УДК 004.04
DOI: http://dx.doi.org/10.21686/1818-4243-2020-4-4-12

К.В. Мулюкова, В.М. Курейчик
Инженерно-технологическая академия Южного федерального университета, Таганрог, Россия

Построение ассоциативной классификационной модели данных на основе метода Apriori*

Целью работы *является исследование современных проблем и перспектив решения интеллектуального анализа больших веб-данных в реальном времени, а также возможность практической реализации технологии Web Mining для больших веб-данных на практическом примере.*
Материалы и методы. *Исследование включало в себя обзор библиографических источников по проблемам интеллектуального анализа больших данных.*
Была применена технология Web Mining для ассоциативного анализа больших веб-данных, а также компьютерное моделирование практической задачи анализа транзакции с помощью скриптового языка общего назначения (PHP).
Результаты. *В ходе работы описана специфика технологии Data Mining, а также был проанализирован современный подход к анализу больших веб-данных – Web Mining. Дана краткая классификация решаемым задачам с помощью технологии Web Mining. Обоснована проблема интеллектуального анализа больших веб – данных на скриптовом языке общего назначения (PHP): отсутствие библиотек для интеллектуального анализа данных, затрудненная нормализация данных к виду необходимому для интеллектуального анализ, взаимодействие с системой управления базой данных.*
Так же был реализован пример, показывающий подход к интеллектуальному анализу больших веб-данных. На основе представления о технологии Web Mining и описанных сложностях анализа веб-данных на языке PHP, были предложены приёмы эффективного решения поставленной практической задачи интеллектуального анализа веб-данных на основе транзакций, совершенных в динамическом веб-приложении.
Был разработан модуль ассоциативного анализа транзакций клиентов на языке программирования PHP. Модуль включает в себя класс интеллектуальной обработки данных. Так же разработана структурная схема модуля, архитектура системы. Построенный модуль позволяет решить основную часть проблемы ассоциативного анализа больших веб-данных по технологии Web Mining с целью решения поставленной задачи выявления закономерностей в большом массиве веб-данных. Ассоциативный анализ веб - данных происходит значительно быстрее благодаря сочетанию скриптового языка общего назначения и объектно-ориентированного подхода.
Заключение. *По результатам проведённого исследования, можно утверждать, что современное состояние технологии анализа больших веб-данных позволяет эффективно обрабатывать объекты данных, выявлять закономерности, получать скрытые данные и получать полноценные статистические данные в реальном времени.*
Полученные результаты могут использоваться как в целях первичного изучения технологий анализа больших веб-данных, так и в качестве дополнения к системе управления содержимым для интеллектуального анализа веб-данных. Использование технологии ассоциативного анализа и созданного универсального класса-обработчика делает созданный модуль гибким, а возможность ручной интеграции делает данный модуль универсальным (не зависит от системы управления базой данных). Методы алгоритма работают с выбранными данными. Данный фактор существенно упрощает дальнейшую разработку программного кода.

Ключевые слова: *большие данные, Data Mining, Web Mining, веб-данные, PHP,структура, анализ данных, big date, интеллектуальная обработка данных, ассоциативный анализ, associative analysis.*

Ksenia V. Mulyukova, Victor M. Kureichik
Engineering and Technological Academy of the Southern Federal University, Taganrog, Russia

Building an Associative Classification Data Model Based on the Apriori Method

The purpose of the work *is to explore the current problems and prospects of mining solution, big web data in real time, as well as the possibility of practical implementation of Web Mining technology for big web data on a practical example.*
Materials and methods. *The study included a review of bibliographic sources on big data mining.*
We used Web Mining technology for associative analysis of large web data, as well as computer modeling of the practical task of transaction analysis using a general-purpose scripting language (PHP).
Results. *During the work, the specifics of the Data Mining technology were described, and a modern approach to the analysis of large web data –Web Mining was analyzed. A brief classification of tasks solved using Web Mining technology is given. The problem of data mining of large web data in a general-purpose scripting language (PHP) has been solved: the lack of libraries for data mining, the difficult normalization of data to the form necessary for associative analysis, interaction with the database management system.*
Also, an example showing an approach to the mining of large web data was implemented. Based on the understanding of Web Mining technology and the described difficulties of analyzing web data in the PHP language, methods for effectively solving the practical problem of analyzing web data based on transactions committed in a dynamic web application have been proposed.
A module for associative analysis of customer transactions in the programming language PHP was developed. The module includes an intelligent data processing class. The structural scheme of the module and system architecture were developed.
The constructed module allows us to solve the main part of the problem of associative analysis of large web data using Web Mining technology in order to solve the problem of identifying patterns in a large array of web data. Associative analysis of web data is much faster because of the combination of a general-purpose scripting language and an object-oriented approach.

* Работа выполнена за счет частичного финансирования по гранту РФФИ ГР №18-07-00050

Conclusion. *According to the results of the study, it can be argued that the current state of the technology for the analysis of large web data allows efficiently process data objects, identify patterns, obtain hidden data and receive complete statistical data in real time.*
The results can be used both for the purpose of the initial research of technologies for analyzing large web data, and as an addition to the content management system for the intelligent analysis of web data. The usage of the technology of associative analysis and the created universal handler class makes the created module flexible, while the possibility of manual integration makes this module universal. With manual integration, the database management system is not important. Algorithm methods work with selected data. This factor greatly simplifies the further development of program code.

