ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«БЕЛГОРОДСКИЙ ГОСУДАРСТВЕННЫЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ»
(Н И У « Б е л Г У »)

ИНСТИТУТ ИНЖЕНЕРНЫХ ТЕХНОЛОГИЙ И ЕСТЕСТВЕННЫХ НАУК
КАФЕДРА ПРИКЛАДНОЙ ИНФОРМАТИКИ И ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

РАЗРАБОТКА ФОРМАЛЬНО-ЛОГИЧЕСКИХ СРЕДСТВ КЛАСТЕРИЗАЦИИ ПОЛЬЗОВАТЕЛЕЙ ДЛЯ АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ УПРАВЛЕНИЯ КОНТЕНТОМ WEB-РЕСУРСА

Выпускная квалификационная работа
обучающегося по направлению подготовки 38.04.05 Бизнес-информатика
очной формы обучения, группы 07001634
Тюха Анастасии Сергеевны

Научный руководитель
доцент кафедры прикладной информатики и информационных технологий,
к.т.н.
Асадуллаев Р.Г.

Рецензент
доцент кафедры информационных и робототехнических систем,
к.т.н.
Шамраев А.А.

БЕЛГОРОД 2018

АННОТАЦИЯ

к магистерской диссертации Тюха Анастасии Сергеевны
«Разработка формально-логических средств кластеризации пользователей для автоматизированной системы управления контентом web-ресурса», представленной к защите по направлению подготовки 38.04.05 «Бизнес-информатика», программа «Управление жизненным циклом информационных систем».

С ростом конкуренции на рынке товаров и услуг становится актуальным вопрос о совершенствовании методов продвижения продукта среди целевой аудитории, а также разработке новых методов представления информации о нем в степени максимально соответствующей ожиданиям и потребностям клиентов в условиях электронной коммерции. Из этого вытекает необходимость разработки подходов дифференцированного формирования контента с учетом специфики пользовательской аудитории и предлагаемых товаров и услуг.

Во введении обоснована актуальность, определены объект, предмет, цель исследования, сформулированы задачи, которые необходимо решить для достижения поставленной цели, приведены используемые методы научного исследования, изложены научная новизна, практическая значимость, определены положения, выносимые на защиту, описаны структура и объем диссертации.

В первой главе проведено исследование систем управления контентом сайта, рассмотрены подходы персонализации web-ресурсов, а также исследованы алгоритмы кластеризации многомерных данных.

Во второй главе разработан подход персонализированного управления сайтом, проведена систематизация параметров сегментации пользователей web-ресурса, разработана структура модифицированной системы управления контентом сайта с функцией персонализации.

В третьей главе разработан алгоритм гиперсегментации пользовательских профилей на основе комбинации модифицированного алгоритма DBSCAN и оригинальных процедур, а также разработан алгоритм подбора персонализированного контента.

В заключении представлены основные выводы и результаты работы.

Автор магистерской диссертации  Тюха А.С.

Научный руководитель  Асадуллаев Р.Г.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

В настоящее время каждая компания, предлагающая к продаже какие-либо товары или услуги, активно использует методы представления и продвижения своего продукта в сети Интернет как минимум при помощи размещения необходимой информации на страницах корпоративного сайта и разработки системы рекламных мероприятий. Самым распространенным способом осуществления продаж в сети является интернет – магазин, содержащий полный категорийный каталог товаров, их развернутое описание, а также формы заказа. Однако любой интернет – продавец может столкнуться с проблемой низкой конверсии сайта ввиду целого ряда причин.

С ростом конкуренции на рынке товаров и услуг становится актуальным вопрос о совершенствовании методов продвижения продукта среди целевой аудитории, а также разработке новых методов представления информации о нем в степени максимально соответствующей ожиданиям и потребностям клиентов. В связи с этим наряду с наиболее популярными инструментами персонализации (e-mail рассылки с персональными предложениями и скидками, показ объявлений контекстной рекламы, отображение баннеров с супер-предложениями и др.) наблюдаются тенденции развития новых подходов, например, персонализации контента сайта путем изменения его дизайна, структуры, способов отображения различных элементов и информации для определенных групп его посетителей [28]. Решение данной задачи влечет за собой серьезные денежные затраты на изучение аудитории сайта как потенциальных клиентов, сегментацию пользователей, подготовку соответствующих материалов контента и программную реализацию алгоритмов его отображения. Все обозначенные подзадачи требуют привлечения нескольких сторонних специалистов – профессиональных маркетологов и программистов, что не всегда возможно для компаний малого и среднего масштаба. Таким образом, научные

исследования в области автоматизации процесса персонализированного управления контентом web-ресурса являются актуальными.

Объектом исследования являются современные системы управления контентом сайта.

Предметом исследования являются средства формирования потребительских групп пользователей.

В связи с этим целью выпускной квалификационной работы является повышение эффективности процесса управления контентом web-ресурса на основе подхода персонализации.

В соответствии с поставленной целью определены следующие задачи исследования:

– исследовать системы управления контентом сайта и методы кластеризации многомерных данных;

– спроектировать формальные средства формирования групп пользователей сайта;

– разработать алгоритм сегментации пользователей web-ресурса.

При выполнении выпускной квалификационной работы были использованы следующие методы исследования: анализ, синтез, сравнение, классификация, формализация и моделирование.

Исследованиями в области персонализации интернет-ресурсов занимаются такие ученые как Царев А.Г., Царева Т.Н., Домрачев В.Г. и Ретинская И.В. [58-64]. В научных трудах этих ученых речь идет о моделях и методах персонализации сайта, исследованиях и сборе пользовательских данных. Методы кластеризации рассматриваются в научных трудах Климовой А.С., Берикова В.С., Нейского И.М., Буховец А.Г. [19-23, 29-35].

Научную новизна исследования представляют:

– алгоритм кластеризации пользовательских профилей сайта, который не требует определения желаемого количества кластеров и позволяет учитывать изменения в предпочтениях пользователей сайта;

– алгоритм подбора персонализированного контента структурных элементов страницы web-ресурса в зависимости от того, к какому кластеру пользователь может быть отнесен.

Разработанные формально-логическое средства могут иметь непосредственное применение в практической деятельности в области автоматизации процесса персонализированного управления web-ресурсом.

Основным положения, выносимые на защиту выпускной квалификационной работы:

– структура модифицированной CMS;

– алгоритм кластеризации пользовательских профилей сайта;

– алгоритм подбора персонализированного контента.

Основные положения выпускной квалификационной работы опубликованы в 5 научных работах автора.

Выпускная квалификационная работа состоит из введения, трех глав, заключения, списка использованных источников из 64 наименования, 16 из которых зарубежные. Основная часть работы изложена на 76 страницах машинописного текста и содержит 18 рисунков и 1 таблицу.

1 ИССЛЕДОВАНИЕ СИСТЕМ УПРАВЛЕНИЯ КОНТЕНТОМ И МЕТОДОВ КЛАСТЕРИЗАЦИИ МНОГОМЕРНЫХ ДАННЫХ

1.1. Обзор систем управления контентом

Под CMS (Content management system) или системой управления контентом понимается программная основа для разработки и редактирования сайта [17]. Современные системы управления контентом подразделяются на коммерческие, системы с открытым исходным кодом и узкоспециализированные.

К системам первого типа относятся те, которые созданы коммерческими организациями и направленное действие которых в том, чтобы получить максимально большее количество прибыли с продажи лицензии на свой продукт. Их нельзя изменять или модифицировать под свои характеристики. Они предназначены только для создания сайта со всеми необходимыми настройками. Наиболее популярные из них, согласно рейтингу компании “iTrack”, это 1С –Битрикс, WebAsyst Shop-Script, UMI.CMS, NetCat, HostCMS, InSales, Simpla, diafan.CMS (рисунок 1.1) [42].

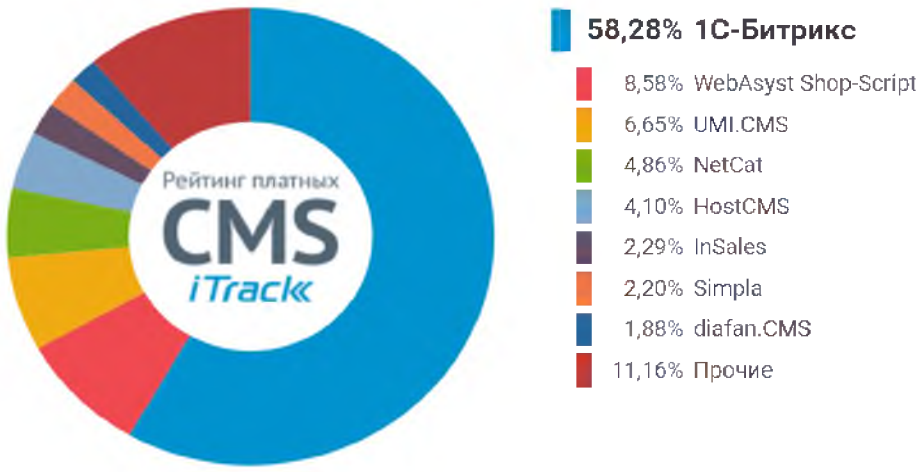

Рисунок 1.1 – Рейтинг платных систем управления контентом

Некоммерческие платформы, наоборот, находятся в бесплатной реализации и распространяются с лицензией Open-Sourse CMS. Главное их отличие в том, что они имеют открытый исходный код, что дает возможность другим разработчикам и пользователям модифицировать эти продукты. Среди систем такого типа самыми популярными являются WordPress, Joomla, OpenCart, Drupal, MODX Revolution, Wix, DataLife Engine, uCoz (рисунок 1.2).

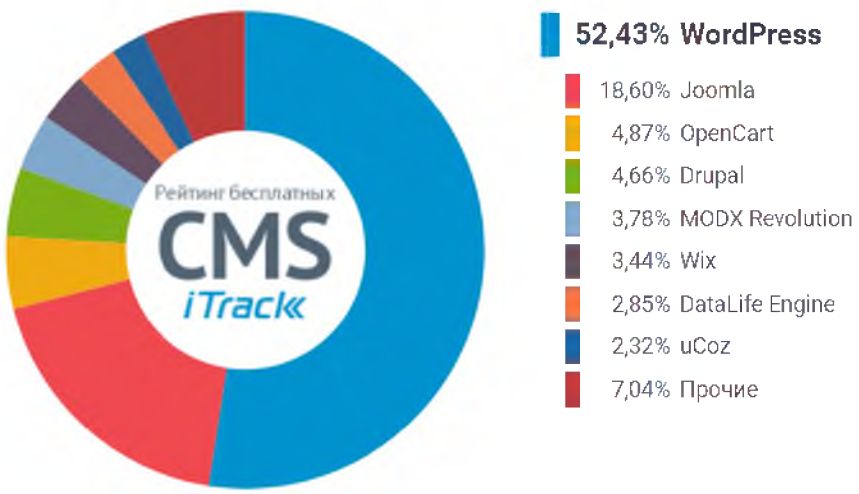

Рисунок 1.2 – Рейтинг бесплатных систем управления контентом

К узкоспециализированным CMS относятся системы, которые используются для создания и сопровождения конкретных видов сайтов, например, интернет-магазинов. К таким системам относятся CMS Sitebill, InSales, ShopCMS, Fast-Sales, Tiu.ru, Melbis Shop, OpenCart, PrestaShop, StoreLand, Zen Cart и другие.

Согласно информации, предоставленной компанией “iTrack”, общий рейтинг для всех видов систем (рисунок 1.3) возглавляет WordPress. Второй по популярности использования является Joomla. Третью позицию рейтинга занимает 1С-Битрикс.

Самая популярная система – WordPress - обладает большим списком преимуществ и положительных сторон [45]. Она бесплатная, имеет множество

плагинов, полностью переведена на русский язык, легка в установке и имеет интуитивно понятную систему управления.

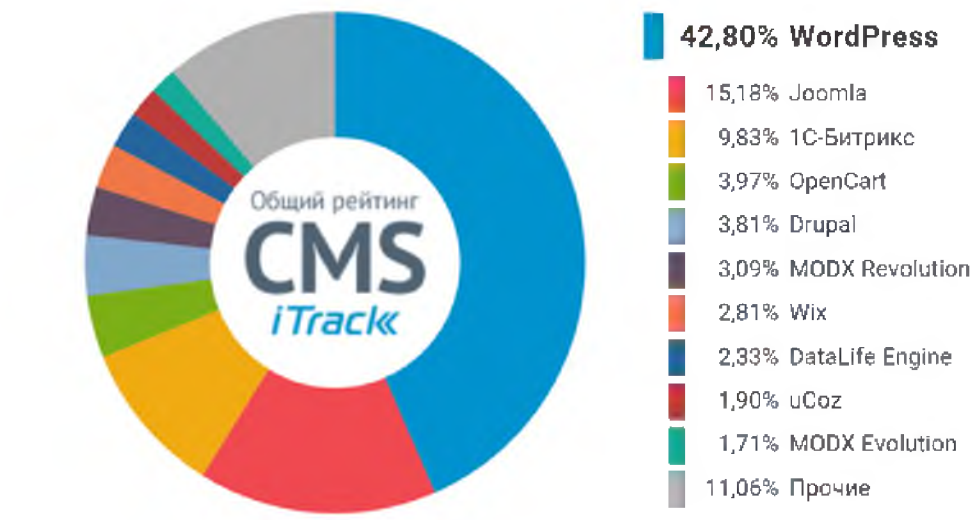

Рисунок 1.3 – Рейтинг систем управления контентом

Среди других преимуществ можно отметить следующие:

– множество бесплатных и платных тем оформления сайта;
– русскоязычный форум и поддержка, огромная база знаний;
– категоризация контента;
– система безопасности;
– предоставления различным пользователям различных прав;
– поддержка тысяч плагинов и расширений, открытый исходный код;
– регулярные обновления с исправлением найденных ошибок;
– модульность;
– отложенная публикация.

WordPress, в основном, применяется для блогов и небольших информационных сайтов, а также, при помощи плагина WooCommerce позволяет создать полноценный интернет-магазин с функциями заказа, корзины, каталога и характеристиками на сайте и простой аналитикой заказов в админ-панели.

Однако у данной системы имеются некоторые недостатки:

– данный движок не предназначен для высоконагруженных проектов или магазинов с многотысячной посещаемостью, так как создает большие нагрузки на сервер;

– из-за популярности системы и большого количества бесплатных плагинов, ее постоянно пытаются взломать, однако, все «лазейки» закрываются в очередном обновлении;

– большинство бесплатных шаблонов для WP имеют огромное количество ошибок как HTML и CSS кода, так и PHP, в идущих к ним виджетах;

– отсутствие личной техподдержки — только через форум;

– структура управления сайтом изначально рассчитана на блог, поэтому коммерческие сайты и интернет магазины, построенные на этой платформе неудобны в управлении таким контентом.

– Joomla - вторая по популярности система управления в мире [44]. Более сложна в установке, управлении и настройке, но вполне доступна для освоения. Также, как и первая CMS, имеет широкий спектр сфер применения, большую базу дополнений, расширений и тем. Данная система имеет следующие возможности:

– увеличение функциональности с помощью дополнительных расширений;

– наличие модуля безопасности, обеспечивающего многоуровневую аутентификацию администраторов и пользователей;

– наличие огромного количества шаблонов позволяет в любое время легко изменять дизайн сайта, расположение элементов страницы, шрифты и пр.;

– схемы расположения разнообразных модулей также разнообразны – возможно указание положения в левом, правом и центральном блоках;

– любые компоненты, шаблоны или модули, размещенные в каталоге расширений, при необходимости могут быть отредактированы или написаны заново;

– регулярно выпускаются новые версии и обновления для них;

– осуществляется поддержка многоязычности;

– поддержка баз данных с каждой новой версией расширяется.

