмотре первого слайда этой презентации на N=3 кластера, мощности которых $L_1 = 782$, $L_2 = 51$, $L_3 = 19$ соответственно. Кластеры a_2, a_3 были определены как кластеры, содержащие аномальные значения времени просмотра первого слайда:

$a_2 = [3812...13621]$, $a_3 = [14586...30587]$.

После удаления аномальных значений первого слайда его темпоральная статистика в виде среднего времени просмотра уменьшилась в 6,7 раза с 1195 секунд до 178 секунд. Проделав подобные действия с наборами данных по всем остальным слайдам, получили новую темпоральную статистику среднего времени просмотра всей презентации, представленную на рисунке 2.

Результаты эксперимента показали, что данный алгоритм успешно обнаруживает и удаляет аномальные значения в наборах данных о просмотрах презентации и предоставляет более адекватные данные о темпоральной статистике просмотра презентации. Из новой статистики видно, что большинство пользователей задерживаются на первом и последнем слайдах и очень мало времени уделяют просмотру 4, 8, 22 и 31 слайдов.

Для сравнения полученных результатов был взят метод поиска и обнаружения аномалий, основанный на среднем значении. В данном случае аномалиями считались все значения, которые больше среднего в два раза. График с темпоральной статистикой среднего времени просмотра слайдов презентации со сглаженными аномалиями при помощи метода средних значений представлен на рисунке 3. Из данного графика видно, что данный метод удалил не только аномальные значения. При этом общий вид гра-

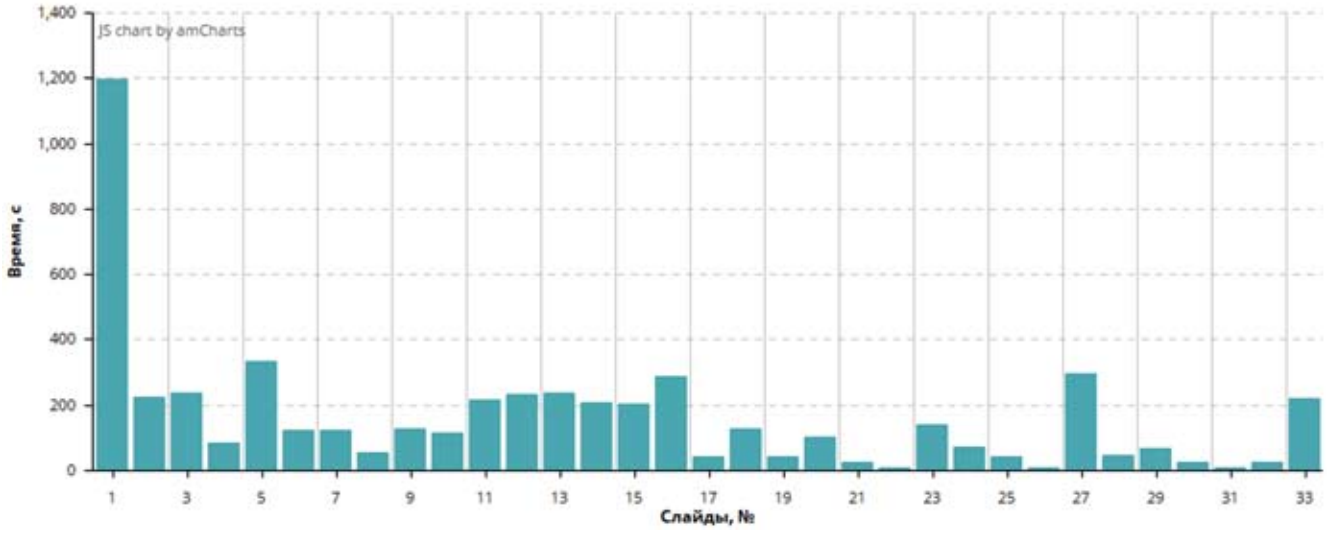

Рис. 1. График среднего времени просмотра слайда презентации, упорядоченный по номеру слайда КП