Keywords: *Data Mining, web data, Web Mining, data analysis, big date, associative analysis, data analysis.*

Введение

Мы живем в такое время, когда Всемирная паутина становится неотъемлемой часть нашей жизни. Нет такого вида деятельности, в котором бы не использовался Интернет. Работая, общаясь, делая покупки, человек использует Всемирную сеть, поэтому развитие современного общества невозможно без современных технологий.

Современный веб-контент растет, данных становится больше, появляется необходимость держать их под контролем и периодически анализировать. В связи с этим наиболее перспективным направлением для анализа больших веб-данных в динамических веб-приложениях в настоящее время является интеллектуальный анализ веб-данных, также известный как Web Mining. Web Mining является частью Data Mining.

Формирование Data Mining, как направления, которое включает в себя алгоритмы и методы для обнаружения в данных ранее неизвестных, нетривиальных практически полезных и доступных интерпретации знаний необходимых для принятия решений в различных сферах человеческой деятельности, началось в конце двадцатого века, в 90х годах. Теоретиком и основоположником концепции стал Григорий Пятецкий-Шапиро. В статье «Knowledge Discovery in Databases: An Overview» он ввел термин Data Mining [1]. Так же значительный вклад в концепцию Data Mining был внесен Клиффордом Линчем. Он ввел понятие «большие данные», которое тесно связано с концепцией Data Mining. Марц Натан и Уоррен Джеймс в своей книге дают представление о математических аспектах концепции Data Mining и о способах реализации на практике технологий и алгоритмов Data Mining [2].

В 2005 году Тим О'Рейлли вводит понятие Веб 2.0. Понятие Веб 2.0 кардинально меняет подход к проектированию веб-сайтов. С появлением концепции Веб 2.0. веб-сайты в большей степени перестают быть простыми визитками или витринами организации, учреждений и т.д., так как на сегодняшний день для успешного развития в сети Интернет клиенту необходимо предложить нечто большее [3], чем обычный сайт визитка. Веб-сайт становится динамической системой с постоянно растущей в объеме базой данных, которую не представляется возможным проанализировать обычными инструментами для анализа данных [4]. Для этого и необходимо внедрение технологии Web Mining в работу динамических веб-систем. Технология Web Mining применяет методы Data Mining для анализа неструктурированной, неоднородной, распределенной и значительной по объему информации, содержащейся в сети Интернет. Успешное применение технологий Web Mining дает серьезные конкурентные преимущества при осуществлении цифрового бизнеса.

В настоящее время уже имеется значительный ряд работ, затрагивающих те или иные аспекты применения технологий Web Mining, например, [5–7] и др. Но данные работы носят, в основном, обзорный или теоретический характер. В данных работах рассматривается либо готовый инструментарий для интеллектуального анализа данных [8–9] либо перспективы применения технологий Web Mining к различным вопросам цифрового бизнеса [10].

Актуальность выбранной темы обусловлена тем, что многим современным компаниям и учреждениям необходим интеллектуальный анализ данных для успешного развития и конкурентоспособности. Но процесс внедрения алгоритмов Web Mining в практическую деятельность динамических веб-систем и веб-приложений для проведения интеллектуального анализа веб-данных в большинстве случаев затратный и трудоемкий.

Основной проблемой является внедрение готовой программы для интеллектуального анализа данных в действующие динамические веб-системы. Большинство современных динамических веб-систем построены на скриптовом языке общего назначения PHP, в отличие от систем интеллектуального анализа данных, которые спроектированы на других языках программирования. Внедрение системы интеллектуальной обработки данных в динамическое веб-приложение приводит к системным ошибкам, потере информации, а качество данных для анализа значительно снижается. Из-за интеграции систем данные появляется с задержкой, у специалиста нет доступа к данным в реальном времени.

Задача разработки методов и моделей интеллектуального анализа веб-данных для динамических веб-систем, разработанных на скриптовом языке программирования PHP, предоставляемых за небольшие деньги и обеспечивающих простое и удобное интегрирование в процесс работы компании в реальном времени, которые могут применяться на практике широким кругом лиц, не имеющих специального образования, является актуальной на сегодняшний день.

Целью работы является исследование технологии Web Mining для интеллектуального анализа веб-данных, разработка модели ассоциативного анализа на скриптовом языке общего назначения PHP для больших веб-данных на практическом примере – электронной транзакции в динамическом веб-приложении, а также реализация в форме программного модуля.

Что такое Data Mining и Web Mining?

Исторически сложилось, что у термина Data Mining есть несколько вариантов перевода (и значений):

– извлечение, сбор данных, добыча данных [11] (ещё используют Information Retrieval или IR);

– извлечение знаний, интеллектуальный анализ данных (Knowledge Data Discovery или KDD, Business Intelligence).