На этой системе отлично расположится сайт-визитка или корпоративный портал, небольшой информационный ресурс или магазин. Среди основных недостатков выделяют недочеты структуры элементов системы управления контентом, уязвимость системы ко взломам, временами наблюдаются проблемы с индексацией, часто код в шаблонах и самой платформе бывает перегружен, обновления системы иногда происходят некорректно, а также система не имеет технической поддержки пользователей.

1С-Битрикс считают по праву одной из самых распространенных коммерческих CMS [37]. Автоматизированная система управления контентом подходит в первую очередь для онлайн-магазинов, а также порталов новостей, информационных и статистических сайтов, социальных сетей и т.п.

Для начала рассмотрим ряд плюсов, которые предоставляет данная CMS:

– безопасность для сайта: разработчики постоянно предлагают усовершенствования в аспекте безопасности коммерческих проектов;

– бесплатная и круглосуточная техническая поддержка, правда, в первый год использования 1С-Битрикс она более оперативна, далее ответа можно ждать и более суток;

– хорошая техническая документация, в которой можно самому найти ответ или решения проблемы;

– возможность интегрирования баз данных в систему;

– возможность совмещать с бухгалтерскими программами, зачастую с 1С;

– стандартный набор действий в администрировании, удобный внутренний текстовый редактор;

– обратная связь: возможность оставлять свои идеи и предложения на сайте для разработчиков данной CMS;

– возможность разрабатывать индивидуальные решения: при определенных знаниях вы можете создать свои систему управления;

– универсальность: разработчики утверждают, что данная CMS подходит для 95% типов сайтов.

Если рассматривать недостатки системы, то большинство разработчиков сходятся во мнении, что Битрикс очень требователен к мощностям хостинг-провайдера и для работы с ним необходимо иметь особые профессиональные навыки. Второй недостаток – это стоимость самой системы Битрикс. На первый взгляд вы столкнетесь с одной стоимостью, но когда вам нужно будет тщательно настраивать модули, то с финансовой стороны это окажется более затратно.

Таким образом, говоря об 1С-Битрикс, можно сделать вывод о том, что этот движок отлично подойдет даже для высоконагруженного интернет-магазина или портала, но ее интерфейс нельзя назвать простым — для того, чтобы освоить все возможности системы, необходимые простому пользователю, придется провести время в справочном центре Битрикс.

Система DLE (DataLife Engine) является платной, но есть возможность использовать ее бесплатно с ограничением в количестве публикаций, достаточно легка в установке, не требовательна к хостингу, создает небольшие нагрузки при большой посещаемости, большое количество платных и бесплатных тем оформления, масштабируется с помощью дополнительных полей. Основное назначение этой системы управления — информационные порталы [39].

Из недостатков можно отметить то, что система будет сложна в освоении для тех, кто впервые начинает администрировать сайт. Казалось бы, такие базовые вещи, как автоматическое или визуальное формирование меню

здесь не предусмотрено — все это придется делать ручками через файлы шаблонов в html.

Всего в DLE есть три типа данных — категории (могут быть вложенными), записи и статические страницы. В базовом функционале нет даже такой опции, как установка превью записи, это необходимо делать с помощью добавления дополнительного поля.

Система Drupal, Есть статистика, что около 20% всех сайтов, созданных в США и Европе, базируется именно на этой CMS, на постсоветском пространстве она не получила такого распространения, но, все же, применяется достаточно широко [41].

На его базе с одинаковым успехом можно сделать и сайт, и интернет магазин, и базу знаний, и социальную сеть, и пространство для совместной работы сотрудников (некий аналог CRM), с помощью тысяч доступных модулей можно расширить функционал сайта до бесконечности. Например, организовать интернет магазин, где клиент сможет оформить заказ и произвести оплату и здесь же, в админ-панели, ваши сотрудники из отдела продаж смогут передать информацию отделу логистики, контролируя статус доставки.

Из минусов можно отметить, что не каждый хостинг может корректно работать с этой системой администрирования — при работе она формирует много запросов к базе данных, что создает некую нагрузку, хотя на рынке хостинг-услуг есть компании, специализирующиеся именно на размещении сайтов на Drupal.

ModX – это достаточно старый движок, который, субъективно, можно оценить как немного устаревший [43]. Система является бесплатной, масштабируемой до бесконечности, проста в освоении (ее структура напоминает простой файловый менеджер — папки, а внутри них записи/страницы), а также безопасной и достаточно часто обновляемой.

Самым большим минусом для новичка становится то, что нет такого понятия, как «Шаблоны для ModX” — внешний вид сайта формируется с помощью CSS и HTML и лишь косвенно связан с сайтом.

Рассмотрев и проанализировав наиболее распространенные системы управления контентом, можно назвать их общие основные функции:

– создание контента сайта (материалов для размещения на его страницах);

– управление контентом сайта (контроль за созданием, изменением и удалением версий документов и файлов, их хранением, осуществление контроля за доступом к файлам, а также осуществление интеграции с иными информационными системами);

– публикация контента и его представление (вывод контента на страницу в соответствии с заданным шаблоном дизайна всего сайта при обращении пользователя к ней).

Использование систем управления контентом имеет ряд преимуществ:

– публикация и обновление информации на сайте, а также определение его визуального представления осуществляется через интерфейс системы, что обеспечивает оперативность модификации его информационного наполнения и не требует от пользователя глубоких знаний языков HTML и php;

– для компаний немаловажным плюсом становится снижение издержек на поддержку сайта, так как не требуется оплачивать услуги внешнего или собственного web-программиста;

– многие системы управления контентом сайта реализуют различные методы интерактивного взаимодействия с пользователями (голосования, форумы и т.д.);

– использование систем управления контентом также позволяет существенно сократить сроки и стоимость разработки сайта за счет своего встроенного функционала;

– гарантия качества разработки сайта при помощи CMS системы обеспечивается неоднократным тестированием и использованием ее различных модулей другими пользователями и программистами;

– разделение данных и их представления в таких системах позволяет оперативно и просто изменять внешний вид сайта или его конкретных элементов.

По характеру представления данных системы управления контентом сайта могут быть объектными, сетевыми или модульными [5, 18].

Главной единицей представления информации в объектных CMS, как ясно из названия, являются объекты: изображения, символы, строки. Объекты объединяются в классы, каждый из которых имеет определённую структуру (иерархическую и сетевую), классы между собой также могут находится в подчинительных и равноправных отношениях. Классы объектов не хранят в себе никакой информации и служат только для облегчения работы с объектами. Объектно-ориентированные CMS наиболее просты в эксплуатации и имеют большой функционал, однако сложные структуры подобных программ отпугивают большинство потенциальных пользователей, предпочитающих более простые системы.

В основе принципа действия сетевых CMS лежит теория графов и причинно-следственные связи между объектами. Основами для данных, помещаемых в сетевую CMS, могут служить как сетевые, так и реляционные СУБД, в которой зафиксированы связи между данными и объектами. Чтобы извлечь информацию из графов, в виде которых представляется информация в сетевых CMS, необходимо использовать рекурсивные алгоритмы: определение списков узлов, идентификация данных узла по данным его «родителя».

В модульных CMS все данные разбиты на определённые разделы (модули), работа с объектами возможна только в пределах его модуля. Типы документов, используемых в модульных CMS, строго фиксированы, а модули полностью независимы друг от друга. Данные системы имеют весьма

ограниченный функционал, однако в силу своей простоты такие системы получили большую популярность главным образом в небольших предприятиях, владеющими небольшими Интернет-сайтами в несколько страниц. Функционал модульных CMS можно серьёзно увеличить, загрузив дополнительные модули из Сети.

Рассматривая процесс функционирования систем управления контентом (рисунок 1.4) важно понимать их главную особенность – визуальный дизайн и информационное наполнение разделены, т.е. разработанные шаблоны страниц хранятся отдельно от данных, а формирование целостного представления любой страницы сайта происходит динамически по запросу пользователя.

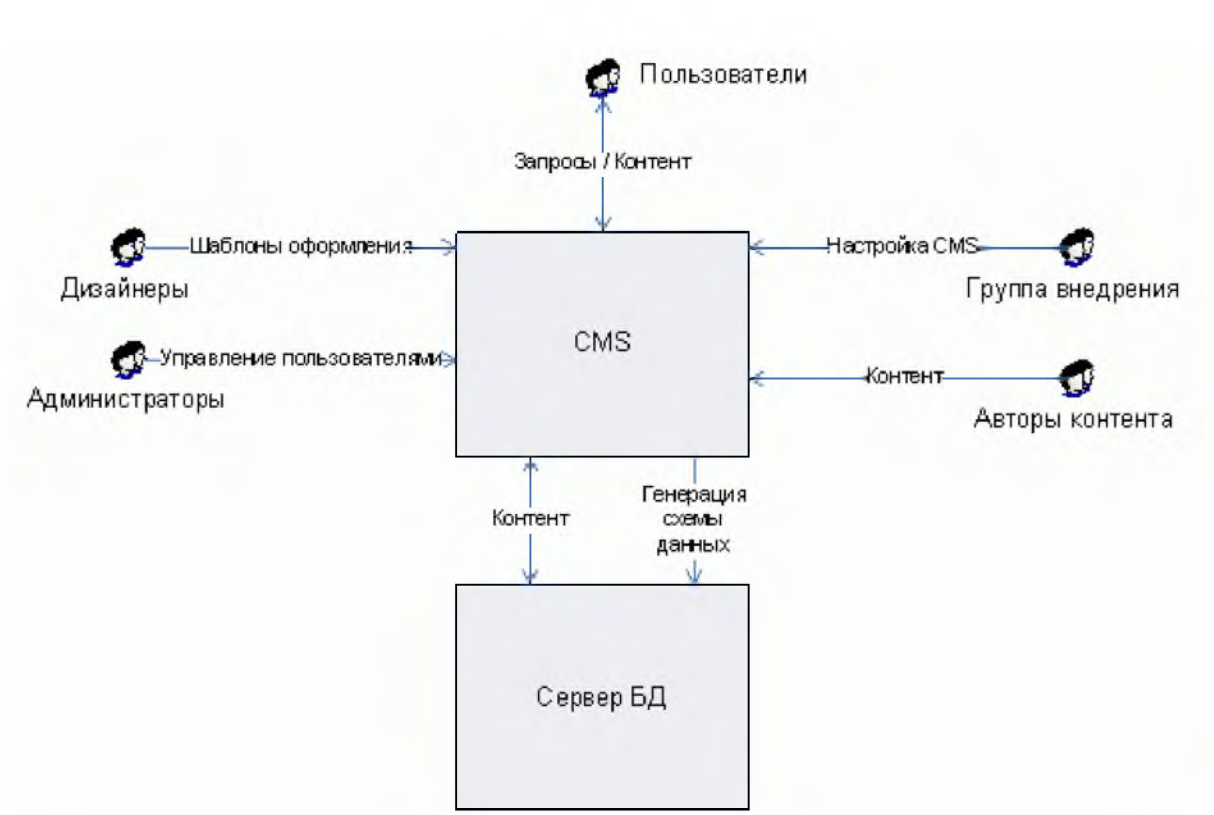

Рисунок 1.4 – Процесс функционирования системы управления контентом сайта

В случае создания сайта при помощи системы управления контентом дизайнеры обеспечивают разработку шаблонов страниц для размещения контента заказчика. В свою очередь разработчик или чаще группа внедрения проводит установку и в случае, если это необходимо, настройку системы в

соответствии с пожеланиями и профессиональной спецификой компании – заказчика, а также создание информационного хранилища данных контента в СУБД. Далее учетные данные администратора передаются клиенту и работа сторонних специалистов в системе прекращается.

Администратор сайта создает, изменяет и удаляет учетные записи отдельных пользователей, назначает им права доступа для работы с различными элементами контента. Наполнение сайта обеспечивают авторы контента. В соответствии с заданными типами контента в системе создаются различные объекты – документы, изображения и другие. В процессе поддержки функционирования сайта авторы также могут изменять или удалять элементы контента. Система также позволяет сохранять информацию о версиях документов, что при необходимости вернуться к любой из них. Более подробно типовая структура системы управления контентом представлена на рисунке 1.5.

Рисунок 1.5 – Типовая структура системы управления контентом

В структуру системы обязательно входят модуль навигации, модуль содержания, модуль контент менеджера, модуль авторизации, файловая система, а также дизайн-шаблон и стили CSS.

Все необходимые для работы системы данные (параметры дизайна, заголовки, метаданные, параметры структуры страницы, настроек модулей, контент и учетные записи пользователей) хранятся в Базе данных. При обращении пользователя к странице сайта соответствующие модули запрашивают необходимые данные и происходит сборка страницы для показа пользователю. Все пользователи сайта в этом случае видят один и тот же контент.

В настоящее время в общем виде архитектура системы управления контентом web – ресурса может быть представления следующим образом (рисунок 1.6).

Рисунок 1.6 – архитектура системы управления контентом сайта

В основу технологии CMS положена архитектура Клиент-сервер, включающая в себя три компонента. В этом случае процесс обработки данных разбит между клиентом, сервером приложений и хранилищем данных. Архитектурой предусмотрено наличие двух хранилищ. Первое, как правило, представляет собой реляционную базу данных и служит для хранения и управления данных для публикации на сайте. Второе – файловая система – содержит различные элементы представления данных (графические изображения, шаблоны и т.д.).

Сервер приложений получает запрос и, чтобы его обработать, связывается с базой данных. В результате клиент получает готовый HTML-

файл. Так сервер приложений служит для динамической доставки контента. Архитектура может предусматривать несколько серверов приложений, связь между которыми происходит через Web-сервер.

Таким образом, проведенный обзор систем управления контентом показал, что для создания и поддержки работы сайтов компании активно используют как коммерческие CMS, так и системы, находящиеся в открытом доступе. Согласно анализу, все они имеют общий функционал - создание контента сайта, управление контентом, его публикация и управление представлением контента. Также были выявлены общие для всех систем управления контентом структурные элементы - модуль навигации, модуль содержания, модуль контент менеджера, модуль авторизации, файловая система, дизайн-шаблон и стили CSS, а также База данных, где хранится необходимая для работы системы информация. В рассмотренных CMS функция персонализации сайта не реализована, либо предлагается доработка/покупка соответствующего модуля системы.

1.2. Исследование подходов реализации функции персонализации

Современные пользователи Интернет предъявляют довольно высокие требования к посещаемым сайтам. Любая система веб-аналитики покажет, что наибольшие конверсии совершаются на узкоспециализированных сайтах с контентом, релевантным пользовательским запросам. Применение технологии персонализации сайта для удовлетворения этих потребностей, является уникальным, эффективным и многообещающим решением [25]. Посетители требуют от современного сайта хороший дизайн, релевантный контент, удобную навигацию. Персонализация сайта предоставляет маркетологам возможность удовлетворить растущие потребности, предоставляя контент и рекламу, которая подстраивается под поведение и характеристики посетителей сайта. Один веб-сайт не может удовлетворить потребности всех пользователей ввиду ряда причин [49]:

1. Намерения посетителей многообразны. Предлагая слишком разнообразный или слишком узкоспециализированный контент, статичные сайты удовлетворяют требованиям одних посетителей, вызывая отчуждение других. Адаптирующийся контент, созданный при помощи персонализации, подстраивается под посетителей, позволяя сайтам эффективно привлекать разнообразных посетителей.

2. Клиенты требуют повышенной релевантности. Сегодняшние потребители проявляют свою осведомленность при каждом взаимодействии с брендом. Они используют интернет для ведения бизнеса и получения информации на каждом этапе потребительского цикла. Клиенты хотят найти то, что ищут быстро и в соответствии со своими нуждами.

3. Фирмы недооценивают возможности конверсий. Не персональный или генерируемый контент прерывает цикл покупки, в связи с невозможностью предоставить правильные предложения и элементы потребительского опыта, которые увеличат продуктивность и предоставят релевантную информацию. Статичным сайтам не хватает подвижности, чтобы предложить похожий персонализированный контент, ограничивая вовлеченность так же, как возможность дополнительных и кросс продаж.

