УДК 004.891.2
DOI: 10.26102/2310-6018/2022.37.2.019

Поддержка принятия решений при анализе эффективности веб-сайта с применением методов Web Usage Mining

А.И. Кокорина[1], Д.А. Петросов[1], А.Н. Зеленина[2]✉

[1]Финансовый университет при Правительстве Российской Федерации (Финансовый университет), Москва, Российская Федерация
[2]Воронежский институт высоких технологий, Воронеж, Российская Федерация
snakeans@gmail.com✉

Резюме. В современных реалиях одним из наиболее эффективных методов не только для поддержания работы своей организации или бизнеса, но и с целью развития является разработка собственного веб-сайта и его дальнейшее использование для коммуникации с пользователями и клиентами. Веб-сайт позволяет систематизировать всю информацию об организации, предоставляет возможность электронной коммерции, а также возможность общения как представителей организации и пользователей, так и самих пользователей между собой для обмена идеями или отзывами о продуктах и услугах. Таким образом, остро ставится вопрос о необходимости анализа эффективности самих веб-сайтов и принятии верного решения по их оптимизации и изменению дизайна, что позволит компании впоследствии достичь поставленные цели. В статье была реализована система поддержки принятия решений для анализа эффективности веб-сайта с применением методов Web Usage Mining. В качестве методов были выбраны статистические, позволяющие улучшать производительность веб-сайта на основе получаемой информации, модифицировать дизайн; и методы интеллектуального анализа данных, в частности, кластеризация и поиск ассоциативных правил, применяемые для персонализации информации и статей, а в случае продающих веб-сайтов – предложений для покупок, что значительно повысит лояльность пользователей и клиентов.

Ключевые слова: система поддержки принятия решений, Web Usage Mining, веб-сайт, лог-файл, машинное обучение, кластеризация, поиск ассоциативных правил.

Для цитирования: Кокорина А.И., Петросов Д.А., Зеленина А. Н. Поддержка принятия решений при анализе эффективности веб-сайта с применением методов Web Usage Mining. *Моделирование, оптимизация и информационные технологии.* 2022;10(2). Доступно по: https://moitvivt.ru/ru/journal/pdf?id=1191 DOI: 10.26102/2310-6018/2022.37.2.019

Support decision-making for analyzing the effectiveness of a website using Web Usage Mining methods

A.I. Kokorina[1], D.A. Petrosov[1], A.N. Zelenina[2]✉

[1]The Financial University under the Government of the Russian Federation (FinU or Financial University), Moscow, Russian Federation
[2]Voronezh Institute of High Technologies, Voronezh, Russian Federation
snakeans@gmail.com✉

Abstract. In the modern world, one of the most effective methods to maintain the functioning of an organization or business with a view to facilitating development is to design a website and then to employ it to communicate with users and customers. The website helps to systematize all information about the organization, provides a means of e-commerce and gives the opportunity for representatives of the organization and users to communicate with each other to exchange ideas or feedback on products or services. Thus, effectiveness analysis of the website and appropriate decision-making, regarding its

optimization and changes to the design, which will allow the company subsequently to achieve its goals, becomes more relevant. In this article, a decision support system was implemented to analyze the effectiveness of a website using Web Usage Mining methods. Statistical methods, which enable performance improvement of the website based on the information received, were chosen as well as data mining methods, in particular, clustering and association rules that are utilized to personalize content and, in the case of selling websites, purchasing offers, which will significantly increase the loyalty of users and customers.

Keywords: decision support system, Web Usage Mining, website, log file, machine learning, clusterization, association rules.

For citation: Kokorina A.I., Petrosov D.A. Support decision making for analyzing the effectiveness of a website using Web Usage Mining methods. *Modeling, Optimization and Information Technology*. 2022;10(2). Available from: https://moitvivt.ru/ru/journal/pdf?id=1191 DOI: 10.26102/2310-6018/2022.37.2.019 (In Russ.).