IR оперирует первыми двумя уровнями информации, соответственно, KDD работает с третьим уровнем. Если же говорить о способах реализации, то первый вариант относится к прикладной области, где главной целью являются сами данные, второй – к математике и аналитике, где важно получить новое знание из большого объёма уже имеющихся данных. Чаще всего извлечение данных (сбор) является подготовительным этапом для извлечения знаний (анализ) [12].

Обработка больших данных представляет собой комплексную задачу, не имеющую однозначного решения и осложнённую рядом факторов. В случае обработки веб-данных, на эти факторы дополнительно накладываются проблемы канала доступа к информации и вопросы сетевых протоколов. Также существенной проблемой анализа веб-данных является их децентрализованность [13], обычно характеризуемая двумя положениями:

– различные записи находятся на различных источниках в сети Интернет, доступ к которым может быть ограничен по скорости или по числу запросов за период;

– структура данных на каждом источнике отличается.

Таким образом, мы видим, как формируется усложнённая задача обработки больших данных в сети Интернет, которые являются распределёнными и неструктурированными, но над которыми, как правило, необходимо выполнить анализ по строгим правилам расчёта. Такие веб-данные, часто называют «скрытыми», т.к. они содержатся в гигабайтах и терабайтах информации, которые человек не в состоянии исследовать самостоятельно. В связи с этим существует высокая вероятность пропустить гипотезы, которые могут принести значительную выгоду. Очевидно, что для обнаружения скрытых знаний необходимо применять специальные методы автоматического анализа, при помощи которых приходится практически добывать знания из «завалов» информации. За этим направлением прочно закрепился термин «добыча данных» или Data Mining. Классическое определение этого термина дал в 1996 г. один из основателей этого направления – Григорий Пятецкий-Шапиро: «Добыча данных – Data Mining являет собой исследование и обнаружение “машиной” (алгоритмами, средствами искусственного интеллекта) в сырых данных скрытых знаний, которые ранее не были известны, нетривиальны, практически полезны, доступны для интерпретации человеком. Эти знания должны быть новые, ранее неизвестные» [14].

На текущем этапе развития технологий, Data Mining позволяет решать следующие задачи обработки массивов данных:

– статистические цели – получение распределений, поиск корреляций, выявление нечётких закономерностей;

– аналитические цели – прогнозирование, поиск необычных значений, моделирование;

– бизнес-аналитика – построение обобщённых данных по массиву детализированных данных.

Данная область знаний всё ещё не имеет чёткого научного определения, поэтому в общем случае под Data Mining понимают любую обработку достаточно большого массива данных, позволяющую получать при помощи Data Mining новые знания, неочевидные без дополнительной, порой весьма трудоёмкой по сложности и затратной по времени обработки.

Для более глубокой обработки и выявления закономерностей в данных используются задачи: регрессии, ассоциации, классификации, последовательности, кластеризации, анализ отклонений, сокращение описания. Такие методы позволяют выявлять закономерности и обрабатывать большие массивы данных.

Для эффективного решения задач обработки больших веб-данных и преодоления ранее сформулированных сложностей, было разработано много эффективных и высокопроизводительных технологий, находящихся в сфере знаний Data Mining. Одна из них технология [15] – Web Mining. Web Mining, – это добыча данных в веб.

В первом приближении технология делится на три группы, в соответствии с классификациями и решаемыми задачами:

- Анализ использования веб-ресурса.
- Извлечение веб-структур.
- Извлечение веб-контента.

Одной из задач извлечение веб-контента является поиск шаблонов в поведении пользователя. Данная задача позволяет выявить закономерности в шаблонах взаимодействия пользователя с динамической веб-системой с целью прогнозирования его дальнейших действий. Найденные паттерны используются в дальнейшем в бизнес-аналитике, веб-аналитике и т.д.

В Web Mining, так же, как и в Data Mining, для поиска шаблонов в поведении пользователя используют задачи ассоциации, классификации, кластеризации [16], анализ последовательностей. Такие методы позволяют выявлять закономерности и обрабатывать большие массивы данных.

Проблема реализации задачи ассоциации в динамических веб-системах, спроектированных на языке программирования PHP

На сегодняшний день язык PHP занимает лидирующие позиции по количеству спроектированных на нем динамических веб-сайтов в сети Интернет. Согласно статистике, каждый второй сайт спроектирован помощью языка программирования PHP.

Но в отличие от языка программирования Python [17], в PHP почти нет библиотек предназначенных для интеллектуального анализа веб-данных.

В динамических веб-системах или системах управления сайтами, спроектированных на языке программирования PHP, приходится использовать готовые системные решения для интеллектуального анализа веб-данных, например, поиска шаблонов в поведении пользователя [18]. Наибольшую сложность вызывает реализация внедрения готового программного продукта в действующую динамическую веб-систему. Это обусловлено высокой степенью затрат на внедрение интеллектуальной системы в готовое динамическое веб-приложение, большими временными и ресурсными затратами. Зачастую приходится «допиливать» готовую систему под приложение, а это в свою очередь приводит к нестабильной работе, системным сбоям и неточным данным. Качество полученных данных, с которыми приходится работать специалистам, является плохим. Из-за интеграции систем двух и более систем появляются пропуски, выбросы, экстремальные значения, а также непригодная для обработки форма представления.