Фирмы инвестируют значительную часть ресурсов в дизайн и привлечение трафика на сайт (SEO, контекстная реклама и др.). Персонализация сайта улучшает процесс превращения посетителей в покупателей и заказчиков (конверсии), т.к. подстраивается под поведенческий опыт и позволяет фирмам сохранять взаимодействие на всех точках пересечения. Чтобы использовать всю ценность приходящего на сайт трафика, фирмы должны использовать проницательность персонализации, как дополнение к кросс-канальной стратегии.

На практике персонализацию сайтов используют для [51]:

1. Предоставления релевантного контента. Персонализация сайта упрощает процесс нахождения контента посетителем. Технология соответствующе определяет ключевые характеристики посетителя и

категоризирует его, основываясь на предустановленных правилах. В то же время посетитель ощущает персонализацию сайта, наслаждаясь преимуществами снижения шума от нерелевантной информации и видит только контент, который его интересует [36].

2. Целевой рекламы. Выстраивание рекламы в зависимости от потребностей клиента, основанной на истории его взаимодействия с сайтом. Во-первых, это повышает удовлетворение клиента, а также приводит к повышению конверсий. В то время как история взаимодействия пользователя с сайтом накапливается, эти данные можно использовать для подготовки уникальных релевантных спец предложений в будущем, а также посетителям со схожими интересами.

Все методы персонализации можно разделить на 2 основные категории [27]:

- основанная на правилах;
- основанная на алгоритмах.

Первый обозначенный метод предполагает сегментно-ориентированный подход принятия решений. На основании различных собранных данных о пользователе (исторических, поведенческих, об окружающей среде) и заранее определенных правил происходит формирование уникального предложения. Примером такого правила служит следующее: «Если пользователь совершает определенное действие, то отображается предложение X».

Второй обозначенный тип персонализации основан на использовании специальных математических систем с целью осуществления мониторинга действий посетителей. На основании полученных сведений осуществляется разработка предикативной модели для определения наиболее релевантного контента для различных пользователей.

Подход, основанный на правилах, наиболее приемлем в ситуациях, где посетители могут поделиться на эксклюзивные, узкоопределенные сегменты, такие как энтузиасты фотографии или молодые мамы. Но таргетинг,

основанный на правилах, может быть нелегким в градации из-за ручного создания настройки правил. Как стратегия, основанная на статистических техниках, персонализация, основанная на алгоритмах, наиболее подходит фирмам с большим трафиком, которые генерируют достаточно данных и конверсий для обучения точных таргетинговых моделей. Персонализация, основанная на алгоритмах, также хорошо подходит в ситуациях, с большим числом перекликающихся предложений. Несмотря на то, что это впечатляющий инструмент, также можно выделить и минусы. Данная техника часто рассматривается как стратегия «черного ящика» и многие компании не могут смириться с мыслью, что какая-то неизвестная математическая формула может персонализировать контент.

Для эффективной работы персонализации устанавливают веб-аналитики (например, Google Analytics или Яндекс.метрику) для анализа результатов. Чтобы персонализация работала верно, оказывала позитивное влияние на конверсии и бизнес, маркетологи должны измерять эффективность таргетинговых попыток в отношении назначенных целей. Кроме того, использовать инструменты таргетинга необходимо на различных уровнях. И основанные на правилах и на алгоритмах стратегии персонализации не обязательно взаимоисключающие. Аналитики должны определить возможности использования таргетинга на сайтах и определить подходящие таргетинговые техники для каждой ситуации. В условиях, где эти две технологии могут действовать независимо и не вызывая конфликтов, аналитики могут использовать различные технологии совместно.

Маркетологи давно определили пользу таргетированных сообщений и отрицательный эффект от массовых рассылок. Большинство из них сходятся во мнении, что наибольший положительный эффект позволяет получить использование тактик персонализации, однако их необходимо грамотно использовать.

Персонализацию уже внедрили крупные технологичные компании. По данным опроса компании CEB (рисунок 1.7), они используют в среднем

5 методик персонализации [38]. Компании с доходом более 1 миллиарда — 6 методик. Для этого чаще всего применяются веб-аналитические программы, CRM и CMS.

Однако даже в компаниях с большим доходом и широкими технологическими возможностями персонализация не функционирует полноценно.

- 78 % опрошенных компанией CEB не хватает инструментов для эффективной реализации персонализации;
- 14 % утверждают, что их персонализация комплексная и совершенная;
- 59 % недовольны действующими методиками персонализации;
- только треть респондентов располагает готовой картой персонализации.

Рисунок 1.7 – Наиболее часто используемые инструменты персонализации

Даже большие компании часто не могут использовать персонализацию эффективно для бизнеса.

Более половины опрошенных CEB тратят от 1 до 10% маркетингового бюджета на инструменты персонализации, но 74% из них не получают существенного возврата вложений.

Около 71 % пользователей выступают за персонализацию, в то время как остальные считают, что она работает плохо.

Самый простой способ персонализации — программа лояльности, но, по мнению клиентов, только в 22% случаев она срабатывает качественно. Другие активности компаний ещё менее эффективны.

Но всё же покупатели хотят персонального обращения. Около 57 % опрошенных согласны с утверждением «Я рассчитываю, что компании, с которыми я работаю, понимают мои нужды и предпочтения».

И чем моложе респонденты, тем больше они ждут от персонализации. Более 70 % молодёжи, которая хорошо знакома с цифровыми технологиями, требуют персонализации на сайтах и других точках коммуникации. Среди старшего поколения процент гораздо ниже: 38 % в возрасте 50–64 лет и 62 % в возрасте более 65 лет считают персонализацию недопустимой. По данным исследования, самый большой раздражитель для клиентов — контент, который им попросту неинтересен.

В ходе исследования методов персонализации интернет-ресурсов было выявлено, что для этого существуют два подхода – основанный на правилах и подход, основанный на алгоритмах. Оба подхода показывают свою эффективность в определенных условиях. Правила лучше работают для узкоспециализированных компаний с небольшим разбросом отличий между клиентами. Алгоритмический же по подход позволяет учитывать большой поток посетителей сайта в сочетании с большим количеством предложений компании. Согласно исследованию компании CEB, крупные компании используют до 12 технологий для персонализации, среди которых

web-аналитика, CRM-система, CMS-система, AB-тестирование и оптимизация компаний и многие другие.

1.3. Анализ методов кластеризации многомерных данных

Одной из самых важных задач анализа данных является кластеризация – объединение объектов из некоторой выборки в сравнительно однородные группы на основании их некоторого сходства [5, 12, 24]. Из чего следует, что главными характеристиками кластеров являются внутренняя однородность и внешняя изолированность. Применение анализа данных распространено крайне широко [15, 50]:

- в маркетинге – для сегментации потребителей, конкурентов и поставщиков;
- в информатике – для упрощения работы с информацией, визуализации данных, сегментации изображений, реализации интеллектуального поиска;
- в экономике – для анализа рынков и финансовых потоков, выявления закономерностей на фондовых биржах;
- в лингвистике – для восстановления эволюционного древа языков;
- в астрономии – для выделения групп звёзд и галактик, автоматической обработки снимков космоса и др.

Ярким отличием кластеризации от классификации является неопределенность классов, их свойств, а иногда и их количества. Все это должно быть определено в процессе работы кластерного анализа [1-3].

Применение подхода кластеризации в Data Minning как начального этапа анализа данных крайне результативно при построении аналитического решения [18,26]. Специалист с большей степенью точности может построить адекватную модель для отдельных групп близких объектов выборки, нежели для всей массы данных сразу.

Общий алгоритм кластеризации включает несколько стадий:

1. Подготовка выборки для проведения анализа.
2. Определение переменных для оценки объектов выборки.
3. Выбор параметров сходства объектов, и вычисление их значений.
4. Этап кластеризации.
5. Обработка и представление результатов кластерного анализа.
6. Интерпретация полученных результатов и их применение.

Для получения оптимального результата, специалист может скорректировать значение выбранной метрики и даже применить несколько разных алгоритмов кластеризации.

Степень сходства между объектами анализа может быть определена путем вычисления значения расстояния между ними с помощью одной из известных метрик.

1. Наиболее известной из них является евклидово расстояние, представляющее собой геометрическое расстояние в многомерном пространстве (1.1):

$$p(x,y) = (\sum_{i=1}^{n}(x_i - y_i)^2)^{1/2} \qquad (1.1)$$

2. В случае, когда необходимо придать больший вес отдаленным друг от друга объектам, часто используется квадрат евклидова расстояния (1.2):

$$p(x,y) = \sum_{i=1}^{n}(x_i - y_i)^2 \qquad (1.2)$$

3. Манхэттенское расстояние (расстояние городских кварталов) представляет собой среднее разностей по координатам (1.3):

$$p(x,y) = \sum_{i=1}^{n}|x_i - y_i| \qquad (1.3)$$

4. В том случае, когда два объекта имеют различные значения одной из координат и нужно определить их как различные, используется расстояние Чебышева (1.4):

$$p(x,y) = max_{i=1..n}|x_i - y_i| \qquad (1.4)$$

5. Степенное расстояние применяется, когда необходимо увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются (1.5):

$$p(x,y) = \sqrt[r]{\sum_{i=1}^{n}|x_i - y_i|^p} \qquad (1.5)$$

В этом случае параметры r и p определяются пользователем. Первый параметр означает постепенное взвешивание разностей по отдельным координатам, второй – за прогрессивное взвешивание больших расстояний между объектами.

Исследователь определяет метрику самостоятельно в зависимости от каждой конкретной задачи кластеризации. Часто бывает, что в процессе исследования могут быть опробованы несколько метрик для получения желаемого результата кластеризации.

Алгоритмы кластеризации принято разделять на иерархические, которые в свою очередь подразделяются на восходящие и нисходящие, и плоские (неиерархические), которые по методу могут быть итеративными, плотностными, модельными, концептуальными и сетевыми. При иерархической кластеризации исследователь получает систему вложенных разбиений или дерево кластеров. Плоские алгоритмы позволяют получить одно разбиение всей исходной выборки объектов на кластеры [6-8, 9, 11].

Согласно другому типу классификации алгоритмы кластеризации могут быть четкими и нечеткими. В первом случае для каждого объекта определен конкретный кластер, а во втором определена степень отношения объекта к тому или иному кластеру [14, 16].

Наиболее распространенными иерархическими алгоритмами являются восходящие [57]. К ним относятся алгоритм ближнего соседа, алгоритмы дальнего и среднего соседа. Каждый объект из выборки помещается в отдельный кластер, затем происходит поэтапное объединение наиболее

близких друг к друг кластеров до момента образования кластерного дерева. Алгоритм можно представить следующим набором шагов:

1. Вычисление расстояния между объектами для формирования матрицы близости.
2. Определение для каждого объекта выборки отдельного кластера.
3. Объединение кластеров, находящихся на минимальном расстоянии друг от друга.
4. Удаление строки и колонок слитых кластеров из матрицы.
5. Возврат к шагу 3 (до момента выполнения критерия остановки).

Результат иерархической кластеризации наглядно отражается при помощи дендрограммы – дерева вложенных кластеров, пример которой приведен на рисунке 1.8 [48].

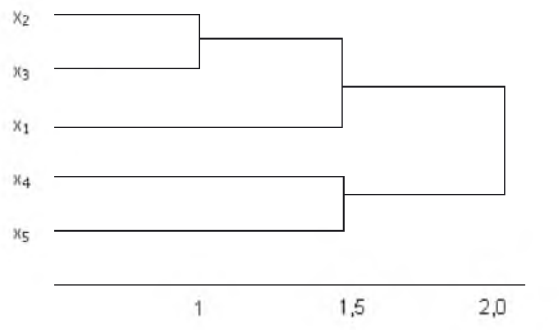

Рисунок 1.8 – Пример дендрограммы – дерева вложенных кластеров

При использовании иерархического алгоритма возможно применение нескольких метрик для вычисления расстояний между кластерами:

1. Расстояние ближайшего соседа (одиночная связь). Расстояние между соседними кластерами определяется самыми близкими друг к другу объектами, отнесенными к разным кластерам.
2. Расстояние дальнего соседа (полная связь). Расстояние между соседними кластерами определяется самыми наиболее удаленно расположенными друг от друга объектами, отнесенными к разным кластерам.

3. Невзвешенное попарное среднее расстояние. Вычисление расстояния между кластерами заключается в расчёте среднего расстояния между всеми парами объектов их разных кластеров.

4. Взвешенное попарное среднее расстояние. Метод вычисления во многом схож с предыдущим, однако в данном случае размер кластеров используется как весовой коэффициент. Метрика используется при работе с кластерами разной размерности.

5. Невзвешенный центроидный метод. Расстояние между кластерами в этом случае определяется как расстояние между их центрами.

6. Взвешенный центроидный метод. Используется, когда кластеры имеют разные размерности. При вычислении расстояния между центрами кластеров учитывается их вес.

Иерархические алгоритмы имеют свои достоинства:

- нет необходимости обучать алгоритм;
- отображают матрицу близости объектов;
- иерархические алгоритмы являются инкрементными.

Среди недостатков выделяют:

- необходимость указывать максимальный размер кластера;
- работа алгоритма не детерминированная, а значит требуется указание определенного порядка для значений близости между кластерами;
- кластеры между собой не пересекаются.

Если говорить о неиерархических методах кластеризации первого типа – итеративных алгоритмах, наиболее известным их них является метод k-средних. Указанный алгоритм ищет заданное количество кластеров исходной выборки, максимально удаленных друг от друга. Работа алгоритма состоит из следующих шагов:

1. Произвольный выбор k точек – центров кластеров.

2. Распределение оставшихся объектов выборки по кластерам с ближайшим центром.

3. В полученных кластерах осуществляется вычисление новых центральных точек (1.6):

$$c_j = \frac{\sum_{i=1}^{L} x_i}{L} \qquad (1.6)$$

где $x_i \in C_j$, $|C_j|$=L.

4. Если условие сходимости не выполнено, возврат к шагу 2.

Критерием остановки часто считают значение среднеквадратической ошибки, которое должно быть минимальным.

Алгоритм k-средних имеет ряд преимуществ:

– он прост в использовании;

– понятен и прозрачен;

– работает быстро на небольших выборках.

К недостаткам можно отнести:

– чувствительность к шуму, который может искажать среднее;

– медленно справляется с анализом больших объемов данных;

– требует указывать количество кластеров, которые необходимо получить.

Разновидностью k-means является метод k-medoids, практической реализацией которого является алгоритм PAM [12]. Медоид – элемент кластера, различие которого с другими объектами кластера минимально. Здесь и кроется главное отличие этих методов – в k-means происходит вычисление центров кластеров, как точек, не обязательно принадлежащих выборке данных, в k-medoids же центр кластера является элементом выборки. PAM также требует изначальное знание k – числа кластеров.

Последовательность действий по выполнению алгоритма следующая:

1. Выбор произвольных k точек в качестве центров кластеров – медоидов.

2. Для каждой точки осуществляется поиск ближайшего центра, с формирование разбиения на кластеры.

Пока медоиды не стабилизируются повторяются следующие шаги:

3. Каждый кластер (C_i) и каждая точка кластера, не являющуюся медоидом (o_i), рассматривается по отдельности.

4. Медоид перебираемого кластера перемещается в текущую точку (o_r) и рассчитывается функция потерь (1.7):

$$S = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, o_i)^2 - d(p, o_r)^2 \qquad (1.7)$$

6. Если S < 0, изменения запоминаются – медоид переносится в новую точку, затем виполняется переход ко второму шагу.

На следующем шаге, можно не перебирать все точки кластеров, а случайным образом выбирать одну из множества.

Следующая группа – это плотностные алгоритмы, определяющие кластер как совокупность кучно расположенных объектов. Другими словами, решение о формировании кластера принимается внутри определенного радиуса находится некоторое минимальное количество объектов. Примером такого алгоритма может выступить DBSCAN [10, 13].