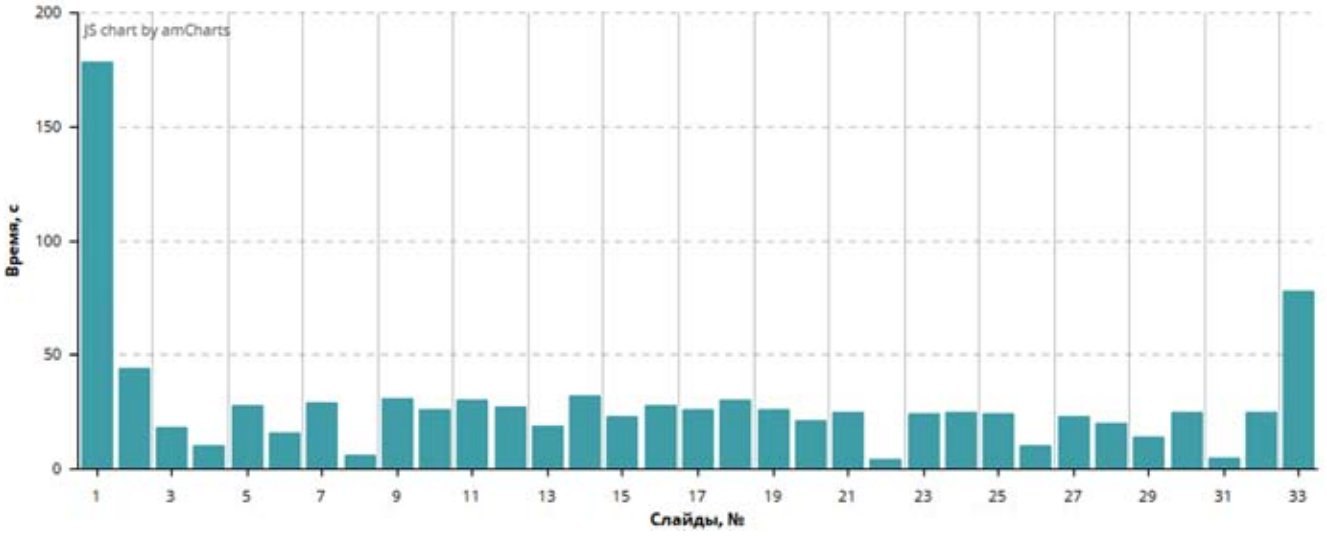

Рис. 2. График среднего времени просмотра презентации со сглаженными аномальными значениями при помощи кластеризации методом k-means

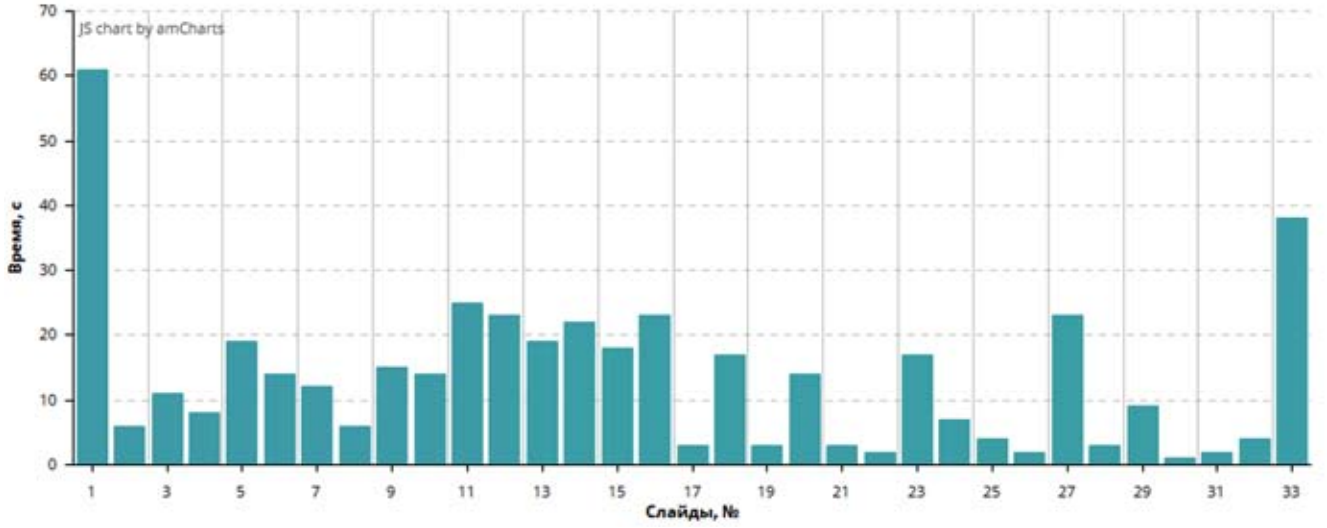

Рис. 3. График среднего времени просмотра презентации со сглаженными аномальными значениями при помощи алгоритма, основанного на средних значениях

фика просмотров слайдов презентации КП не изменился, и темпоральная статистика не стала нагляднее.

Заключение

В данной работе рассмотрена задача поиска и обнаружения аномалий в сервисе отправки и трекинга коммерческих предложений B2BFamily (далее сервис), а также предложен алгоритм, на основе алгоритма кластеризации k-means, для улучшения темпоральной статистики просмотра КП.

В рамках поставленной задачи применение предложенного алгоритма показало, что темпоральная статистика просмотра КП стала наглядной и адаптированной для восприятия клиентом сервиса. Стало видно, каким слайдам при просмотре КП клиенты уделяют значительно меньше времени, что является практической важной бизнес-целью разработки данного алгоритма. В процессе эксперимента были установлены следующие номера слайдов: 4, 8, 22 и 31, которые были классифицированы алгоритмом как аномальные. Это оказался вполне ожидаемый результат, так как при визуальном анализе на данных слайдах было обнаружено мало ценной информации для клиентов.