На примере постановки задачи ассоциации для анализа больших веб-данных, мы опишем решение подобной проблемы и разработаем архитектуру системы, спроектированную на скриптовом языке программирования PHP.

Пример практической реализации алгоритма ассоциативного правил для анализа больших веб-данных на основе транзакций в интернет-магазине

Практическая задача анализа веб-данных – Ассоциативно классификационную модель часто используют для определения часто встречающихся транзакции. Под транзакцией понимается набор товаров, купленных клиентом за один сеанс в сети интернет.

Формулировка: «Пусть $I=\{i_1, i_2, i_3, \ldots i_n\}$ – множество товаров (элементы). Пусть D – множество транзакций, где каждая транзакция T – это набор элементов из I, $T \subseteq I$. Каждая транзакция представляет собой бинарный вектор, где $t[k] = 1$ (если i_k элемент присутствует в транзакции), иначе $t[k] = 0$. Транзакция T содержит X, некоторый набор элементов из I, если $X \subset T$.

Ассоциативное правило определяется, как импликация $X \Rightarrow Y$, где $X \subset I$, $Y \subset I$ и $X \cap Y = \varnothing$. Поддержкой правила $X \Rightarrow Y$ называется величина *support s*, если $s\%$ транзакций из D, содержат $X \cup Y$,

$$supp(X \Rightarrow Y) = supp(X \cup Y)$$

Достоверность правила определяет, какова вероятность того, что из X следует Y. Достоверностью правила $X \Rightarrow Y$ называется величина *confidence c*, если $c\%$ транзакций из D, содержащих X, также содержат Y, по формуле (1):

$$conf\left(X \Rightarrow Y\right) = \frac{\sup p(X \cup Y)}{\sup p(X)} \quad (1)$$

Характеристиками алгоритма являются достоверность и поддержка правила.

Практический пример. Имеется 70% заказов, в которых купили чай, и также купили медовый смузи. В 8% всех заказов имеются и чай, и медовый смузи.

8% – является поддержкой;

70% – “Чай” → “Медовый смузи” с примерной вероятностью 70%. Это является достоверностью правила.

Вывод: цель алгоритма – определение взаимных связей: если в транзакции встретился набор компонентов X, то можно сделать вывод, что в той же транзакции должен появиться набор компонентов Y [19].

У алгоритмов ассоциативных правил имеются границы поддержки и достоверности [20] правила. Зачастую количество правил нужно ограничить установленными порогами (*threshold*): минимальная поддержка и достоверность – *minsupp* и *minconf*.

При высоком значении поддержки в ходе анализа будут находиться очевидные правила. При низком значении в

ходе анализа будет находиться большое количество правил, которые являются не очевидными и необоснованными.

Работа алгоритма происходит в два этапа:

- поиск и формирование набора элементов, удовлетворяющих *minsupp threshold*. Данные наборы элементов называют часто встречающимися;
- создание правил из установленных наборов компонентов с точностью соответствующей *minconf threshold*.

Некоторые виды алгоритма ассоциативных правил вводят дополнительные величины. Одной из таких величин является важность (*importance*) [21]. Вычисляют данную величину по формуле (2):

$$\mathrm{Imp}(X \Rightarrow Y) = \\ = \log\left(\frac{p(X \mid Y)}{p(X \mid notY)}\right) \qquad (2)$$

На практике параметр важности используется больше, чем параметр достоверности. Параметр важности помогает отследить изменения вероятности появления товара Y при купленном товаре X.

Если X приобретен, то продукт Y имеет наибольшую вероятность быть приобретенным при условии, что параметр важности положителен. Если параметр важности отрицателен, то Y имеет наименьшую вероятность быть приобретенным вместе с X. Если же параметр важности равен нулю, то продукты друг с другом не связаны.

Для применения алгоритма ассоциативного правила к задачам анализа транзакций покупателя, необходимо выполнить следующие подготовительные шаги.

Для наглядности разработаем структурную схему, в которой укажем данные шаги рис. 1.

Для наглядности приведем на рис. 2 обобщенную архитектуру приложения для решения указанной задачи.