Объекты рассматриваются, как точки в n- мерном пространстве, где n-количество атрибутов. Две точки p и q называются соседними, если расстояние между точкой p и точкой q меньше некоторого ε. Значение ε называют радиусом соседства. Точка q называется главной точкой подкруга, если количество ее соседних точек больше либо равно минимальному количеству соседних точек MinPts.

Точка p называется прямо достижимой по плотности (directly density-reachable) из точки q, если p и q – соседние точки и q – центральная точка подкруга.

Точка p называется достижимой по плотности (density-reachable) из точки q, если существует такая точка m, которая прямо достижима из точки p и прямо достижима из точки q (рисунок 1.9).

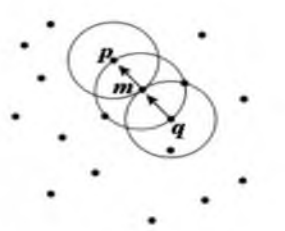

Рисунок 1.9 - Точки p и q достижимы по плотности

Точка p называется соединенной по плотности (density-connected) с точкой q, если существует такая точка o, что p и q достижимы по плотности из точки o (рисунок 1.10).

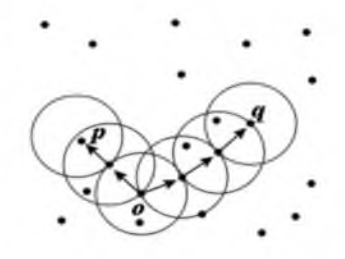

Рисунок 1.10 - Точки p и q соединены по плотности

Алгоритм DBSCAN базируется на следующих принципах:

1. Принцип максимальности: если точки p и q достижимы по плотности, то обе точки принадлежат одному кластеру

2. Принцип связи: каждая точка соединена по плотности со всеми точками, принадлежащими тому же кластеру.

Таким образом, легко идентифицируются выбросы – точки, не достижимые и не соединенные по плотности ни с одной из точек исходного множества.

В общем виде алгоритм представляет собой набор следующих шагов:

1. Выбор параметров: окрестность ε и минимальное количество элементов в кластере Minpts.

2. Проверка окрестности для произвольно выбранного объекта на наличие минимального количества элементов. Если объектов меньше заданного минимума, точка помечается как шум. Переход к другой точке.

3. Если число объектов равно или более минимального, точка помечается как корневая и принадлежащая к кластеру, а окружающие заносятся в отдельную категорию.

4. Каждая точка из категории помечается как принадлежащая к текущему кластеру, и затем, проверяется согласно пункту 2. Если в окрестность попадают новые точки, они также заносятся в категорию.

5. После завершения проверки точка из категории удаляется. В конечном итоге в данной категории точки закончатся, тогда осуществляется переход к шагу 2.

6. Действие алгоритма заканчивается, когда все точки из выборки будут пройдены и помечены.

Среди основных преимуществ алгоритма DBSCAN можно отметить:

- результативен при обработке в меру больших объемов данных;
- применяется, когда количество кластеров заранее не известно;
- позволяет получать кластеры разного размера и формы;
- не является чувствительным к шуму;
- алгоритм хорошо поддается модифицированию.

Главным недостатком DBSCAN — считается неспособность соединять кластеры через проёмы, и, наоборот, способность связывать явно различные кластеры через плотно населённые перемычки.

Статистические алгоритмы кластеризации основаны на предположении, что кластеры описываются некоторым семейством вероятностных распределений. Часто при описании кластерного анализа приводят две гипотезы о байесовском подходе к разделению смеси вероятностных распределений, их упоминание позволит лучше понять идею статистических алгоритмов кластеризации.

Пусть $y \in Y$ – множество кластеров, $x \in X$ – выборка.

Гипотеза о вероятностной природе данных. Объекты выборки появляются случайно и независимо согласно вероятностному распределению, представляющему собой смесь распределений (1.8):

$$P(x) = \sum_{y \in Y} W_y P_y(x) \qquad (1.8)$$

где $P(x)$ – вероятность появление элемента выборки x, $P_y(x)$ – функция плотности распределения кластера y, W_y - неизвестная априорная вероятность появления элементов из кластера y.

Гипотеза о форме кластеров. Объекты описываются n-числовыми признаками $f_1(x), \dots, f_n(x)$, $X = R^n$. Для описания каждого кластера используется функция n-мерной гауссовской плотности $P_y(x) = N(x; \mu_y, \sum_y)$ с центром $\mu_y = (\mu_{y1}, \dots, \mu_{yn})$ и диагональной ковариационной матрицей $\sum_y = diag(\sigma_{y1}^2, \dots, \sigma_{yn}^2)$.

Таким образом, из гипотез следуют следующие краткие выводы: элементы выборки случайны и независимы; каждый кластер описывается центром и ковариационной матрицей.

Первым шагом идёт инициализация начальных кластеров, необходимо знать искомое число кластеров перед выполнением алгоритма. Следует помнить, что n – размерность пространства, l – число элементов выборки.

1. $w_y = \frac{1}{|Y|}$, указание начальных весов кластеров;

2. Назначение центров каждого кластера μ_yкак случайный элемент из выборки;

3. Подсчет элементов ковариационных матриц кластеров (1.9):

$$\sigma_{yj}^2 = \frac{1}{l|Y|} \sum_{i=1}^{l} (f_j(x_i) - \mu_{yj})^2, j=1,\dots,n; \qquad (1.9)$$

в формуле внутри сумматора осуществляется рассчет квадрата разности между j-ым числовым признаком элемента выборки и j-ым числовым признаком центра кластера. Дальнейшее выполнение алгоритма состоит из повторения двух шагов: шага ожидания (Е-шаг) и шага максимизации (М-шага). Их следует повторять, пока характеристики кластеров не перестанут изменяться.

4. На Е-шаге происходит вычисление вероятности принадлежности i-ого элемента выборки к кластеру y (1.10):

$$g_{iy} = \frac{w_y P_y(x_i)}{\sum_{z \in Z} w_z p_z(x_i)}, y \in Y, i = 1, .., l; \quad (1.10)$$

5. На М-шаге пересчитываются веса кластеров – суммированием вероятности каждого элемента выборки принадлежать к заданному кластеру y, делим итоговую сумму на число элементов выборки (1.11):

$$w_y = \frac{1}{l} \sum_{i=1}^{l} g_{iy}, y \in Y; \quad (1.11)$$

6. На М-шаге продолжается пересчет центров кластеров (1.12):

$$\mu_{yj} = \frac{1}{l w_y} \sum_{i=1}^{l} g_{iy} f_j(x_i), y \in Y, j = 1, .., n; \quad (1.12)$$

В формуле выше внутри сумматора берется *j*-ая числовая характеристика *i*-ого кластера и умножается на вероятность принадлежности этого элемента к кластеру y.

7. Последним действием М-шага является вычисление ковариационной матрицы (1.13):

$$\sigma_{yj}^2 = \frac{1}{l w_y} \sum_{i=1}^{l} g_{iy} (f_j(x_i) - \mu_{yj})^2, j = 1, .., n; \quad (1.13)$$

8. После выполнения обоих шагов, необходимо отнести каждый элемент выборки к тому кластеру, вероятность принадлежать к которому максимальна.

Главными достоинствами алгоритма являются:

- способен анализировать малые объемы данных;
- может быть использован с другими алгоритмами;
- не требует определения метрик.

Существенным недостатком является зависимость результата работы алгоритма от первоначального вида распределения.

Гиперсегментация пользователей сайта является примером задачи кластеризации многомерных данных и требует особого подхода к ее решению.

Алгоритм сегментации пользовательских профилей должен удовлетворять нескольким требованиям:

1. Способен работать с большим объемом многомерных данных.
2. Не требует указания искомого количества кластеров.
3. Не требует указания максимальной величины кластера.
4. Не чувствителен к шумам.
5. Не требует обучения.
6. Легко поддается модификации и комбинированию с другими алгоритмами и процедурами.

Среди рассмотренных видов алгоритмов кластеризации указанным требованиям в полной мере не соответствует ни один. Однако в наибольшей степени оказались близки неиерархические плотностные алгоритмы, примером которого является алгоритм DBSCAN.

Таблица 1.1- Сравнение алгоритмов кластеризации многомерных данных

Метод	Способность работать с большими объемами данных	Отсутствие необходимости указания максимального количества искомых кластеров	Отсутствие необходимости указания максимального размера кластеров	Устойчивость работы не смотря на наличие шума	Отсутствие необходимости обучения	Легкость модификации и комбинирования с другими алгоритмами и процедурами
Иерархическая кластеризация			нет		да	нет
Неиерархические – итеративные методы (к-средних)	нет	нет	да	нет	да	
Неиерархические – плотностные алгоритмы (DBSCAN)	да	да	да	да	да	да
Неиерархические - статистические алгоритмы кластеризации (EM-алгоритм)	нет	да	да		да	да

В результате проведенного анализа методов кластеризации различных типов – иерархических и неиерархических (плотностных, статистических, итеративных) – было выявлено, что алгоритмы иерархического типа на

начальном этапе требуют указания максимального размера кластера (максимально возможного количества элементов), что нежелательно для задачи разбиения пользователей сайта по группам. Алгоритмы неиерархические итеративные не могут быть использованы в связи с необходимостью указания количества искомых кластеров, медленно работают при анализе больших объемов данных, а также чувствительны к шумам и могут давать недостоверные результаты в связи с этим. Неиерархические статистические алгоритмы также плохо справляются с анализом больших данных. Наиболее подходящими оказались плотностные алгоритмы, которые хорошо работают с большим объемом данных, не требуют указания количества кластеров и их размеров. Главными преимуществами подобных алгоритмов является устойчивость к данным шума и легкость их модификации и комбинирования с другими алгоритмами и процедурами. Таким образом, разработка нового алгоритма гиперсегментации пользовательских профилей, удовлетворяющего всем указанным требованиям, будет осуществляться на основании плотностного алгоритма DBSCAN, который предполагается дополнительно модифицировать.

2 ПРОЕКТИРОВАНИЕ ФОРМАЛЬНЫХ СРЕДСТВ ФОРМИРОВАНИЯ ГРУПП ПОЛЬЗОВАТЕЛЕЙ САЙТА

2.1 Разработка подхода персонализированного управления сайтом

В условиях сильной конкуренции на рынке одним из самых важных условий достижения успеха наряду с такими параметрами, как ассортимент, стоимость продукта или услуги, месторасположение, наличие дополнительных услуг и качество для любой компании, является клиентоориентированность.

Под клиентоориентированностью понимают:

- интерес к потребностям и желаниям потребителя;
- способность взаимодействовать с потребителем с учетом его ожиданий:
- способность выявлять и удовлетворять потребности потребителя.

Таким образом, принципы работы маркетинга в классическом его понимании нацелены на достижение клиентоориентированности организации:

1. Исследование рынка.
2. Сегментация рынка.
3. Гибкое реагирование производства и сбыта.
4. Инновация.

Однако реализация данного подхода при активном использовании инструментов электронной коммерции без его модификации не представляется возможной. Ввиду отсутствия прямого контакта с потребителями при осуществлении продаж через Интернет, компании вынуждены использовать иные инструменты анализа потребительских интересов, нежели при личном его присутствии. Это говорит о необходимости разработки новой концепции персонализированного управления контентом сайта.

В связи со спецификой электронной коммерции персонализированный подход к потребителю возможно реализовать, используя следующие шаги:

1. Предварительный сбор информации о потребителях.
2. Проведение гиперсегментации пользователей.
3. Анализ результатов гиперсегментации.
4. Генерация контента для соответствующих сегментов пользователей.
5. Автоматический подбор контента при обращении пользователя к сайту.
6. Параллельный сбор и анализ информации о новых пользователях.
7. Оценка результатов применения персонализированного для разных групп пользователей контента.

Таким образом, подход персонализированного управления контентом сайта, представленный на схеме (рисунок 2.1), значительно отличается от классического подхода к управлению сайтом [56]. Типовая структура системы управления контентом нуждается в доработке и расширении функционала с целью реализации дополнительных функций – гипергесментации пользовательских профилей и подбора персонализированного контента для разных групп пользователей при обращении к странице сайта.

Прежде чем запустить работу Интернет-ресурса с персонализированным контентом, необходимо осуществить установку системы управления контентом сайта или интегрировать дополнительный функционал (гиперсегментации и персонализации контента) в уже имеющуюся CMS. Установку и первичную настройку осуществляет группа внедрения. В том случае, если сайт еще не публиковался в сети, созданием и загрузкой шаблонов представления элементов страниц сайта в систему занимаются дизайнеры.

Для создания и загрузки динамически изменяющегося контента, а также настройки подмен необходимо понимать, для каких групп

пользователей и каким образом необходимо изменять содержимое web-страницы.

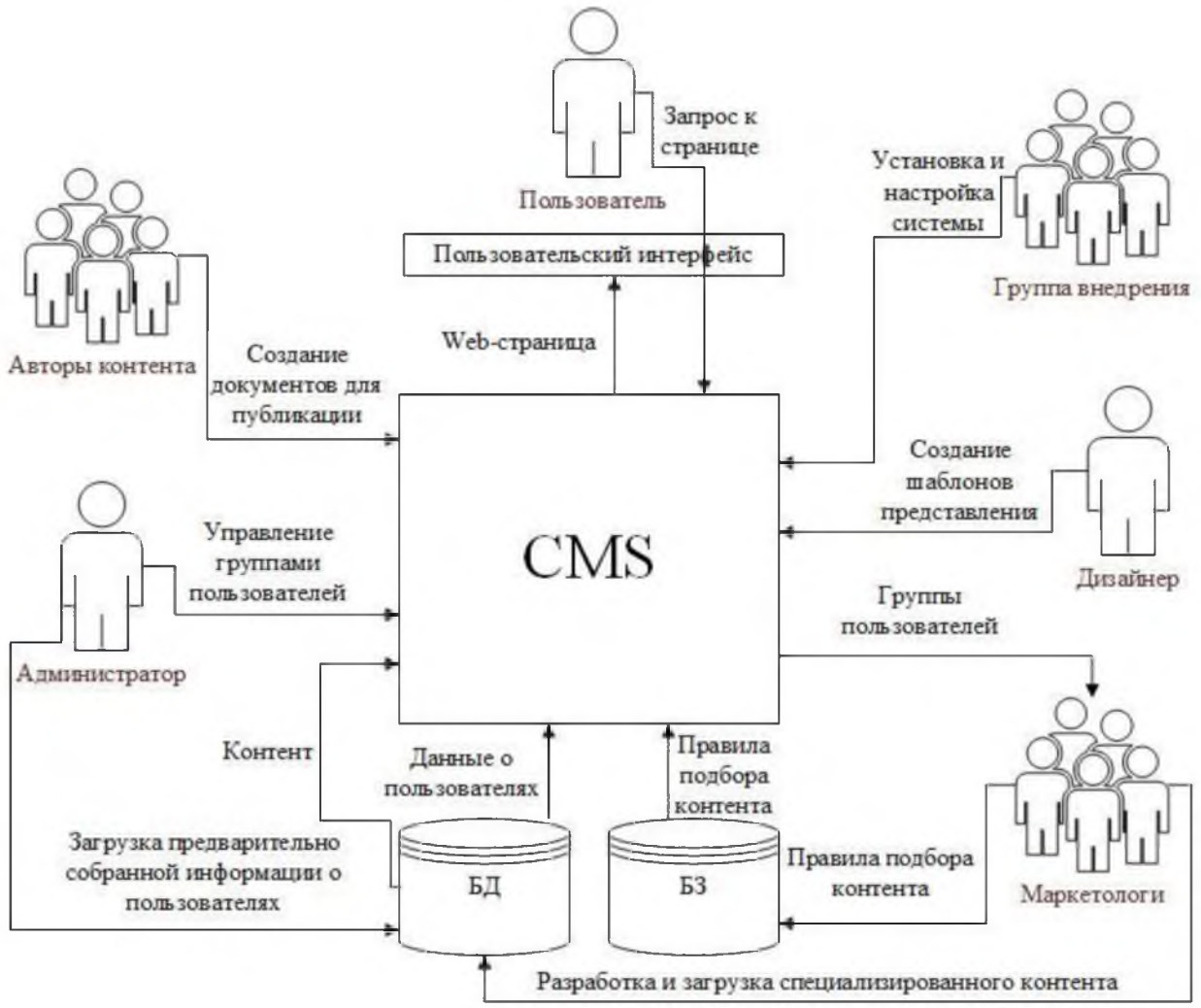

Рисунок 2.1 – Схема персонализированного управления сайтом

На данном шаге предварительно собранные данные о пользователях, относящихся к предметной области, в рамках которой осуществляется деятельность предприятия, должны быть загружены в соответствующий модуль системы для осуществления кластеризации пользовательских портретов. Результатом работы модуля являются выделенные группы пользователей, отличающиеся по некоторым из признаков их характеризующих. Для полученных групп разрабатываются варианты информационного наполнения блоков страницы сайта и загружаются в Базу данных. Также на основании полученной в результате анализа имеющихся данных о пользователях информации в Базу знаний вносятся

соответствующие правила подбора содержания информационных блоков для пользователей, принадлежащих к разным выявленным группам.