Введение

В связи с развитием информационных технологий [1] и растущей конкуренцией на рынке возрастает актуальность не только использования веб-сайта в интернете, представляющего интересы компании, но и оценки и повышения его эффективности с целью увеличения дохода или привлечения новых пользователей услуг (покупателей товаров) [2].

Для оценки эффективности веб-сайта (пользуется ли сайт популярностью, есть ли в нем критические ошибки, мешающие течению бизнес-процессов, отталкивает ли что-то потенциальных клиентов и т. д.) необходимо использовать специальные инструменты, позволяющие предоставить статистику и предложить возможные пути решения для дальнейшего принятия решения. На современном рынке распространены такие инструменты, как «Яндекс. Метрика», Google Analytics, Analog Stats и Web Log Explorer [3], позволяющие рассчитать базовые метрики: от количества лидов и посещений веб-сайта до отчетов, указывающих, какие тематики и материалы наиболее популярны, какое количество посетителей совершило целевое действие (например, отправили форму, подписались на рассылку и оставили свои контактные данные). Затем полученная информация анализируется и на ее основе делаются выводы и принимаются решения.

Однако на основе только статистических методов невозможно оптимизировать структуру веб-сайта, персонализировать информацию или обнаружить проблемы в функционировании того или иного блока. Решением, позволяющим расширить функционал подобных инструментов, стало использование методов интеллектуального анализа данных в сети Интернет, включая классификацию, кластеризацию и иные методы машинного обучения – Web Mining [4]. В частности, обработкой журналов использования веб-сайтов занимается подраздел Web Usage Mining, которому и посвящена данная работа.

Основными направлениями применения Web Usage Mining являются:

– Персонализация веб-контента, цель которой основана на предложении ссылок на последующие страницы, в которых пользователи могут быть заинтересованы [5].

– Предварительная выборка и кэширование, которые используются для уменьшения нагрузки на сервер и времени отклика.

– Изменение дизайна. Результаты, полученные с помощью методов Web Usage Mining, могут служить руководством для улучшения дизайна сайтов, приложений с помощью динамической реорганизации согласно данным, полученным из поведения пользователей.

– Электронная коммерция, в которой использование методов Web Usage Mining может влиять на управление взаимоотношениями с клиентами (CRM). Основная цель – увеличить продажи с помощью привлечения клиентов и их удержания, включая перекрестные продажи.

Поэтому целью исследования стала разработка системы поддержки принятия решений с применением методов Web Usage Mining для повышения эффективности веб-сайта.

Применяемые методы и алгоритмы Web Usage Mining

Для проведения дальнейшего исследования и реализации задуманной идеи с точки зрения наивысшей эффективности предлагаем ввести жизненный цикл интеллектуального анализа данных (ИАД). Для этого воспользуемся методологией CRISP-DM (CRoss-Industry Standard Process for Data Mining) [6]. Методология CRISP-DM рассматривает Data Mining как бизнес-процесс, в ходе которого применение технологий интеллектуального анализа данных фокусируется на решении конкретных проблем бизнес составляющей (Рисунок 1).