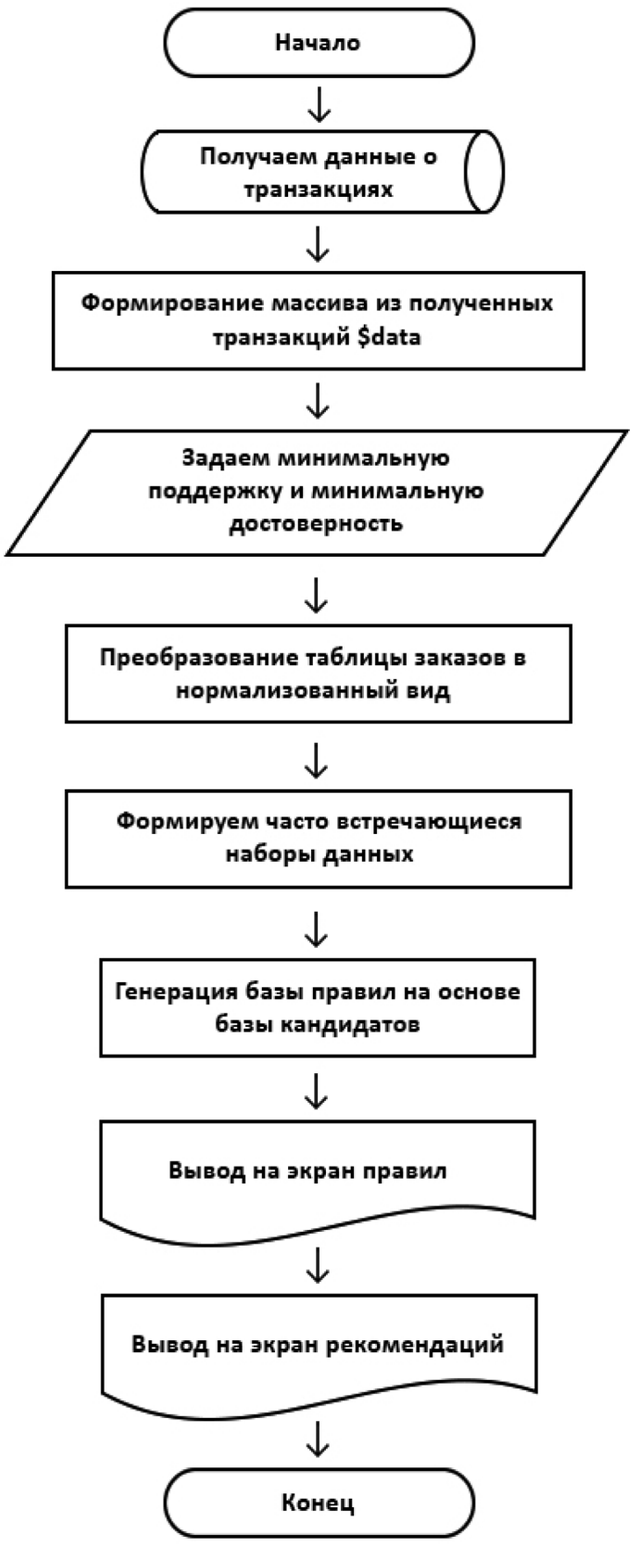

Рис. 1. Структурная схема алгоритма

Хотя данная статья не предполагает полноценной разработки программного решения, мы опишем алгоритм на языке PHP в виде класса [22]. Это позволит более конкретно представить технологию решения задачи с использованием алгоритма ассоциативных правил в области больших веб-данных.

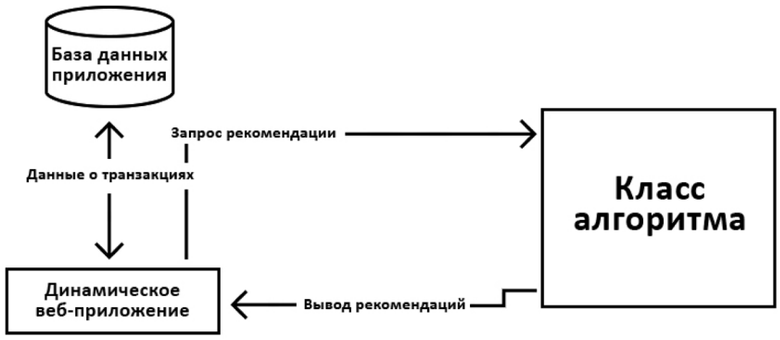

Рис. 2. Архитектура приложения

Проектирование модуля, решающего задачу анализа больших веб-данных на основе транзакций в динамическом веб-приложении

Модуль разработан на скриптовом языке общего назначения и реализован в виде PHP класса. Класс содержит следующие основные методы:

- *arrayAdd* – объединить два или более массива путем добавления значений в элементы с одинаковыми ключами. Результатом является массив, который содержит элементы из всех входных массивов с суммированными значениями;
- *toValidKey* – возвращает форматированный ключ ассоциированной таблицы;
- *setTransactionsFromDB* – заполняет данные из входного массива;
- *getTransactions* – возвращает таблицу транзакций;
- *displayTransactions* – вывод на экран таблицы транзакций;
- *aprioriGen* – создает доверительные наборы из часто используемых наборов элементов;
- *solve* – генерация наиболее частых наборов с учетом минимальной поддержки;
- *getFrequentItems* – возвращает частотный массив;
- *addRule* – добавляет новое правило ассоциации в набор;
- *generateRules* – генерация ассоциативных правил с учетом уровня минимального доверия;
- *getRules* – возвращает ассоциативные правила;
- *displayRules* – вывод на экран ассоциативных правил;
- *displayRecommendations* – вывод рекомендаций;
- *getMinSupport* – возврат минимальной поддержки;
- *getMinConfidence* – возврат минимального доверия;
- *setMinSupport* – установка минимального уровня поддержки;
- *setMinConfidence* – установка минимального уровня доверия;
- *setMinSupport* – установка минимального уровня поддержки;
- *setMinConfidence* – установка минимального уровня доверия.

Публичные свойства класса приведены на рис. 3.

Последние два метода необходимы для интеграции модуля в любое динамическое веб-приложение.