Для любых элементов страницы сайта возможна разработка вариантов содержания:

- заголовок сайта;
- подзаголовок сайта;
- текстовый блок, поясняющий заголовок/подзаголовок;
- рекламный баннер;
- заголовок формы обратной связи/оставления заявки;
- кнопка с призывом к действию и другие.

Решение о том, какие именно элементы будут персонализированы, принимает специалист – маркетолог, который будет работать с системой, либо сторонние специалисты, хорошо знакомые с предметной областью. Функции администратора сайта остаются теми же – он управляет группами пользователей, а также отвечает за публикацию статичного контента, который в свою очередь готовится авторами контента – сотрудниками организации.

Предложенный подход обеспечивает персонализированный подбор контента для каждого пользователя, посетившего сайт компании, в соответствии с особенностями каждой группы, к которой он относится. Таким образом, вероятность удовлетворения его интереса при посещении сайта возрастает, а значит растет и вероятность совершения им покупки.

Стоит отметить, что одновременно с запуском работы сайта начинается сбор собственной статистической информации о его посетителях, которая также записывается и принимается к анализу. Данные о пользователях непрерывно накапливаются - чем больше ресурс, тем с большей скоростью происходит их поступление. Чтобы избежать избытка ненужных данных, а также с целью учёта актуальных предпочтений пользователя, устаревшие данные необходимо очищать. При таком подходе в зависимости от пропускной способности ресурса необходимо задать максимальный период хранения данных, по истечении которого они будут удалены. В таком ключе

изменения в интересах пользователей не останутся незамеченными и у компании появляется возможность максимально быстро на них реагировать путем доработки контента и правил для вновь появившихся групп пользователей и удаления действующего функционала для исчезнувшего кластера.

Таким образом предложенный подход персонализированного управления сайтом направлен на подбор и показ релевантного для разных групп пользователей контента страницы сайта с учетом интересов и потребностей пользователей сайта, а также изменения их предпочтений во времени. Подход предполагает сбор и обработку информации о пользователях, гиперсгментацию пользовательских профилей, разработку специализированного контента определенных блоков структуры страницы и настройку отображения необходимого содержания для той или иной группы.

2.2 Систематизация показателей и параметров сегментации интернет-пользователей

Современные компании, продвигающие и продающие свои товары и услуги в сети Интернет, в условиях большой конкуренции вынуждены все больше внимания уделять исследованиям своей целевой и потенциально целевой аудитории, а также стараться максимально близко подстраиваться под их интересы. По факту этот многоэтапный процесс оказывается трудо и время затратным. В данном процессе они сталкиваются с необходимостью решения нескольких задач:

- кластеризация пользователей – группирование пользователей схожих по общим признакам в группы.
- построение поведенческих профилей пользователей.
- построение расширенных профилей пользователей с учетом социально демографических данных.

– сегментация клиентской базы согласно расширенным профилям по анкетным и поведенческим данным.

– прямой маркетинг – предоставление рекламных и маркетинговых предложений пользователю в соответствии с его расширенным профилем.

– персонализация контента – предоставление пользователю сайта наиболее интересной ему информации.

Современные интернет-маркетологи выделяют несколько видов сегментации пользователей.

Сегментация по категории продуктов — один из самых надежных типов сегментации. Он основывается на товарах, рассматриваемых потенциальным покупателем, чтобы рекламодатели могли размещать объявления уникальных кампаний в небольших сегментах. Например, спортивный онлайн-магазин, пытающийся добраться до любителей летних пробежек, может легко разделить активности, продвигая выбранные модели обуви или подходящего оборудования людям, просматривавшим предметы в рамках одной категории.

Рекламные активности можно также сосредоточить на определенных типах пользователей, например, тех, кто еще не подписался на рассылку интернет-магазина. Благодаря информации, собранной с помощью специального тега или клиентского репоста, персонализированная кампания ретаргетинга может легко идентифицировать и информировать эту группу неохваченных пользователей о специальных акциях для уже зарегистрированных пользователей. Это полезно для роста новых подписчиков, избегая двойного участия для тех, кто уже подписан.

Продвинутые решения ретаргетинга позволяют запускать дополнительную кампанию, ориентированную на пользователей, которые не посещали веб-сайт магазина в течение длительного периода времени (например, 14 или 30 дней) или тех, кто купил телевизор за последние три месяца. Такого рода тактика помогает повысить узнаваемость бренда, поддерживать долгосрочные отношения и наблюдать за потенциальными

покупателями. Показывая заметки «спасибо» в объявлениях, эксклюзивных скидках или сообщении о предстоящих продажах, современные руководители кампаний могут планировать долгосрочные стратегии, чтобы сохранить свой бренд на слуху, одновременно отвлекая внимание от брендов конкурентов.

Кампании также могут быть сегментированы по интенсивности покупки или количеству предложений, которые просматривают посетители, шопперы и покупатели, а также добавляют в корзину или покупают. Маркетологи могут запускать отдельные кампании с различными сообщениями и рекламными объявлениями, которые используются для пользователей, которые часто покупают (например, направляют специальную скидку для возвращающихся клиентов) и теми, кто редко или никогда не покупает на веб-сайте электронного магазина (например, отправляет код купона для первой покупки).

Если целью является увеличение продаж товаров с высокой ценностью или высокой ценой, то современный персонализированный ретаргетинг может быть оптимизирован и для конкретных позиций или товарных групп. В таких случаях самообучающиеся алгоритмы будут выбирать для отображения варианты того же ценового диапазона, что и продукты, ранее просмотренные пользователем.

Теперь маркетологам можно сегментировать потребителей в соответствии с устройством, на котором они выполняют поиск: десктопом, мобильным телефоном, телевизором или любым другим устройством, используемым для запроса продукта в Интернете. Рекламные кампании на разных платформах позволяют маркетологам запускать кросс-платформенные кампании, таргетировать перемещающихся в пространстве людей и использовать в своих предложениях не только тайминги, но и варианты экранов, которые, скорее всего, будут использоваться.

Во время работы пользователя Интернет вместе с его запросами и через счетчики посещений передаются некоторые первичные данные о нем:

– адрес компьютера;

– источник перехода на сайт,

– информация о действиях пользователя.

На серверах такие данные записываются в специальные журналы и и могут быть использованы для проведения анализа. В базу данных сайта также в профили пользователей, а также в журналы работы Web-приложений записываются различные дополнительные данные о нём:

– информация из регистрационных форм;

– продукты и их рейтинги;

– рейтинги статей;

– сделанные пользователем покупки.

На основании анализа подобного рода информации возможно определение явных и скрытых предпочтений пользователя. Сбор и хранение обозначенных данных не представляет никакой сложности, так как практически все Web-серверы поддерживают функцию сохранения журналов запросов. Серверы также способны предоставлять отчеты о собранных данных.

Для сбора данных о данных в Web-аналитике необходимо использовать так называемые метрики сайтов. При проведении сегментации пользователей эти метрики являются основой, так как соотносятся с маркетинговыми переменными. Традиционно в процессе сегментации потребителей используются переменные двух групп:

Из описательных характеристик часто используют:

1. Географические:

– страна;

– район;

– область проживания и т.п.;

2. Демографические:

– возраст,

– пол,

– семейное положение,

– доходы,

– социальный класс и т.п.;

3. Психографические: тип личности и т.п.;

Также применяются характеристики поведения, среди которых могут быть способы использования продукта и др.

Среди метрик сайтов могут быть выделены:

1. Сведения, которые передаются с компьютера пользователя автоматически при посещении сайта:

– данные компьютера, получаемые через поля заголовка HTTP-запроса (характеристики ПО, системный язык, источник перехода на сайт, поисковый запрос пользователя, который привёл на сайт или страницу, географическое расположение провайдера, Cookies и т.п.);

– данные компьютера, которые могут быть получены из Web-браузера с помощью счётчиков посещений (характеристики монитора, история просмотров страниц в текущем сеансе работы браузера и др);

2. Дополнительная информация с сайта: ключевые слова просмотренного содержимого и атрибуты интересующих продуктов или услуг;

3. Обобщённая Интернет-статистика:

– глобальная и региональная Интернет-статистика, которую можно найти на сайтах W3Counter, Bigmir)net, SpyLog и др.;

– метрики отраслевой статистики (benchmarking), включающие в себя сведения о посетителях сайтов в зависимости от их отраслевой принадлежности и предоставляемые такими Web-службами, как Google Ad Planner, Google Trends, Google Benchmarking, Coremetrics, ClickZ Stats, Fireclick и др.

Таким образом, проведя теоретико-множественное описание для гиперсегментации пользователей возможно использование:

Множества метрик сайта MS=$\{ms_1, ms_2, ms_3\}$, где

ms_1 – данные компьютера, получаемые через поля заголовка HTTP-запроса;

ms_2 – данные компьютера, которые могут быть получены из Web-браузера;

ms_3 – дополнительная информация с сайта;

Элементами подмножества данных, полученных через поля заголовка, $MS_1 = \{dcz_1, dcz_2, \dots, dcz_k\}$ являются:

dcz_1 – характеристики ПО;

dcz_2 – системный язык;

dcz_3 – источник перехода на сайт;

dcz_4 – поисковый запрос пользователя;

dcz_5 – географическое расположение провайдера;

dcz_6 – Cookies и т.д.

Элементами подмножества данных, полученных через Web-браузер, $MS_2 = \{dcb_1, dcb_2, \dots, dcb_l\}$ являются:

dcb_1 – характеристики монитора;

dcb_2 – история просмотров страниц в текущем сеансе работы браузера и другие;

Элементами подмножества дополнительных данных, полученных с сайта, $MS_3 = \{dcd_1, dcd_2, \dots, dcd_m\}$ являются:

dcd_1 – ключевые слова;

dcd_2 – атрибуты интересующих продуктов;

dcd_3 – атрибуты интересующих услуг;

dcd_4 – информация о купленных товарах или услугах;

dcd_5 – рейтинги просмотренных продуктов и другие.

В различных областях, где используется сегментация пользователей - существуют свои методики её выполнения.

В маркетинге сегментация потребительских рынков выполняется в три этапа:

1. Выбор критериев (переменных) сегментации;
2. Подбор метода сегментации;
3. Выбор целевых сегментов.

Однако этого набора шагов недостаточно для сегментации пользователей Web-сайтов.

Во-первых, в данном случае при выборе метрик и методов необходимо учитывать область применения результатов сегментации: это может быть повышение эффективности работы сайта, его персонализация, или же уточнение потребительских сегментов.

Во-вторых, первичные данные о пользователях, сохраняемые в журналах серверов, непригодны для непосредственного использования и нуждаются в дополнительной обработке.

Чтобы их использовать, необходимо:

– предварительно их очистить от несущественной информации вроде загрузок изображений или же записей про посещение сайта Web-агентами, сжать и трансформировать в удобный для поиска и анализа важной и полезной информации;

– из этих данных построить многомерный массив, где в качестве измерений будут использованы URL, время, IP-адреса, информация о содержании посещённых Web-страниц, дополнительные данные о пользователе из журналов Web-приложений, чтобы затем иметь возможность определить характеристики и последовательности действий пользователей, вычислить поведенческие метрики сайта и выполнить сегментацию.

С учётом необходимости дополнительной обработки первичных данных о пользователе, сегментация пользователей может выполняться как итеративная и адаптивная последовательность фаз:

1. Фаза определения бизнеса или постановки исследования.

2. Фаза сбора, анализа и выборки данных.

3. Предварительная очистка, объединение и интеграция данных.

4. Фаза моделирования и сегментации.

5. Фаза оценки результатов.

6. Фаза применения результатов.

Некоторые фазы сегментации могут зависеть от результатов предыдущих фаз. В свою очередь любая фаза может быть повторно выполнена с новыми условиями, если этого будет нужно для удовлетворительного выполнения последующих фаз. Например, в зависимости от поведения и характеристик модели сегментации может появиться необходимость вернуться к фазе подготовки данных для их дополнительной очистки перед фазой оценки результатов.

В ходе исследования были выявлены источники информации о пользователе, который обращается к сайту – данные из полей заголовка http, данные из web-браузера, дополнительные данные с сайта. Было дано теоретико-множественное описание показателей, по которым могут быть сегментированы пользователи ресурса.

2.3 Проектирование структуры модифицированной CMS

Использование систем управления контентом сайта для его создания и сопровождения имеет ряд неоспоримых преимуществ. CMS-системы позволяют создавать контент ресурса, легко управлять им (хранить, редактировать, удалять) и доступом к нему, осуществлять публикацию контента в форме необходимого представления даже тем людям, которые не имеют соответствующей квалификации в области информационных технологий. Однако в последнее время далеко не все web-ресурсы способны удовлетворить потребности различных пользователей в той информации, которую они ищут.

Причина подобного явления кроется в том, что все посетители сайта, имея разные интересы видят один и тот же контент. Соответственно, получив

информацию, не соответствующую ожиданиям частично или вовсе далекую от искомой, пользователь, как правило, менее охотно идет на выполнение целевых действий на сайте или же вовсе покидает его. В условиях жесткой рыночной конкуренции в сфере электронного бизнеса остро встает вопрос об изменении подхода к взаимодействию с клиентом – от одинакового для всех к персонализированному.

Задача подготовки и настройки вывода персонализированного содержимого страницы ресурса для каждого пользователя может показаться нерешаемой, однако учитывая современные возможности сбора и анализа информации в сети Интернет, возможна разработка и реализация подхода, основанного на разбиении пользователей со схожими интересами на группы и настройка работы ресуса под каждую из них.

В связи с тем, что типовая структура CMS систем не предназначена для реализации функции персонализации контента сайта, необходимо осуществить доработку модуля контент менеджера и структуры БД системы управления контентом с учетом новых требований к ней, а также подключить Базу знаний, в которую заносятся правила подбора персонализированного контента (рисунок 2.2) [52, 53].

В связи с этим функционал новой системы получается следующий:

– создание контента сайта (материалов для размещения на его страницах, а также вариантов наполнения персонализируемых блоков);

– управление контентом сайта (контроль за созданием, изменением и удалением версий документов и файлов, их хранением, осуществление контроля за доступом к файлам, а также осуществление интеграции с иными информационными системами);

– гиперсегментация пользовательских профилей (разбиение пользователей со схожими параметрами на группы);

– персонализированная публикация контента и его представление (вывод персонализированного для разных групп пользователей контента

блоков структуры страницы ресурса в соответствии с заданным шаблоном дизайна всего сайта при обращении пользователя к ней).