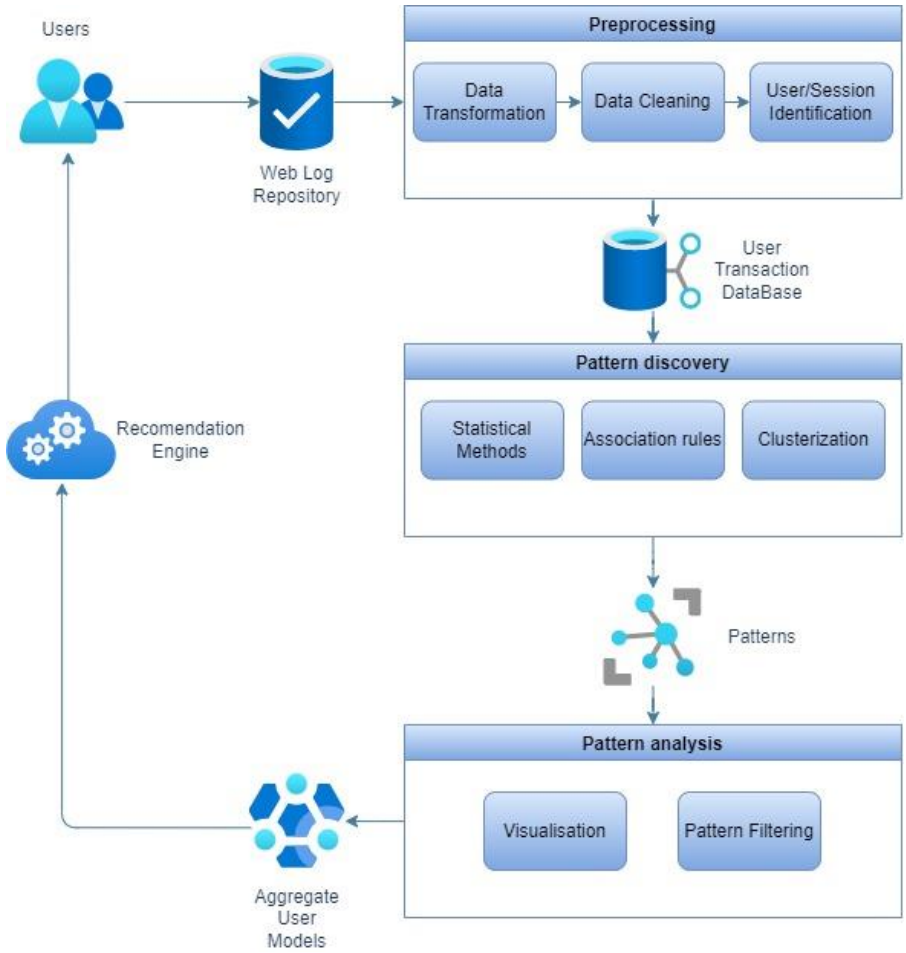

Рисунок 1 – Функциональные модули согласно жизненному циклу ИАД
Figure 1 – Functional modules according to the IAD lifecycle

Этап получения данных Data Understanding. Исходными данными для методов Web Usage Mining являются взаимодействия пользователя (Users) с конкретной страницей или веб-сайтом. На стороне сервера каждый отклик системы на действие

пользователя записывается в веб-лог (или лог-файл). Такой текстовый файл является некой базой данных (Web Log Repository), расположенной на сервере и содержащей каждое действие.

В зависимости от сервера доступны несколько различных форматов журнала лог-файла. Большинство из них имеют общие базовые поля (Таблица 1) и отличаются лишь наличием дополнительной информации. За основу возьмем форматы CERN (European Laboratory for Particle Physics), т. е. CLF, его расширенную версию ECLF и Combined Log Format.

Таблица 1 – Поля лог-файла форматов CERN (European Laboratory for Particle Physics)
Table 1 – Log file fields for CERN (European Laboratory for Particle Physics)

Поле	Характеристика	Пример
remotehost	Название удаленного хоста или IP-адрес.	125.125.125.125
RFC931	Идентификатор по протоколу RFC931.	-
authuser	Идентификация пользователя, прошедшего аутентификацию на веб-сайте.	Dsmith
Date	Дата и время, когда был отправлен запрос.	10/Oct/2021:21:17:05 +0400
request	Запрос клиентского браузера: метод запроса; унифицированный индикатор ресурса (URI); протокол.	GET «/index.html HTTP/1.0
status	Стандартный код ответа HTTP [7].	200
bytes	Размер передаваемых данных.	1876
referrer	URL предыдущего сайта, с которого был перенаправлен клиент.	-
user_agent	Программное обеспечение (веб-браузер), его версия и операционная система.	Mozilla/5.0 (Windows; U; Windows NT 6.0; ru; rv:1.9.0.3) Gecko/2008092417 Firefox/3.0.3
cookies	Файлы, предназначенные для аутентификации личности пользователя.	USERID=CustomerA;IMPID=01234

Этап предобработки данных Data Preparation (Preprocessing). Первоначальным этапом обработки данных является очистка и фильтрация данных. При загрузке страницы, кроме самой страницы, генерируется еще и контент, расположенный на ней, т. е. файлы gif, jpeg, js и подобные. Однако в связи с тем, что этот контент является частью загружаемой страницы, это не оказывает влияния на путь пользователя, пользователь остается на той же странице без переходов. Следующим этапом будут отфильтрованы записи, которые не отражают реальную активность пользователей, т. е. запросы, метод которых отличен от «GET» или «POST» и запросы, выполняющиеся поисковыми ботами. Поведение таких программ отличается от человеческого и не интересно с точки зрения анализа. Крайним этапом является идентификация пользователей и их сессий.