Среди методов класса основными являются:

setTransactionsFromDB,
setMinSupport,
setMinConfidence,
solve, generateRules,
displayRules,
displayRecommendations.

Метод *setTransactionsFromDB* предназначен для подготовительного этапа работы алгоритма: сбор данных в виде, который необходим для работы и нормализации таблицы данных [23].

Метод *solve* формирует наборы кандидатов и оставляет часто встречающиеся на основе вычисленной поддержки набора.

Метод *generateRules* реализует в себе основную часть работы алгоритма Априори из сформированных наборов, создает правила взаимосвязи между кандидатами и вычисляет достоверность для каждого правила. На основе полученных правил, мы и получаем рекомендации по выбору товара – с помощью метода *displayRecommendatios.*

Построенный модуль-класс позволит провести интеллектуальную обработку по технологии Web Mining больших веб-данных с целью решения поставленной задачи ассоциативного анализа на основе транзакций в динамическом веб-приложении (интернет-магазине). Сочетание скриптового языка общего уровня, объектно-ориентированного подхода и алгоритма ассоциативного анализа позволяет максимально ускорить процесс анализа и получения результатов задачи в реальном времени.

Приведенная модель позволяет выполнять эффективный ассоциативный анализ больших веб-данных, получаемых из транзакций клиентов. Кон-

```
класс Apriory
    __construct(&$dataSrc = null, $minSupp = 5, $minConf = 33)
getTransactions()
displayTransactions()
solve($minSupp = null)
generateRules($minConfidence = null)
getRules()
displayRules()
getRecommendations($set)
displayRecommendations($set)
```

Рис. 3. Класс Априори. Публичные свойства

кретное решение может отличаться в деталях и зависеть от используемой системы управления динамическим веб-приложением. В общем случае, при реализации данного модуля ассоциативного анализа веб-данных на уровне программного кода, необходимо дополнительно передать массив данных, который включает список транзакций со списком товаров в следующем формате:

array(*Порядковый_номер* ⇒ ⇒ *array*(*ИД_транзакции, Элемент*), *Порядковый_номер* ⇒ ⇒ *array*(*ИД_транзакции, Элемент*), *array*(*ИД_транзакции, Элемент*)...).

Заключение

Технология Data Mining всё еще является достаточно сложной с практической точки зрения и требующей дополнительного изучения в теории.

В процессе исследования:

1) Сформулированы основные направления обработки больших веб-данных. Описан современный подход к анализу веб-данных – технология Web Mining.

2) Описана специфика ассоциативной классификационной модели.

3) Предложена экспериментальная задача по интеллектуальному анализу веб-данных, основанная на обработке и анализе транзакций в динамическом веб-приложении.

4) Разработан модуль на скриптовом языке общего назначения ассоциативной классификационной модели для анализа больших веб-данных на основе анализа транзакций в динамическом веб-приложении.

Полученные результаты могут использоваться как в целях первичного изучения технологии Web Mining, так и в качестве основы разработки уже реального приложения для ассоциативного анализа веб-данных. Предложен усовершенствованный алгоритм ассоциативных правил, который разработан на скриптовом языке общего назначения. Разработка алгоритма на скриптовом языке общего назначения значительно облегчит внедрение данного модуля в существующие системы управления сайтами или системы, которые разработаны на собственном движке управления.

Результаты вычислительного эксперимента по внедрению данного алгоритма в систему управления сайтом показали, что упростилась интеграция разработанного модуля с динамическим веб-приложением, сократились расходы и время на анализ веб-данных на основе транзакций в интернет-магазине. Данные анализа можно получать в реальном времени.

Созданный модуль является гибким и универсальным для внедрения в динамическое веб-приложение.

Литература

1. Акимушкин В.А., Поздняков С.Н. Обзор методов educational data mining для анализа протоколов взаимодействия обучаемого с «научными играми» // Компьютерные инструменты в образовании. 2013. № 6. С. 26–32.

2. Марц Н., Уоррен Д. Большие данные. Принципы и практика построения масштабируемых систем обработки данных в реальном времени. М.: Вильямс, 2017. 368 с.

3. Кошик А. Веб-аналитика 2.0 на практике. Тонкости и лучшие методики. М.: Вильямс, 2014. 528 с.

4. Novikova G.M., Azofeifa E.J. Semantics of big data in corporate management systems // Discrete and Continuous Models and Applied Computational Science. 2018. № 4 (26). С. 383–392.

5. Паклин Н., Орешков В. Бизнес-аналитика: от данных к знаниям. СПб.: Питер, 2013. 704 с.

6. Благирев А. П., Хапаева Н. Big Data простым языком. М.: АСТ, 2019. 256 с.

7. Кычкин А.В., Квитко Я.И. Архитектурно-функциональная организация информационной системы управления большими данными в промышленности и энергетике // Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. 2018. № 25. С. 109–125

8. Касторнова В.А. Технология использования программных сред информационно образовательного пространства предметной области «Информатика» в осуществлении контроля знаний // Управление образованием: теория и практика. 2018. № 3 (31). С. 33–49.