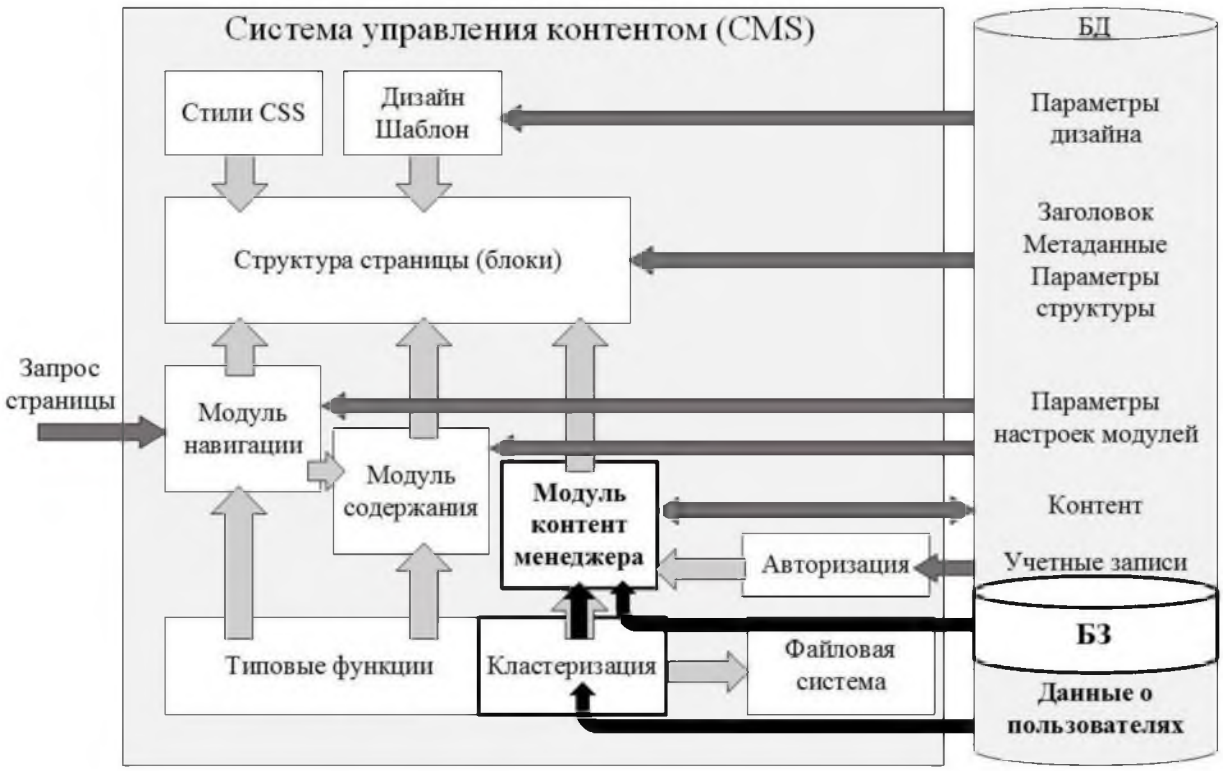

Рисунок 2.2 – Структура модифицированной CMS

В структуру новой системы входят следующие элементы:

1. Модуль навигации (необходим для формирования структуры сайта – задания взаимосвязей и подчиненности страниц, настройки меню и его отображения).

2. Модуль содержания (содержит информацию о содержании структурных элементов страницы).

3. Модуль контент менеджера (в новой системе в этом модуле реализуется функция гиперсегментации профилей пользователей, а также с его помощью осуществляется управление процессом подбора контента для отображения страницы).

4. Стили CSS (необходимы для определения оформления элементов страниц сайта при помощи языка гипертекстовой разметки).

5. Дизайн-шаблон (внешний вид страниц сайта, всех его элементов).

6. Файловая система (модуль предназначен для управления скачиваемыми файлами с сервера – прайсы или каталоги моделей, которые посетители сайта могут скачать, фотографии товаров и т.п).

7. Модуль авторизации (необходим для управления регистрацией и авторизацией для пользователей).

8. База данных (в базе данных системы хранятся параметры дизайна, заголовок, мета-данный, параметры структуры, параметры настроек модулей, контент, учетные записи и данные о пользователях сайта).

9. База знаний (содержит правила подбора элементов контента страницы для разных групп пользователей).

Кроме функций, типичных для систем управления контентом, новая система осуществляет кластеризацию пользовательских портретов, которая направлена на распределение схожих между собой пользователей на группы. Каждая выявленная группа может получать различный контент при обращении к странице.

Для проведения анализа данных о пользователях, информация о них должна храниться в общей Базе данных системы. Соответственно, данные о новых пользователях должны поступать и храниться там же. При запуске модуля, отвечающего за управление контентом в рамках нового подхода – персонализации – данные о пользователях должны поступать в модуль для анализа.

В результате процедуры гиперсегментации пользователей информация о выявленных кластерах и их основных характеристиках становится доступной специалистам-маркетологам, которые на основании своих профессионального опыта и знаний разрабатывают контент, релевантный для специфических групп посетителей Интернет-ресурса. Варианты контента также хранятся в Базе данных. Специалисты также генерируют и сохраняют правила, согласно которым будет происходить подбор контента для пользователей разных кластеров. Все правила сохраняются в Базе знаний и

при обращении к ней конкретное правило передается в модуль управления контентом, который соответствующим образом выбирает нужный контент из Базы данных и собирает страницу для пользователя.

Процесс взаимодействия элементов новой системы в ходе персонализированного управления контентом ресурса показан на рисунке 2.3 при помощи диаграммы последовательности в нотации UML.

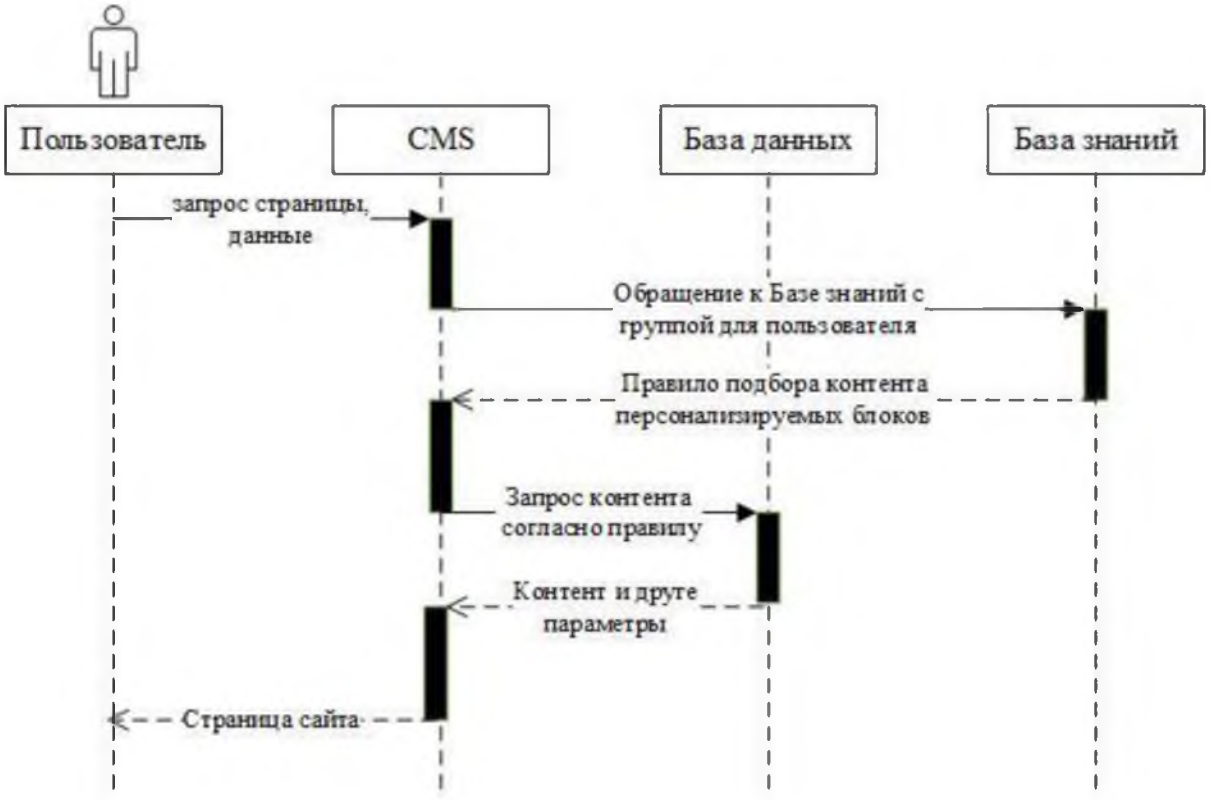

Рисунок 2.3 – Структура модифицированной CMS

1. Запрос пользователя к странице сайта поступает в систему управления контентом сайта. Вместе с ним в систему поступают и данные о нём (источник перехода на сайт, поисковый запрос пользователя, который привёл на сайт или страницу, географическое расположение провайдера, Cookies, история просмотров страниц в текущем сеансе работы браузера и другие). В модуле контент менеджера системы происходит определение группы для пользователя или наиболее близкого к нему кластера.

2. Система управления контентом на основании выявленного кластера осуществляет запрос к Базе знаний.

3. В ответ на запрос в систему поступает правило подбора содержимого персонализируемых блоков страницы сайта.

4. Система направляет запрос контента в Базу данных.

5. Вместе с необходимым контентом в систему поступают все необходимые параметры – дизайна, структуры, а также настроек модулей. В системе происходит сборка страницы на основе всей полученной из Базы данных информации, шаблона страницы и настроек дизайна.

6. Система управления контентом сайта передает собранную персонализировано для пользователя передается ему.

В ходе эксплуатации системы происходит сбор данных о новых пользователях, а также об изменяющихся потребностях уже имеющихся. Подобный подход позволяет фиксировать новообразовавшиеся кластеры и распад выделенных ранее кластеров. В этом случае, в систему вносятся соответствующие изменения – для нового кластера разрабатывается контент и правила, для кластера, который распался – происходит удаление правила и вариантов контента.

Таким образом, предложенная структура новой системы управления контентом сайта позволяет реализовать подход персонализированного управления web-ресурсом благодаря модификации модуля контент менеджера, расширению Базы данных с целью хранения информации о пользователях и вариантов контента структурных блоков страницы ресурса, а также подключению и настройке Базы знаний, необходимой для подбора специального для разных групп пользователей контента.

3 РАЗРАБОТКА АЛГОРИТМА СЕГМЕНТАЦИИ ПОЛЬЗОВАТЕЛЕЙ ИНТЕРНЕТ-РЕСУРСА

3.1 Разработка алгоритма сегментации пользователей ресурса

На основе модифицированного плотностного алгоритма DBSCAN в сочетании с оригинальными процедурами был разработан новый алгоритм гиперсегментации пользователей web-ресурса (рисунок 3.1), у которого отсутствуют обозначенные ранее слабые стороны, выявленные при анализе алгоритмов кластеризации многомерных данных [55].

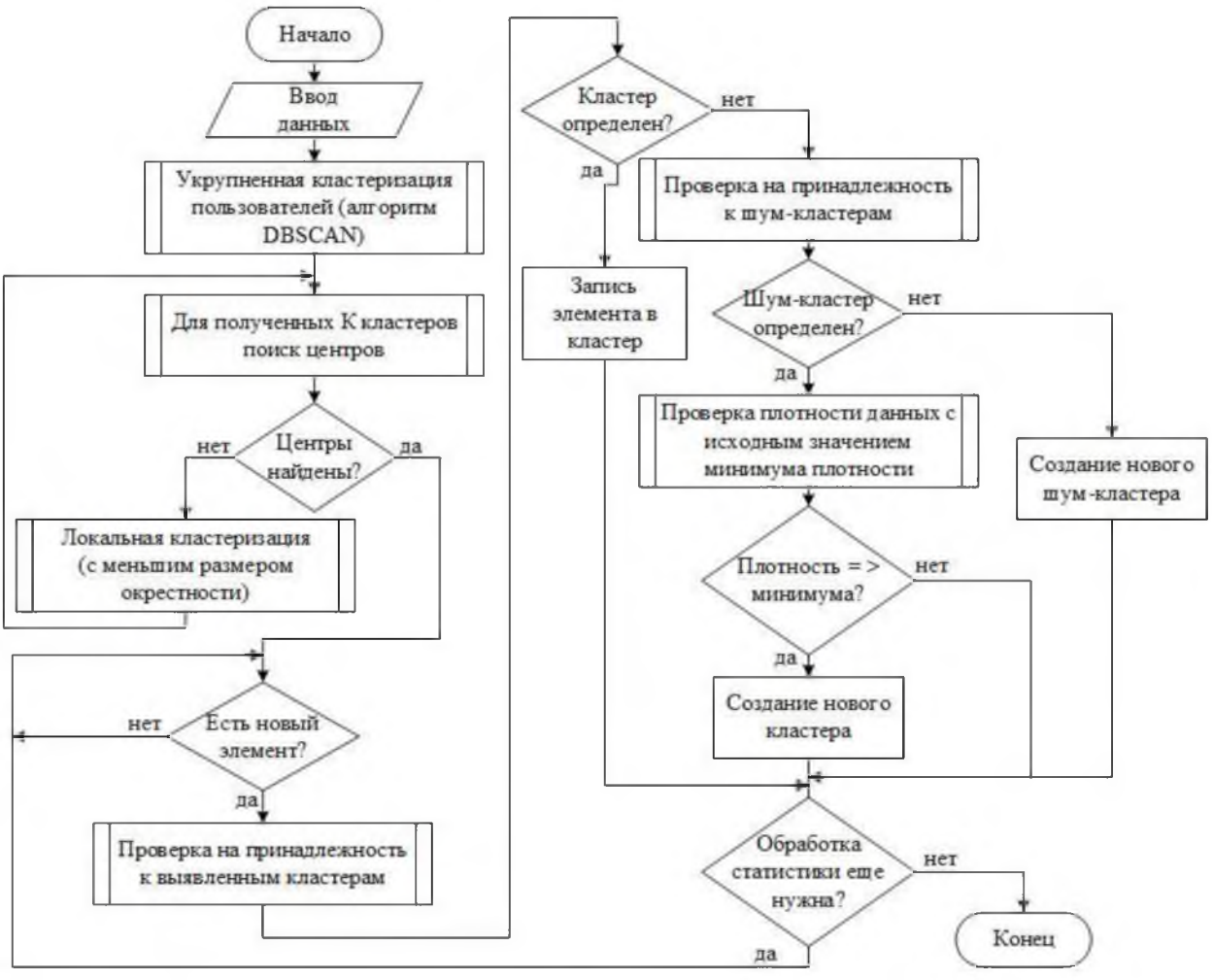

Рисунок 3.1 – Алгоритм гиперсегментаци пользователей Интернет-ресурса

Предложенный алгоритм направлен на получение кластеров (групп пользователей со схожими интересами) различных размеров, а также шум-кластеров, и позволяет оперативно учитывать изменяющиеся во времени потребности пользователей за счет постоянного сбора и анализа новых данных. Алгоритм представляет собой следующую последовательность процедур:

1. Укрупненная кластеризация профилей пользователей.
2. Поиск центров кластеров.
3. Для кластеров неправильной формы – локальная кластеризация объектов.
4. Для нового элемента поиск кластера или шум-кластера.
5. При достижении минимума плотности шум-кластера – преобразование в новый кластер.

Первым шагом осуществляется укрупненная кластеризация профилей пользователей при помощи модифицированного алгоритма DABSCAN (рисунок 3.2). Суть метода остается неизменной, однако его отличительной особенностью в данном случае является распределение точек шума по шум-кластерам в зависимости от их координат, вместо записи в одну единственную категорию «Шум». Таким образом, при поступлении новых данных возможно увеличение плотности разброса пользовательских портретов в определенной области – шум-кластере, что повлияет на принятие решения о создании полноценного кластера.

Алгоритм запускается после ввода всех данных и определения значений:

X = $\{x_1,\ x_2, \dots, x_n\}$ – множество значений для анализа.

ε – значение радиуса окрестности.