Этап построения моделей Modeling (Pattern discovery и Pattern analysis). В качестве базового алгоритма для выделения ассоциативных правил используется алгоритм Apriori.

Пусть $A = \{A_1, A_2, \dots A_j\} \subset I$, $B = \{B_1, B_2, \dots B_k\} \subset I$. I – объекты, составляющие исследуемые наборы и $A \cap B = \varnothing$. Таким образом, формальное описание правила (1) будет соответствовать следующему виду:

$$A \rightarrow B \tag{1}$$

Первым этапом алгоритма Apriori является обнаружение часто встречающихся наборов элементов:

– Объединение – чтение базы данных и определение частоты вхождения отдельных элементов.

– Отсечение – удаление тех наборов, что не удовлетворяют минимальной поддержке и уверенности.

– Повторение – предыдущие два пункта повторяются для каждой размерности набора до тех пор, пока повторно не получим заранее определенный размер.

– После того, как все часто встречающиеся наборы элементов обнаружены, необходимо приступить к извлечению ассоциативных правил:

– Найти все непустые подмножества часто встречающегося набора F.

– Для каждого подмножества s сформулировать правило $s \rightarrow (F - s)$, если уверенность правила, рассчитанная согласно формуле (2) не меньше минимального порога.

$$conf\left(s \rightarrow (F - s)\right) = \frac{supp(F)}{supp(s)} \tag{2}$$

Анализ ассоциативных правил генерирует достаточно большое количество правил, большинство из которых могут оказаться тривиальными или не представляющими интереса для пользователей. В связи с этим необходимо использовать дополнительные показатели – меры интереса (interestingness measures), которые смогут отфильтровать правила, а также выбрать и ранжировать шаблоны в соответствии с их потенциалом для пользователей (3), (4).

$$supp\ (A \rightarrow B) = \ P(AB) = \ \frac{N(AB)}{|D|}, \tag{3}$$

где N(AB) – доля транзакций, содержащих A и B, |D| – общее количество транзакций в базе данных.

$$conf\ (A \rightarrow B) = \ P\ (B \mid A)\ = \ \frac{P\ (AB)}{P\ (\mathrm{A})} = \frac{supp\ (A \rightarrow B)}{supp\ (A)} \tag{4}$$

Такие показатели, как поддержка support (3) и уверенность confidence (4) являются наиболее распространенными объективными мерами, однако имеют существенный недостаток, поскольку контролируются искусственным пороговым значением, а соответственно не учитывают редкие наборы элементов, которые могут иметь потенциальную ценность. Поэтому в работе предлагается использовать измененные (улучшенные) меры интереса: Bi-Improvement (5) и Bi-Confidence (6) [8].

Для того, чтобы исключить влияние антецедента при его высокой вероятности появления, необходимо применение поправки на отношение вероятности появления антецедента к вероятности не наступления данного события.

$$Bi - imp(A \rightarrow B) = [P(B|A) - P(B)] \cdot \frac{P(A)}{P(\bar{A})} = \frac{P(AB) - P(A)P(B)}{P(\bar{A})} \tag{5}$$

Достоверность, как мера интереса, указывает, что появление одних наборов элементов неизбежно приведет к появлению других. Подобный расчет принимает во внимание только вероятность появления события B при появлении события A, но не учитывает отношения событий A и B, когда событие A не происходит, что делает многие правила недействительными. С целью устранения представленного недостатка предлагается применять улучшенную версию достоверности Bi-conf:

$$Bi-conf(A \to B) = conf(A \to B) - conf(\bar{A} \to B) = \frac{P(AB) - P(A)P(B)}{P(A)\big(1 - P(A)\big)}. \quad (6)$$

Используя структуру Bi-Improvement и Bi-Confidence вместо стандартных мер интереса, можно не только эффективно определять наиболее интересные правила ассоциации, но и уменьшать количество правил со слабой корреляцией.