9. Филяк П.Ю., Байларли Э.Э.О., Растворов В.В., Старченко В.И. Инструментальные средства для использования Big Data и Data Mining в целях обеспечения информационной безопасности – подходы, опыт применения // Вестник Московского финансово-юридического университета. 2017. № 2. С. 210–220

10. Павлов Н. В. Советующая интеллектуальная система как инструмент решения маркетинговых проблем и обучения маркетологов-практиков // Практический маркетинг. 2018. № 3 (253). С. 3–9.

11. Большие Данные [Электрон. ресурс] // Толковый словарь на Академике. Режим доступа: https://dic.academic.ru/dic.nsf/ruwiki/1422719 (Дата обращения: 16.06.2020).

12. Data Mining: что внутри [Электрон. ресурс] // Habr. Режим доступа: https://habr.com/ru/post/95209/ (Дата обращения: 24.06.2020).

13. Мулюкова К.В., Курейчик В.М. Проблема анализа больших веб-данных и использование технологии Data Mining для обработки и поиска закономерностей в большом массиве

веб-данных на практическом примере // Открытое образование. 2019. № 23 (2). С. 42–49.

14. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных. Data Mining, Visual Mining, Text Mining, OLAP. 2 изд. Спб.: БХВ-Петербург, 2007. 384 с.

15. Суркова А.С., Буденков С.С. Построение модели и алгоритма кластеризации в интеллектуальном анализе данных // Вестник Нижегородского университета им. Н.И. Лобачевского. 2012. № 2 (1). С. 198–202.

16. Григораш А.С., Курейчик В.М., Курейчик В.В. Программный комплекс решения задачи кластеризации // Программные продукты и системы. 2017. № 2(30). С. 261–269.

17. Валитова Ю. О., Фазанова А. Д. Алгоритм автоматизированного сбора и анализа данных для формирования модели личности специалиста, востребованного рынком труда // Вестник евразийской науки. 2017. № 2 (9). С. 1–9.

18. Сытник А.А., Шульга Т.Э., Данилов Н.А., Гвоздюк И.В. Математическая модель активности пользователей программного обеспечения // Программные продукты и системы. 2018. № 1(31). С. 79-84.

19. Пивоварова Н.В., Видунова С.И. Интеллектуальный анализ данных в фармацевтическом бизнесе // Вестник евразийской науки. 2016. № 6 (8). С. 1–8.

20. Биллиг В.А., Иванова О.В., Царегородцев Н.А. Построение ассоциативных правил в задаче медицинской диагностики // Программные продукты и системы. 2016. № 2 (114). С. 146–157.

21. Оляныч И.А. Сравнение алгоритмов построения ассоциативных правил на основе набора данных покупательских транзакций // Известия Самарского научного центра Российской академии наук. 2018. № 6-2 (20). С. 379–382.

22. Свиридов А.С., Лазарев В.С. Разработка базовой абстракции действий по выполнению математических операций на языке программирования PHP // Известия Южного федерального университета. Технические науки. 2015. № 4 (165). С. 217–224.

23. Лагерев Д.Г., Савостин И.А., Герасимчук В.Ю., Полякова М.С. Исследование склонности пользователей интернет-магазина к покупке на основе технических данных о визитах посетителей интернет-магазина // Современные информационные технологии и ИТ-образование. 2018. № 4 (14). С. 911–922.

References

1. Akimushkin V.A., Pozdnyakov S.N. Review of educational data mining methods for analyzing the protocols of student interaction with «scientific games». Komp'yuternyye instrumenty v obrazovanii = Computer tools in education. 2013; 6: 26-32. (In Russ.)

2. Marts N., Uorren D. Bol'shiye dannyye. Printsipy i praktika postroyeniya masshtabiruyemykh sistem obrabotki dannykh v real'nom vremeni = Big data. Principles and practice of building scalable real-time data processing systems. Moscow: Williams; 2017. 368 p. (In Russ.)

3. Koshik A. Veb-analitika 2.0 na praktike. Tonkosti i luchshiye metodiki = Web analytics 2.0 in practice. Subtleties and best practices. Moscow: Williams; 2014. 528 p. (In Russ.)

4. Novikova G.M., Azofeifa E.J. Semantics of big data in corporate management systems. Discrete and Continuous Models and Applied Computational Science. 2018; 4(26): 383 - 392.

5. Paklin H., Oreshkov V. Biznes-analitika: ot dannykh k znaniyam = Business analytics: from data to knowledge. Saint Petersburg: Peter; 2013. 704 p. (In Russ.)

6. Blagirev A. P., Khapayeva N. Big Data prostym yazykom = Big Data in simple language. Moscow: AST; 2019. 256 p. (In Russ.)

7. Kychkin A.V., Kvitko YA.I. Architectural and functional organization of the information system for managing big data in industry and energy. Vestnik Permskogo natsional'nogo issledovatel'skogo politekhnicheskogo universiteta. Elektrotekhnika, informatsionnyye tekhnologii, sistemy upravleniya = Bulletin of the Perm National Research Polytechnic University. Electrical engineering, information technology, control systems. 2018; 25: 109-125 (In Russ.)