$MinPts$ – минимальное количество точек-соседей для формирования кластера. При этом параметры ε и $MinPts$ выбираются индивидуализировано для каждого ресурса в зависимости от интенсивности поступления новых данных, отражающих масштаб ресурса.

Случайным образом осуществляется выбор объекта для анализа из массива загруженных данных. При помощи евклидова расстояния (3.1):

$$p(x,y) = (\sum_{i=1}^{n}(x_i - y_i)^2)^{1/2} \qquad (3.1)$$

Проводится проверка окрестности точки на наличие соседей – точек, расстояние до которых меньше заданного радиуса окрестности ($p(x,y) \leq \varepsilon$).

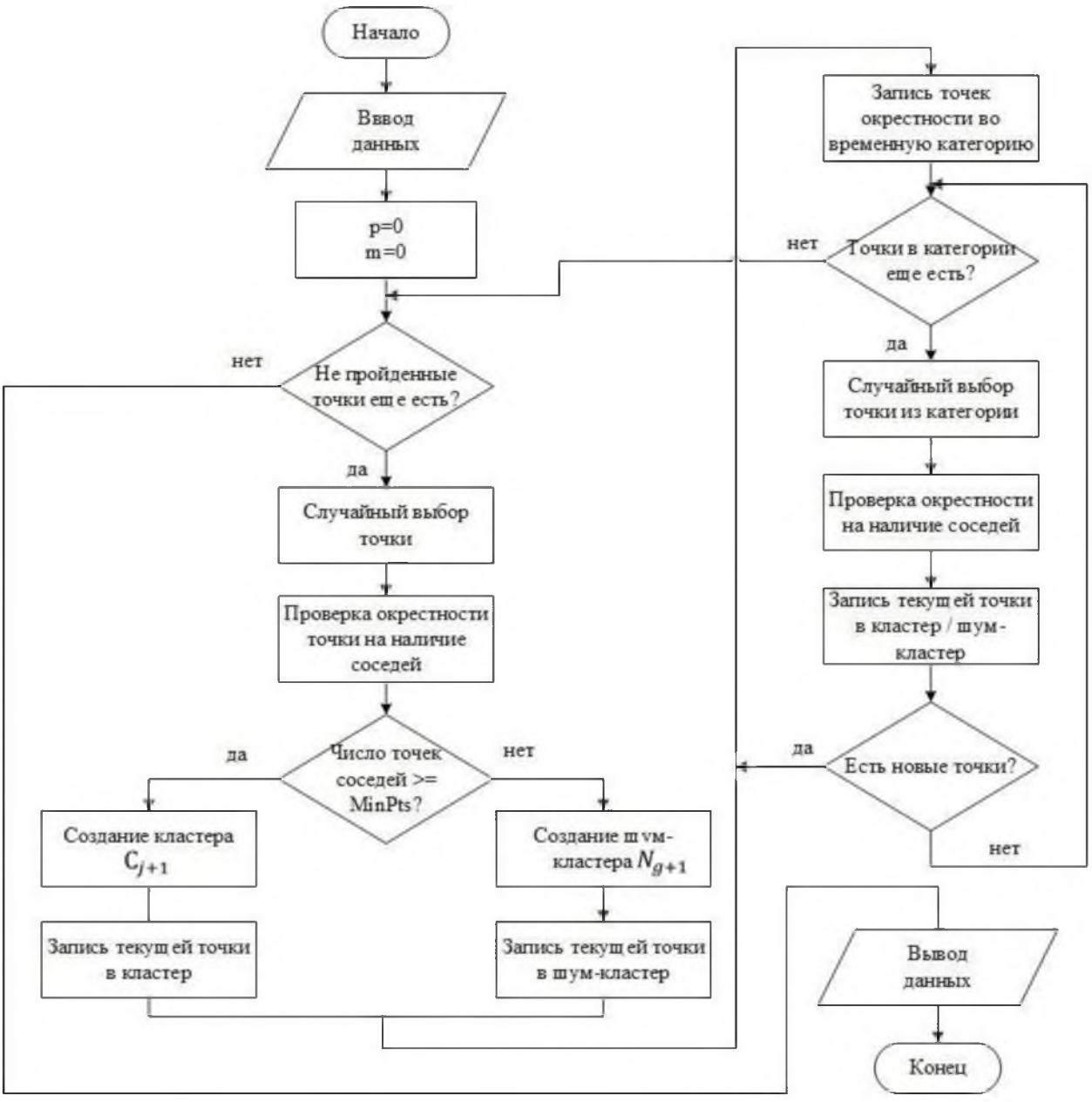

Рисунок 3.2 – Модифицированный алгоритм DBSCAN

Если количество точек – соседей в окрестности больше минимально заданного значения (MinPts), то создается кластер (C_{j+1}), где j – текущее количество кластеров, соответствующее m. Точка, с которой анализ был начат, помечается, как пройденная и принадлежащая к кластеру. Все остальные точки из окрестности помещаются во временную категорию.

Случайным образом осуществляется переход к любой точке из категории. Для новой точки осуществляется аналогичная проверка окрестности. Если в эту окрестность попали точки, ранее не занесенные в категорию, то они также записываются в эту временную категорию. Точка помечается как пройденная и принадлежащая к текущему кластеру. Действие повторяется до тех пор, пока в категории не закончатся точки, т.е. пока алгоритм не дойдет до границы кластера. Таким образом формируется текущий кластер.

Если в массиве входных данных еще остались не пройденные точки, то случайным образом выбирается одна из них и процедура проверки окрестности повторяется для нее. В том случае, если количество соседствующих точек меньше минимально заданного MinPts, принимается решение о создании шум-кластера (N_{g+1}). Тем же образом, что и для полноценного кластера, который описан выше. Если в окрестности анализируемого объекта не оказалось точек-соседей, шум-кластер создается из одного элемента – текущей точки.

Таким образом, в результате работы алгоритма для процедуры «Укрупненная кластеризация» получится m кластеров и p шум-кластеров.

Далее для выявленных кластеров проводится поиск центроидов (3.2):

$$c_j = \frac{\sum_{i=1}^{L} x_i}{L} \qquad (3.2)$$

где c_j - это центр j-го кластера, L – количество элементов (точек) в j-м кластере. Эта процедура необходима для определения областей наибольшего скопления элементов выборки. Кроме того, в результате работы модифицированного плотностного алгоритма DBSCAN в случае особенностей

разброса элементов исходной выборки в n-мерном пространстве возможно получение кластеров неправильной формы – вытянутых, лентовидных, спиральных и т.д. В этом случае весовой центр кластера определить не удастся, для чего алгоритмом предусмотрена локальная кластеризация, направленная на распределение объектов выявленного кластера на более мелкие – с меньшим значением величины окрестности для объектов. Процедура определения центральных точек в этом случае производится для каждого вновь полученного кластера.

После сегментации данных о пользователях из начальной выборки в процессе работы Интернет-ресурса в систему продолжают поступать новые данные. Для каждого вновь полученного элемента проводится проверка на принадлежность к выделенным кластерам согласно тому же подходу, который используется при разбиении выборки объектов на группы – при помощи проверки окрестности нового объекта.

Когда элемент не соответствует ни одному из кластеров, проводится проверка на принадлежность к выделенным шум-кластерам. В том случае, если и шум-кластер не определен, создается новый шум-кластер, состоящий из этой точки. Если же элемент отнесен к оному из шум-кластеров, то для элементов шум-кластера осуществляется проверка плотности окрестности с учетом начального значения минимума плотности. Решение о создании нового кластера, когда минимальное значение равно минимуму или больше него ($p(x, y) \leq \varepsilon$).

Таким образом, предложенный алгоритм гиперсегментации пользовательских профилей представляет собой комбинацию из модифицированного алгоритма DBSCAN и набора оригинальных процедур. Относительно алгоритмов кластеризации многомерных данных, рассмотренных в ходе исследования, предложенный имеет важные преимущества:

– не требует задания ожидаемого количества кластеров;

– позволяет учитывать изменяющиеся во времени потребности и интересы пользователей.

3.2 Разработка алгоритма динамического подбора персонализированного контента

Как правило, реализация подхода персонализации web-ресурса происходит на основании заранее заданных алгоритмов или правил. Оба предложенных подхода имеют свои сильные стороны, в связи с чем, предлагается их одновременное использование: на этапе определения кластера для посетителя сайта срабатывает алгоритм соответствующей процедуры «Определение кластера», далее, на основании результата (выявленной группы пользователей) согласно правилам, определяющим значения элементов контента для выявленной группы, производится сборка страницы сайта (рисунок 3.3) внутри CMS-системы [54].

Рисунок 3.3 – Алгоритм подбора персонализированного контента

При обращении пользователя к странице сайта система получает набор первичных данных о нём, на основании которых возможно отнесение пользовательского портрета к той или иной выделенной группе. Процедура идентификации аналогична той, которая используется на этапе кластеризации пользователей (рисунок 3.4) – для текущего объекта осуществляется проверка окрестности на наличие достаточного количества соседей.

Рисунок 3.4 – Алгоритм поиска кластера для нового пользователя

Если число соседствующих точек оказывается больше минимально заданного значения, тогда точка записывается в кластер, в котором уже находятся окружающие точки. Соответственно, формирование контента происходит согласно заданному для выявленного кластера правилу.

В том случае, если группу (кластер) определить не удается, т.е. точка отнесена к одному из шум-кластеров, система осуществляет поиск ближайшего к объекту кластера и, в соответствии с правилами подбора контента для найденной группы осуществляется сборка элементов контента. Определение ближайшего к рассматриваемой точке кластера происходит путем расчета расстояния до центров каждого из известных кластеров. В результате выполнения процедуры находится центроид, расстояние до которого наименьшее. Соответственно, контент будет формироваться на основании правил, заданных для кластера, центром которого он является.

Как было обозначено ранее элементы контента, разбитые на множества, хранятся в Базе данных системы соответствующими боками, выбор для которых определяется правилами, хранящимися в Базе знаний этой же системы управления контентом.

Объекты подмены контента для множества элементов страницы A=$\{A_1, A_2, \dots, A_i\}$ хранятся в четком соответствии с выявленными группами профилей пользователей P=$\{p_1, p_2, \dots, p_j\}$. Для каждого $p_j \in P$ заданы элементы в подмножествах $A_1, A_2, \dots, A_i$, поэтому правила, хранящиеся в Базе знаний, продукционного типа имеют следующую структуру:

Если пользователь P = p_1, то A_1=a_{11}, $A_2 = a_{21}$, …, $A_i = a_{ij}$.

Во множество элементов страницы A по решению специалиста, ответственного за персонализацию ресурса могут быть включены такие элементы как:

1. Множество заголовков сайта $A_1 = \{a_{11}, a_{12}, \dots, a_{1j}\}$.
2. Множество основных подзаголовков сайта $A_2 = \{a_{21}, a_{22}, \dots, a_{2j}\}$.
3. Множество текстовых блоков, поясняющих подзаголовок $A_3 = \{a_{31}, a_{32}, \dots, a_{3j}\}$.
4. Множество надписей кнопки с призывом к действию $A_4 = \{a_{41}, a_{42}, \dots, a_{4j}\}$.
5. Множество заголовков формы $A_5 = \{a_{51}, a_{52}, \dots, a_{5j}\}$.

6. Множество прочих элементов $A_i = \{a_{i1}, a_{i2}, \dots, a_{ij}\}$.

Таким образом, предложенный подход сочетает алгоритмический и основанный на правилах способы реализации персонализированного подбора контента сайта под разные категории его посетителей. Причем алгоритмическая часть позволяет определить, каким именно правилом система будет руководствоваться при построении страницы. Для пользователей, которые не могут быть отнесены к уже выявленным группам, предлагается использование правил ближайшего кластера.

3.3 Обоснование эффективности разработанных средств

В связи с научной и теоретической направленностью выпускной квалификационной работы, а также отсутствием программной реализации алгоритмического обеспечения персонализированного управления контентом web-ресурса, оценка эффективности работы предложенных алгоритмов путем проведения эксперимента не представляется возможной. В связи с этим предлагается применение вероятностного метода оценки.

Оценить эффект от применения подхода персонализированного управления сайтом возможно по нескольким его характеристикам, например, конверсии и прибыльности.

Конверсия интернет-ресурса представляет собой отношение количества пользователей, совершивших какие-либо целевые действия (регистрация, покупка, заявка подписка, переход по рекламной ссылке и т.д.) на нём, к общему числу посетителей сайта. Чаще всего этот показатель выражен в процентах (3.3):

$$\text{Конверсия сайта} = \frac{\text{Количество целевых действий}}{\text{Количество посетителей}} * 100\% \qquad (3.3)$$

Обычно результаты находятся в пределах от 0,5% (для мебели, музыкальных инструментов и спортивных товаров) до 15% (для сервисов

доставки еды). Исследование данных, предоставляемых маркетинговыми компаниями сети Интернет показало, что величина конверсии сайта зависит от многих факторов одновременно (дизайн и удобство сайта, качество текстов, источников трафика, цена, спрос на товары или услуги и многие другие) однако существуют некоторые тенденции показателей конверсии для разных отраслей. По данным компании DATA insight самые высокие показатели конверсии показывают сайты по доставке еды и продажи билетов на мероприятия, самый низкий показатель имеют сайты по продаже алкоголя и музыкальных инструментов (рисунок 3.5) [40].

Рисунок 3.5 – Показатели конверсии сайтов в зависимости от отраслей

Исследование источников в Интернет показало, что некоторые компании уже начали успешно применять подход персонализированного управления сайтом и открыто делятся своими результатами. Например, компании «Альянс Форест» из Санкт-Петербурга, которая занимается поставкой оборудования, запчастей и инструмента из Европы для промышленных предприятий России, сообщает о повышении конверсии сайта в 7 раз [47].

До начала использования нового подхода ситуация была следующей:

- конверсия по некоторым направлениям – менее 1%;
- одна посадочная страница;
- задача страницы — получить заявку от клиента.

При помощи сервиса YAGLA[1] была настроена подмена шести элементов страницы (заголовка страницы, призыва к действию, заголовка формы оставления заявки, а также некоторых других элементов) относительно продаваемых брендов [46]. В результате, конверсия по бренду Heidenhain в 2017 году выросла в среднем до 13,7%. Конверсия по остальным брендам колеблется от 5 до 23%, что зависит как от объемов трафика, так и специфики конкретного бренда. Общая конверсия сайта компании составила 7,45%. Таким образом, в результате применения персонализированного контента некоторых элементов станицы сайта, его конверсия выросла в 7 раз или на 645%.

Еще одним примером успешного применения персонализированного контента можно назвать опыт компании Lenovo. Совместно с компанией Neustar была проведена сегментация посетителей в «профилях» на основе анонимных данных на уровне домохозяйства в Интернете. В результате посетители ресурса компании были определены в несколько ключевых групп

[1] Официальный сайт компании YAGLA [Электронный ресурс] / [б.и.], 2014-2018. – режим доступа: https://yagla.ru/, свободный.

аудитории на основе склонности к продукту и приоритетов компании, а затем было настроено отображение этих продуктов в заголовке заголовок главной страницы, ориентированный на предпочтения той или иной группы. Теперь web- дизайнеры компании создают рекламные баннеры, ориентируясь на выделенные сегменты аудитории сайта. Результатом нововведений стал рост конверсии сайта на 40 % и на 25 % увеличился доход от посещений потребителей.

Таким образом, можно сделать вывод о том использование персонализированного подхода управления web-ресурсом оказывает положительное влияние на основной показатель интернет-сайта (конверсию), что влечет за собой увеличение объемов продаж и прибыли организации. Ключевыми факторами успеха персонализации ресурса является гиперсегментация пользователей сайта – выявление групп посетителей со схожими интересами, а также разработка и настройка необходимых подмен элементов контента. Реализация подхода позволяет значительно повысить конверсию сайта – от 40% и выше. Исследование успешных примеров реализации подхода от ведущих отраслевых компаний доказывает результативность метода и позволяет предположить, что применение предложенных в работе средств также будет способствовать увеличению конверсии на 20-100%.