Для более глубокого анализа предпочтений пользователей и персонализации контента необходимо применить кластеризацию пользователей. Одними из главных характеристик любого алгоритма кластеризации являются набор входных данных и форма получающихся кластеров. Наиболее популярный алгоритм k-средних хоть и является таковым, но имеет существенные недостатки, связанные с разбиением элементов векторного пространства на заранее известное число кластеров k, а также повышенной чувствительностью к выбору начальных центров кластеров. Поэтому для исследования предпочтений пользователей используется алгоритм иерархической агломеративной кластеризации [9], который позволяет визуально изучить количество получаемых кластеров, не определяя изначально их количество.

В качестве метода вычисления расстояния между кластерами (7) рассматривается алгоритм минимизации Уорда [10]:

$$d(U,V) = \sqrt{\frac{|V| + |S|}{Q} * d(V,S) + \frac{|V| + |T|}{Q} * d(V,T) - \frac{|V|}{T} * d(S,T)^2}, \quad (7)$$

где U – кластер, полученный путем объединения кластеров T и S, V – неиспользованный кластер, $Q = |V| + |T| + |S|$.

Входными данными для кластеризации пользователей по интересам является статистика их посещения определенных страниц. Такая статистика представляет многомерное пространство, каждая координата которого соответствует числу посещений конкретного пользователя определенной страницы. В связи с тем, что количество посещаемых страниц, а соответственно и их уникальных URI, могут достигать больших размеров, которые трудно интерпретировать, необходимо первоначально сузить пространство до меньшей размерности. В качестве одного из предлагаемых решений может служить переход от подсчета количества посещений по URI к подсчету количества посещений по тематикам (группам) страниц. Подобное разделение страниц на классы позволит, во-первых, сократить размерность входных данных, во-вторых, повысить универсальность метода и учесть уникальность определенных веб-сайтов при использовании для иного набора данных.

Результаты

В качестве примера лог-файла использовался набор данных «NASA Server Access Logs», содержащий порядка 2 млн записей за месяц. Для применения алгоритма поиска ассоциативных правил после прохождения этапа Preprocessing произведена идентификация сессий пользователей, что привело к получению матрицы размерности 33694 rows × 615 columns, где 33694 – количество сессий за установленный промежуток

времени, 615 – количество уникальных страниц, посещенных пользователями с учетом успешных HTTP-запросов. После расчета часто встречающихся наборов элементов для фильтрации правил используем обновленные меры интереса (Рисунок 2).

antecedents	(/ksc.html)	antecedents	(/history/history.html, /ksc.html)
consequents	(/shuttle/missions/missions.html)	consequents	(/shuttle/missions/missions.html)
antecedent support	0.346382	antecedent support	0.036891
consequent support	0.205942	consequent support	0.205942
support	0.058942	support	0.021547
confidence	0.170165	confidence	0.584071
lift	0.826279	lift	2.836098
Bi-Imp	0.090178	Bi-Imp	0.022372
Bi-Conf	0.346382	Bi-Conf	0.392613
length_antecedents	1	length_antecedents	2
length_consequents	1	length_consequents	1
Name: 78, dtype: object		Name: 260, dtype: object	

Рисунок 2 – Пример работы алгоритма поиска ассоциативных правил Apriori
Figure 2 – Example of the algorithm for associative rules Apriori

Если пользователь посещал страницу «ksc.html» космического центра Кеннеди или интересовался историей миссий «/history/history.html» и в принципе всей космической программы NASA, то согласно извлеченным правилам, ему будет интересна страница «/shuttle/missions/missions.html», посвященная текущим миссиям, осуществляемым на многоразовом транспортном космическом корабле «Спейс шаттл».