8. Kastornova V.A. The technology of using software environments of the educational information space of the subject area «Informatics» in the implementation of knowledge control. Upravleniye obrazovaniyem: teoriya i praktika = Education management: theory and practice. 2018; 3(31): 33-49. (In Russ.)

9. Filyak P.YU., Baylarli E.E.O., Rastvorov V.V., Starchenko V.I. Tools for using Big Data and Data Mining in order to ensure information security - approaches, application experience. Vestnik Moskovskogo finansovo-yuridicheskogo universiteta = Bulletin of the Moscow University of Finance and Law. 2017; 2: 210-220 (In Russ.)

10. Pavlov N.V. The advising intellectual system as a tool for solving marketing problems and training marketing practitioners. Prakticheskiy marketing = Practical marketing. 2018; 3(253): 3-9. (In Russ.)

11. Bol'shiye Dannyye = Big Data [Internet]. Explanatory Dictionary on Academician. Available from: https://dic.academic.ru/dic.nsf/ruwiki/1422719 (cited 16.06.2020). (In Russ.)

12. Data Mining: chto vnutri = Data Mining: What's Inside [Internet]. Habr. Available from: https://habr.com/ru/post/95209/ (cited 24.06.2020). (In Russ.)

13. Mulyukova K.V., Kureychik V.M. The problem of analyzing big web data and the use of Data Mining technology for processing and searching for patterns in a large array of web data on a practical example. Otkrytoye obrazovaniye = Open Education. 2019; 23(2): 42-49. (In Russ.)

14. Barsegyan A.A., Kupriyanov M.S., Stepanenko V.V., Kholod I.I. Tekhnologii analiza dannykh. Data Mining, Visual Mining, Text Mining, OLAP. 2 izd = Data analysis technologies. Data Mining, Visual Mining, Text Mining, OLAP. 2nd ed. Saint Petersburg: BHV-Petersburg; 2007. 384 p. (In Russ.)

15. Surkova A.S., Budenkov S.S. Building a model and a clustering algorithm in data mining. Vestnik Nizhegorodskogo universiteta im. N.I. Lobachevskogo = Bulletin of Nizhny Novgorod University. N.I. Lobachevsky. 2012; 2(1): 198-202. (In Russ.)

16. Grigorash A.S., Kureychik V.M., Kureychik V.V. Software complex for solving the clustering problem. Programmnyye produkty i sistemy = Software products and systems. 2017; 2(30): 261-269. (In Russ.)

17. Valitova YU.O., Fazanova A.D. Algorithm of automated data collection and analysis for the formation of a personality model of a specialist demanded by the labor market. Vestnik yevraziyskoy nauki = Bulletin of Eurasian Science. 2017; 2(9): 1-9. (In Russ.)

18. Sytnik A.A., Shul'ga T.E., Danilov N.A., Gvozdyuk I.V. Mathematical model of software users' activity. Programmnyye produkty i sistemy = Software products and systems. 2018; 1(31): 79-84. (In Russ.)

19. Pivovarova N.V., Vidunova S.I. Data mining in pharmaceutical business. Vestnik yevraziyskoy nauki = Bulletin of Eurasian Science. 2016; 6(8): 1-8. (In Russ.)

20. Billig V.A., Ivanova O.V., Tsaregorodtsev N.A. Construction of associative rules in the problem of medical diagnostics. Programmnyye produkty i sistemy = Software products and systems. 2016; 2(114): 146 -157. (In Russ.)

21. Olyanich I. A. Comparison of algorithms for constructing associative rules based on a set of data of consumer transactions. Izvestiya Samarskogo nauchnogo tsentra Rossiyskoy akademii nauk = Bulletin of the Samara Scientific Center of the Russian Academy of Sciences. 2018; 6-2(20): 379 - 382. (In Russ.)

22. Sviridov A.S., Lazarev V.S. Development of a basic abstraction of actions to perform mathematical operations in the PHP programming language. Izvestiya Yuzhnogo federal'nogo universiteta. Tekhnicheskiye nauki = News of the Southern Federal University. Technical science. 2015; (165): 217 – 224. (In Russ.)

23. Lagerev D.G., Savostin I.A., Gerasimchuk V.U., Polyakova M.S. Research of the propensity of users of an online store to purchase based on technical data on visits of visitors to an online store. Sovremennyye informatsionnyye tekhnologii i IT-obrazovaniye = Modern information technologies and IT -education. 2018; 4 (14): 911-922. (In Russ.)

Сведения об авторах

Ксения Валериановна Мулюкова
Аспирант,
Инженерно-технологической академии Южного федерального университета
Таганрог, Россия
Эл. почта: mu.ksusha@yandex.ru

Виктор Михайлович Курейчик
Д.т.н., профессор
Инженерно-технологическая академия Южного федерального университета
Таганрог, Россия
Эл. почта: vmkureychik@sfedu.ru

Information about the authors

Ksenia V. Mulyukova
Postgraduate student
Engineering and Technological Academy of the Southern Federal University
Taganrog, Russia
E-mail: mu.ksusha@yandex.ru

Victor M. Kureichik
Dr. Sci. (Technical sciences), Professor
Engineering and Technological Academy of the Southern Federal University
Taganrog, Russia
E-mail: vmkureychik@sfedu.ru