Дальнейшим развитием данной работы является исследование заинтересованности клиента по данным из просмотра КП, а также по другим точкам контакта с клиентом, таким как количество просмотров презентации, время суток, когда были эти просмотры, и прочее. Это позволит оповещать менеджеров о целевых клиентах и корректировать стратегию общения в зависимости от уровня заинтересованности клиента.

СПИСОК ЛИТЕРАТУРЫ

1. Сервис отправки и трекинга коммерческих предложений B2BFamily. – URL : https://b2bfamily.com/.

2. Deepthi Cheboli. Anomaly Detection of Time Series. Facility Of The Graduate School Of The University Of Minnesota. – 2010. – URL : http://conservancy.umn.edu/.

3. Антипов С.Г. Исследование и разработка методов и алгоритмов обобщения знаний для систем поддержки принятия решений в реальном времени : дис. ... канд. техн. наук. – М., 2016. – 215 с.

4. Петренко С.А. Методы обнаружения вторжений и аномалий функционирования киберсистем // Тр. ИСА РАН. – М., 2009. – Т. 41. – С. 194–202.

5. Браницкий А.А., Котенко И.В. Анализ и классификация методов обнаружения сетевых атак // Тр. СПИИРАН. – 2016. – № 2 (45). – С. 207–244.

6. Зубков Е.В., Белов В.М. Методы интеллектуального анализа данных и обнаружение вторжений // Вестник СибГУТИ. – 2016. – № 1. – С. 118–133.

7. Афанасьева Т.В., Сибирев И.В., Заварзин Д.В. Алгоритм поиска и удаления аномалий временных рядов на основе применения кластеризации // Радиотехника. – 2015. – № 6. – С. 59–62.

8. Афанасьева Т.В., Заварзин Д.В. Алгоритм поиска аномалий в процессах на основе нечётких тенденций временных рядов // Тр. XIV нац. конф. по искусственному интеллекту с междунар. участием КИИ–2014, 24–27 сентября 2014 г., Казань. – Казань : Изд-во РИЦ «Школа», 2014. – Т. 3. – С. 5–12.

REFERENCES

1. *Servis otpravki i trekinga kommercheskikh predlozhenii B2BFamily* [Commercial Offer Sending and Tracking Service]. Available at: https://b2bfamily.com/.

2. Deepthi Cheboli. *Anomaly Detection of Time Series. Facility of the Graduate School of the University of Minnesota*. 2010. Available at: http://conservancy.umn.edu/.

3. Antipov S.G. *Issledovanie i razrabotka metodov i algoritmov obobshcheniia znanii dlia sistem podderzhki priniatiia reshenii v realnom vremeni. Dis. kand. tekhn. nauk* [Researching and Developing of Knowledge Generalization Methods and Algorithms for Decision-Making Support Systems in Real Time. Cand. eng. sci. diss.]. Moscow, 2016. 215 p.

4. Petrenko S.A. *Metody obnaruzheniia vtorzhenii i anomalii funktsionirovaniia kibersistem* [Detection Methods of Cybersystem Functioning Attacks and Anomalies]. *Tr. ISA RAN* [Proc. of Institute for System Analysis of RAS], Moscow, 2009, vol. 41, pp. 194–202.

5. Branitskii A.A., Kotenko I.V. Analiz i klassifikatsiia metodov obnaruzheniia setevykh atak [Analysis and Classification of Methods for Network Attack Detection]. *Tr. SPIIRAN* [SPIIRAS Proceedings], 2016, no. 2 (45), pp. 207–244.

6. Zubkov E.V., Belov V.M. Metody intellektualnogo analiza dannykh i obnaruzhenie vtorzhenii [Techniques for Dynamic Dependence Detection between Groups of Events]. *Vestnik SibGUTI* [Vestnik SibGUTI], 2016, no. 1, pp. 118–133.

7. Afanaseva T.V., Sibirev I.V., Zavarzin D.V. Algoritm poiska i udaleniia anomalii vremennykh riadov na osnove primeneniia klasterizatsii [Algorithm based on Clustering Applications for Searching and Removing Time Series Anomalies]. *Radiotekhnika* [Radioengineering], 2015, no. 6, pp. 59–62.

8. Afanaseva T.V., Zavarzin D.V. Algoritm poiska anomalii v protsessakh na osnove nechetkikh tendentsii vremennykh riadov [The Algorithm of Anomalies Searching in Processes Based on Fuzzy Time Series Trends]. *Tr. XIV nats. konf. po iskusstvennomu intellektu s mezhdunar. uchastiem KII–2014, 24–27 sent. 2014, Kazan* [Proc. of the 14th National Conf. on Artificial Intelligence with Int. Participation KII–2014, 24–27, September, 2014, Kazan]. Kazan, Izd-vo RITs Publ., School Publ., 2014, vol. 3, pp. 5–12.