Иерархическая кластеризация позволяет построить дендрограмму на основе матрицы пространственных расстояний, по которой можно определить наиболее близких по интересам пользователей (Рисунок 3).

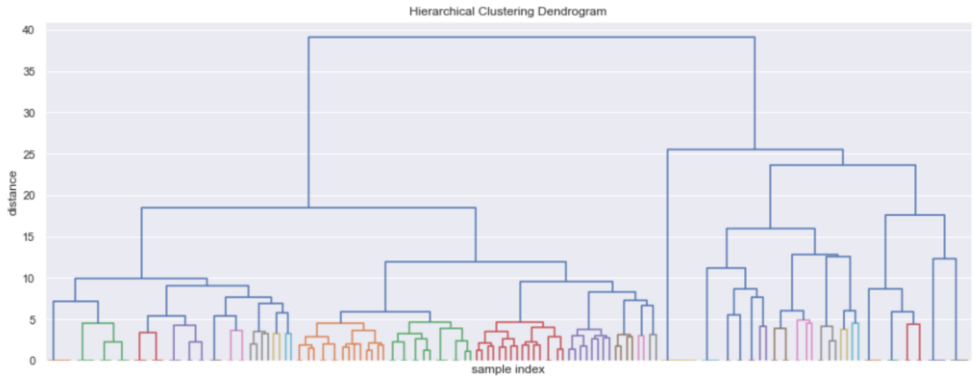

Рисунок 3 – Дендрограмма для приведенного набора данных
Figure 3 – Hierarchical clustering dendrogram

Для выделения количества кластеров применялись оценочные метрики Коэффициент Силуэта (Silhouette Coefficient) [11] и Индекс Дэвиса-Болдина (Davies-Bouldin Index) [12]. Наименьшее целевое значение индекса Дэвиса-Болдина и наибольшее целевое коэффициента Силуэта показывают одинаковые результаты, определяя 7 кластеров как наиболее предпочтительное значение. Поэтому представленный набор данных может быть разделен на 7 кластеров по посещениям веб-страниц с определенными корневыми метками (Рисунок 4), информацию по которым удобно представить в табличной форме для дальнейшего анализа и принятия решения по персонализации. Таким образом, определенному пользователю со своим набором страниц будет соответствовать кластер, в который объединены все посетители с похожими интересами.

Table with clustering results

remotehost	request_URI	group
133.43.96.45	/shuttle/missions/sts-69/mission-sts-69.html	1
133.43.96.45	/shuttle/resources/orbiters/endeavour.html	1
133.43.96.45	/shuttle/missions/sts-72/mission-sts-72.html	1
133.43.96.45	/shuttle/missions/sts-49/mission-sts-49.html	1
133.43.96.45	/shuttle/missions/sts-57/mission-sts-57.html	1
133.43.96.45	/facilities/lc39a.html	1
133.43.96.45	/history/apollo/apollo.html	1
133.43.96.45	/history/apollo/apollo-13/apollo-13.html	1
133.43.96.45	/ksc.html	1

Рисунок 4 – Результаты кластеризации
Figure 4 – The results of clusterization

Обсуждение

С помощью представленного алгоритма Apriori при использовании дополнительных разработанных метрик мы можем отслеживать интересные перемещения пользователя по веб-сайту, которые затем применять для увеличения его эффективности, а именно реорганизации, добавления колонки с рекомендациями и повышения удобства интерфейса. В рамках реализации метода иерархической кластеризации мы можем определить следующие основные направления его использования:

– Обнаружение групп пользователей со схожими предпочтениями; рекомендовать на данной основе те страницы или продукты, которые могли бы их заинтересовать;

– Осуществление поиска отклонения от стандартного поведения конкретного пользователя в связи со сменой интересов или системной ошибкой на веб-сайте, что поможет быстро ее обнаружить и исправить.

Результатом данного исследования стал разработанный сервис поддержки принятия решения для анализа эффективности веб-сайта с использованием методов Web Usage Mining.

Заключение

Для оценки эффективности веб-сайтов применяются системы, позволяющие обрабатывать лог-файлы и анализировать информацию, предоставляемую веб-сайтами. Однако чаще всего подобные инструменты являются «счетчиками», которые фиксируют действия пользователей в текущий момент времени. Для увеличения функционала необходимо применять не только статистические методы, показывающие численное выражение определенных показателей, но и методы интеллектуального анализа данных, включая классификацию, кластеризацию и иные методы машинного обучения.

Используя разработанные алгоритмы, компании смогут повышать эффективность собственного веб-сайта за счет оптимизации его структуры, обнаружения проблем в функционировании блока при изменении стандартного поведения пользователей; уделении внимания тем страницам-продуктам, которые в данный момент времени не являются востребованными; персонализации информации и предложений о покупке/услуге, что увеличивает конкурентоспособность компании на рынке.

Одним из возможных путей развития представленной работы является применение классификации текстов контента веб-сайта как задачи информационного поиска и определение тональности на основе лингвистического анализа.

СПИСОК ИСТОЧНИКОВ

1. Singh D.K. et al. Computational Intelligence in Web Mining. *Innovative Trends in Computational Intelligence*. 2022:197–215. DOI: 10.1007/978-3-030-78284-9_9.
2. Sharma S. et al. Performance Evaluation of Secure Web Usage Mining Technique to Predict Consumer Behaviour (SWUM-PCB). *Intelligent Computing and Networking*. 2022;301:136–145. DOI: 10.1007/978-981-16-4863-2_12.
3. Kandpal N., Singh H.P., Shekhawat M.S. Application of web usage mining for administration and improvement of online counseling website. *Int J Appl Eng Res*. 2019;14(7):1431–1437.
4. Kumar V., Thakur R. S. Web log analysis tools: at a glance. *Proceedings of International Conference on Recent Advancement on Computer and Communication*. 2018;34:135–142. DOI:10.1007/978-981-10-8198-9_14.
5. Haridasan A.C., Fernando A.G. Online or in-store: unravelling consumer's channel choice motives. *Journal of Research in Interactive Marketing.* 2018;12(2):215–230. DOI: 10.1108/JRIM-07-2017-0060.
6. Schröer C., Kruse F., Gómez J. M. A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*. 2021;181:526–534. DOI: 10.1016/J.PROCS.2021.01.199.
7. Hypertext Transfer Protocol (HTTP/1.1) RFC7231: Semantics and Content. Internet Engineering *Task Force (IETF)*. Доступно по: https: // datatracker.ietf.org/doc/html/rfc7231 (дата обращения: 15.04.2022).
8. Somyanonthanakul R., Theeramunkong T. Scenario-based Analysis for discovering Relations among Interestingness Measures. *Information Sciences*. 2022;590:346–385. DOI: 10.1016/j.ins.2021.12.121.
9. Shetty P., Singh S. Hierarchical Clustering: A Survey. International Journal of Applied Research. 2021;7(4):178–181. DOI: 10.22271/ALLRESEARCH.2021.V7.I4C.8484.
10. Jarman A.M. *Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method*. Georgia Southern University. 2020. DOI: 10.13140/RG.2.2.11388.90240.
11. Dinh, D.T., Fujinami, T., & Huynh, V.N. Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. *In International Symposium on Knowledge and Systems Sciences*.2019:1–17. DOI: 10.1007/978-981-15-1209-4_1.
12. Sitompul, Bernad J.D., Opim S.S., and Poltak S. Enhancement clustering evaluation result of davies-bouldin index with determining initial centroid of k-means algorithm. *Journal of Physics: Conference Series.* 2019; 1235(1). DOI: 10.1088/1742-6596/1235/1/012015.

REFERENCES

1. Singh D.K. et al. Computational Intelligence in Web Mining. *Innovative Trends in Computational Intelligence*. 2022:197–215. DOI: 10.1007/978-3-030-78284-9_9.
2. Sharma S. et al. Performance Evaluation of Secure Web Usage Mining Technique to Predict Consumer Behaviour (SWUM-PCB). *Intelligent Computing and Networking*. 2022;301:136–145. DOI: 10.1007/978-981-16-4863-2_12.
3. Kandpal N., Singh H.P., Shekhawat M.S. Application of web usage mining for administration and improvement of online counseling website. *Int J Appl Eng Res*. 2019;14(7):1431–1437.
4. Kumar V., Thakur R. S. Web log analysis tools: at a glance. *Proceedings of International Conference on Recent Advancement on Computer and Communication*. 2018;34:135–142. DOI:10.1007/978-981-10-8198-9_14.

5. Haridasan A.C., Fernando A.G. Online or in-store: unravelling consumer's channel choice motives. *Journal of Research in Interactive Marketing.* 2018;12(2):215–230. DOI: 10.1108/JRIM-07-2017-0060.
6. Schröer C., Kruse F., Gómez J. M. A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*. 2021;181:526–534. DOI: 10.1016/J.PROCS.2021.01.199.
7. Hypertext Transfer Protocol (HTTP/1.1) RFC7231: Semantics and Content. Internet Engineering Task Force (IETF). Available at: //datatracker.ietf.org/doc/html/rfc7231 (accessed on 15.04.2022).
8. Somyanonthanakul R., Theeramunkong T. Scenario-based Analysis for discovering Relations among Interestingness Measures. *Information Sciences*. 2022;590:346–385. DOI: 10.1016/j.ins.2021.12.121.
9. Shetty P., Singh S. Hierarchical Clustering: A Survey. International Journal of Applied Research. 2021;7(4):178–181. DOI: 10.22271/ALLRESEARCH.2021.V7.I4C.8484.
10. Jarman A.M. *Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method*. Georgia Southern University. 2020. DOI: 10.13140/RG.2.2.11388.90240.
11. Dinh, D.T., Fujinami, T., & Huynh, V.N. Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. *In International Symposium on Knowledge and Systems Sciences*.2019:1–17. DOI: 10.1007/978-981-15-1209-4_1.
12. Sitompul, Bernad J.D., Opim S.S., and Poltak S. Enhancement clustering evaluation result of davies-bouldin index with determining initial centroid of k-means algorithm. *Journal of Physics: Conference Series.* 2019; 1235(1). DOI: 10.1088/1742-6596/1235/1/012015.

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Кокорина Анастасия Игоревна, магистрант, департамент анализа данных и машинного обучения, Финансовый университет при Правительстве Российской Федерации (Финансовый университет), факультет информационных технологий и анализа больших данных, Москва, Российская Федерация.
e-mail: anas.kokorina2017@yandex.ru

Anastasiia I. Kokorina, Undergraduate Student, Data Analysis and Machine Learning Department, The Financial University under the Government of the Russian Federation (FinU or Financial University), Faculty of Information Technology and Big Data Analysis, Moscow, Russian Federation.

Петросов Давид Арегович, кандидат технических наук, доцент, департамент анализа данных и машинного обучения, Финансовый университет при Правительстве Российской Федерации (Финансовый университет), факультет информационных технологий и анализа больших данных, г. Москва, Российская Федерация.
e-mail: DAPetrosov@fa.ru
ORCID: 0000-0002-8214-052X

David A. Petrosov, Candidate of Technical Sciences, Associate Professor of Data Analysis and Machine Learning Department of The Financial University under the Government of the Russian Federation (FinU or Financial University), Faculty of Information Technology and Big Data Analysis, Moscow, Russian Federation.

Зеленина Анна Николаевна, кандидат технических наук, доцент, ведущий специалист проектного отдела, Воронежский институт высоких технологий, Воронеж, Российская Федерация.
e-mail: snakeans@gmail.com
ORCID: 0000-0001-9052-7540

Anna N. Zelenina, Candidate of Technical Sciences, Associate Professor, Leading Specialist of the Project Department, Voronezh Institute of High Technologies, Voronezh, Russian Federation.

Статья поступила в редакцию 23.05.2022; одобрена после рецензирования 14.06.2022; принята к публикации 27.06.2022.

The article was submitted 23.05.2022; approved after reviewing 14.06.2022; accepted for publication 27.06.2022.