ЗАКЛЮЧЕНИЕ

Для достижения цели выпускной квалификационной работы были решены все поставленные задачи:

1. Исследованы системы управления контентом сайта и методы кластеризации многомерных данных. Проведенный обзор систем управления контентом показал, что для создания и поддержки работы сайтов компании активно используют как коммерческие CMS, так и системы, находящиеся в открытом доступе. Согласно анализу, все они имеют общий функционал - создание контента сайта, управление контентом, его публикация и управление представлением контента. Также были выявлены общие для всех систем управления контентом структурные элементы - модуль навигации, модуль содержания, модуль контент менеджера, модуль авторизации, файловая система, дизайн-шаблон и стили CSS, а также База данных, где хранится необходимая для работы системы информация.

В ходе исследования методов персонализации интернет-ресурсов было выявлено, что для этого существуют два подхода – основанный на правилах и подход, основанный на алгоритмах. Оба подхода показывают свою эффективность в определенных условиях. Правила лучше работают для узкоспециализированных компаний с небольшим разбросом отличий между клиентами. Алгоритмический же по подход позволяет учитывать большой поток посетителей сайта в сочетании с большим количеством предложений компании.

В результате анализа методов кластеризации различных типов – иерархических и неиерархических (плотностных, статистических, итеративных) – было выявлено, что алгоритмы иерархического типа на начальном этапе требуют указания максимального размера кластера (максимально возможного количества элементов), что нежелательно для задачи разбиения пользователей сайта по группам. Алгоритмы неиерархические итеративные не могут быть использованы в связи с

необходимостью указания количества искомых кластеров, медленно работают при анализе больших объемов данных, а также чувствительны к шумам и могут давать недостоверные результаты в связи с этим. Неиерархические статистические алгоритмы также плохо справляются с анализом больших данных. Наиболее подходящими оказались плотностные алгоритмы, которые хорошо работают с большим объемом данных, не требуют указания количества кластеров и их размеров. Главными преимуществами подобных алгоритмов является устойчивость к данным шума и легкость их модификации и комбинирования с другими алгоритмами и процедурами. Таким образом, было принято решение разработке нового алгоритма гиперсегментации пользовательских профилей, на основании модифицированного плотностного алгоритма DBSCAN.

2. Спроектированы формальные средства формирования групп пользователей сайта. Предложен подход персонализированного управления сайтом, направленный на подбор и показ релевантного для разных групп пользователей контента страницы сайта с учетом интересов и потребностей пользователей сайта, а также изменения их предпочтений во времени. Подход предполагает сбор и обработку информации о пользователях, гиперсгментацию пользовательских профилей, разработку специализированного контента определенных блоков структуры страницы и настройку отображения необходимого содержания для той или иной группы.

В ходе исследования были выявлены источники информации о пользователе, который обращается к сайту – данные из полей заголовка http, данные из web-браузера, дополнительные данные с сайта. Было дано теоретико-множественное описание показателей, по которым могут быть сегментированы пользователи ресурса.

Спроектирована структура новой системы управления контентом сайта, которая позволяет реализовать подход персонализированного управления web-ресурсом благодаря модификации модуля контент менеджера, расширению Базы данных с целью хранения информации о

пользователях и вариантов контента структурных блоков страницы ресурса, а также подключению и настройке Базы знаний, необходимой для подбора специального для разных групп пользователей контента

3. Разработан алгоритм сегментации пользователей web-ресурса. Предложенный алгоритм гиперсегментации пользовательских профилей представляет собой комбинацию из модифицированного алгоритма DBSCAN и набора оригинальных процедур. Относительно алгоритмов кластеризации многомерных данных, рассмотренных в ходе исследования, предложенный имеет важные преимущества:

– не требует задания ожидаемого количества кластеров;

– позволяет учитывать изменяющиеся во времени потребности и интересы пользователей.

Разработан алгоритм подбора контента структурных элементов страницы сайта на основании отнесения пользователя к одной из выделенных групп. Предложенный подход сочетает алгоритмический и основанный на правилах способы реализации персонализированного подбора контента сайта под разные категории его посетителей. Причем алгоритмическая часть позволяет определить, каким именно правилом система будет руководствоваться при построении страницы. Для пользователей, которые не могут быть отнесены к уже выявленным группам, предлагается использование правил ближайшего кластера.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Aggarwal C.C. Fast Algorithms for Projected Clustering [текст]/ C.C. Aggarwal, C. Procopiuc. – In Proc. ACM SIGMOD Int. Conf. on Management of Data, Philadelphia, PA, 2010. – 12 с.

2. Agrawal, R. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications [текст]/ R. Agrawal, J. Gehrke, D. Gunopulos, P. Raghavan. – In Proc. ACM SIGMOD Int. Conf. on Management of Data, Seattle, Washington, 2014. – 9 с.

3. Ankerst M. OPTICS: Ordering Points To Identify the Clustering Structure [текст]/ M. Ankerst, Markus M. Breunig, H-P Kriegel, J. Sander.Proc. – ACM S, 2011. – 16 с.

4. Brecheisen S. Efficient DensityBased Clustering of Complex Objects [текст]/ S. Brecheisen, H-P. Kriegel, M. Pfeifle, Proc. 4th IEEE International Conference on Data Mining, 2004 – 7 с.

5. Date C. J. Databases, Types, and the Relational Model. The Third Manifesto. Addison Wesley; 3th edition [текст]/ C. J. Date, Hugh Darwen. - Addison Wesley, 2014. – 604 с.

6. Demster, A. Maximum Likelihood from Incomplete Data via the EM Algorithm [текст]/A.P. Demster, N.M. Laird, D.B. Rubin. – JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B, Vol. 39, No. 1, 2007. – 38 с.

7. Ester, M. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise [текст]/ M. Ester, H.-P. Kriegel, J. Sander, X. Xu. – In Proc. ACM SIGMOD Int. Conf. on Management of Data, Portland, OR, 2010. – 5 с.

8. Ester M. Clustering for Mining in Large Spatial Databases [текст]/ M. Ester, H.-P. Kriegel, J. Sander, X. Xu. – Issue on Data Mining, KI-Journal, ScienTec Publishing, 2008. – 8 с.

9. Fisher, D.H. Knowledge acquisition via incremental conceptual clustering [текст]/ D.H. Fisher. – Machine Learning 2, 2007. – 33 с.

10. Kailing, K. Density-Connected Subspace Clustering for High-Dimensional Data [текст]/ K. Kailing, H.-P. Kriegel, P. Kröger. – In Proceedings of the 4th SIAM International Conference on Data Mining (SDM), 2014. – 7 с.

11. Karypis, G. CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling [текст]/ G. Karypis, E.-H. Han, V. Kumar. – Journal Computer Volume 32 Issue 8. IEEE Computer Society Press Los Alamitos, CA, 2010. – 9 с.

12. Kaufman L. Clustering by means of Medoids, in Statistical Data Analysis Based on the l–Norm and Related Methods [текст]/ L. Kaufman, P.J. Rousseeuw, Y. Dodge, 2011. – 416 с.

13. MacQueen J. Some methods for classification and analysis of multivariate observations [текст]/ J. MacQueen. – In Proc. 5th Berkeley Symp. On Math. Statistics and Probability, 2007. – 6 с.

14. Nagesh, H. MAFIA: Efficient and Scalable Subspace Clustering for Very Large Data Sets [текст]/ H. Nagesh, S. Goil, A. Choudhary. – Technical Report Number CPDC-TR-9906-019, Center for Parallel and Distributed Computing, Northwestern University, 2009. – 20 с.

15. Ng, R.T. Efficient and Effective Clustering Methods for Spatial Data Mining [текст]/ R.T. Ng, J. Han. – Proc. 20th Int. Conf. on Very Large Data Bases. Morgan Kaufmann Publishers, San Francisco, CA, 2014. – 11 с.

16. Zhang T. BIRCH: An Efficient Data Clustering Method for Very Large Databases [текст]/ T. Zhang, R. Ramakrishnan, M. Linvy. – In Proc. ACM SIGMOD Int. Conf. on Management of Data. ACM Press, New York, 2016. – 11 с.

17. Бабаев А. Создание сайтов [текст]/ А. Бабаев, Н. Евдокимов, М. Бодэ. – Спб.: Питер, 2013. – 304 с.

18. Барсегян А.А. Методы и модели анализа данных: OLAP и Data Mining [текст]/ А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И Холод. – Спб.: БХВ-Петербург, 2004. – 336 с.

19. Бериков, В.С. Современные тенденции в кластерном анализе [текст]/ В.С. Бериков, Г.С. Лбов. – Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы», 2008. – 26 с.

20. Буховец А.Г. Аналитические оценки плотности в многомерных классификационных задачах [текст]/ А.Г. Буховец, Т.Я. Бирючинская. – Современные методы теории краевых задач. Материалы Воронежской весенней математической школы. – Воронеж: ВГУ, 2008. – 4 с.

21. Буховец А.Г. Использование систем итеративных функций в решении прикладных задач [текст]/ А.Г. Буховец, П.В. Москалев, Т.Я. Бирючинская. – Актуальные проблемы прикладной математики, информатики и механики: сборник трудов Международной конференции. – Воронеж: ВГУ, 2010. – 5 с.

22. Буховец А.Г. Фрактальный подход в классификационных задачах [текст]/ А.Г. Буховец, Т. Я. Бирючинская. – Экономическое прогнозирование: модели и методы: материалы VIII Международной научно-практической конференции. Воронеж, 12 мая 2012 г. – Воронеж: ВГУ, 2012. – 4 с.

23. Буховец А.Г. Фрактальный подход к анализу данных в моделях многомерной классификации [текст]/ А.Г. Буховец, Т. Я. Бирючинская. – Современная экономика: проблемы и решения, 2011.- №7(19). – 11 с.

24. Воронцов К.В. Алгоритмы кластеризации и многомерного шкалирования. Курс лекций. К.В. Воронцов. – Москва: МГУ, 2007. — 120 с.

25. Деревицкий А. А. Персонализация продаж. Как найти путь к сердцу каждого клиента [текст]/ А.А. Деревицкий. – Москва: Манн, Иванов и Фербер, 2014 – 321 с.

26. Дюк В. Data Mining. Учебный курс [текст]/ В. Дюк, А. Самойленко. – Спб: Питер,2001. – 386с.

27. Заррелла Д. Интернет-маркетинг по науке. Что, где и когда делать для получения максимального эффекта [текст]/ Д. Заррелла. - Москва: Манн, Иванов и Фербер, 2014 – 192 с.

28. Как новые медиа изменили журналистику. 2012—2016 [текст]/ А. Амзин, А. Галустян, В. Гатов, М. Кастельс, Д. Кульчицкая, Н. Лосева, М. Паркс, С. Паранько, О. Силантьева, Б. ван дер Хаак; под науч. ред. С. Балмаевой и М. Лукиной. — Екатеринбург: Гуманитарный университет, 2016. — 304 с.

29. Климова А.С. Применение методов гибридной кластеризации к анализу нефтяных скважин [текст]/ А.С. Климова, И.З. Батыршин, Н.К. Шайдуллина. – Вестник Казанского технологического университета. – Казань, 2013.- Т. 11. – 4 с.

30. Нейский И.М. Адаптивная кластеризация на основе дивизимных и итерационных методов [текст]/ И.М. Нейский. – Сборник трудов третьей международной научно-практической конференции «Информационные технологии в образовании, науке и производстве» под редакцией Ю.А. Романенко. – МО.: 2009. – 4 с.

31. Нейский И.М. Докластеризация как способ оптимизации времени анализа исходных данных [текст]/ И.М. Нейский. – Научная школа для молодых ученых «Компьютерная графика и математическое моделирование (Visual Computing)»: тезисы и доклады. – М.: 2009. – 21 с.

32. Нейский И.М. Интеграция дивизимных и итерационных методов для адаптивной кластеризации фактографических данных [текст]/ И.М. Нейский, А.Ю. Филиппович. – Труды конференции «Телематика`2009» – М.: 2009. – 3 с.

33. Нейский И.М. Методика адаптивной кластеризации фактографических данных на основе интеграции алгоритмов MST и Fuzzy C-means [текст]/ И.М. Нейский, А.Ю. Филиппович. – Известия высших учебных заведений. Проблемы полиграфии и издательского дела. – М.: Изд-во МГУП, 2009. – №3 – 14 с.

34. Нейский И.М. Сегментация клиентов брокерского обслуживания [текст]/ И.М. Нейский, А.Ю. Филиппович. – Бизнес-аналитика. Вопросы теории и практики. Использование аналитической платформы Deductor в

деятельности учебных заведений: сборник материалов межвузовской научно-практической конференции. – Рязань: Лаборатория баз данных, 2010. – 10 с.

35. Нейский, И.М. Экспериментальные исследования адаптивной кластеризации фактографических данных [текст]/ И.М. Нейский. – Материалы научной межвузовской конференции преподавателей, аспирантов, молодых ученых и специалистов «Печатные средства информации в современном обществе (к 80-летию МГУП)». Секция «Электронные средства информации в современном обществе. Сб. тезисы докладов. – М.: 2010. – 4 с.

36. Одден Ли. Продающий контент. Как связать контент-маркетинг, SEO и социальные сети в единую систему [текст]/ Ли Одден. - Москва: Манн, Иванов и Фербер, 2012. – 374 с.

37. Официальный сайт компании 1С-Битрикс [Электронный ресурс] / [б.и.], 2001 - 2018. – режим доступа: https://www.1c-bitrix.ru/, свободный.

38. Официальный сайт компании CEB [Электронный ресурс] / [б.и.], 2018. – режим доступа: https://www.cebglobal.com/, свободный.

39. Официальный сайт компании DataLife Engine SoftNews Media Group [Электронный ресурс] / Красноярск: [б.и.], 2001 - 2018. – режим доступа: https://dle-news.ru/, свободный.

40. Официальный сайт компании DATA insight [Электронный ресурс] / Красноярск: [б.и.], 2018. – режим доступа: http://www.datainsight.ru/, свободный.

41. Официальный сайт Drupal на русском языке [Электронный ресурс] / [б.и.], 2018. – режим доступа: https://drupal.ru/, свободный.

42. Официальный сайт компании «iTrack» [Электронный ресурс] / Москва: [б.и.], 2004 - 2018. – режим доступа: https://itrack.ru/research/cmsrate/#!cms-overall-tab, свободный.

43. Официальный сайт компании MODX creative freedom [Электронный ресурс] / [б.и.], 2005 - 2018. – режим доступа: https://modx.ru/, свободный.

44. Официальный сайт Joomla! [Электронный ресурс] / [б.и.], 2006 - 2018. – режим доступа: http://joomla.ru/, свободный.

45. Официальный сайт WordPress.org на русском языке [Электронный ресурс] / [б.и.], 2018. – режим доступа: https://ru.wordpress.org/, свободный.

46. Официальный сайт компании YAGLA [Электронный ресурс] / [б.и.], 2014-2018. – режим доступа: https://yagla.ru/, свободный.

47. Официальный сайт компании «Альянс Форест» [Электронный ресурс] / [б.и.], 2010-2018. – режим доступа: http://alforest.ru/, свободный.

48. Райзин Дж. Вэн. Классификация и кластер [текст]/ Дж. Вэн Райзин. – Москва: Мир, 2008. – 390 с.

49. Роуз Р. Управление контент-маркетингом. Практическое руководство по созданию лояльной аудитории для вашего бизнеса [текст]/ Р. Роуз, Д. Пулицци. – Москва: Манн, Иванов и Фербер, 2014. – 229 с.

50. Симчера В.М. Методы многомерного анализа статистических данных [текст]/ В.М. Симчера. – Москва: Финансы и статистика, 2008. – 398 с.

51. Стелзнер М. Контент-маркетинг. Новые методы привлечения клиентов в эпоху Интернета [текст]/ М. Стелзнер. Москва: Манн, Иванов и Фербер, 2012 – 134 с.

52. Тюха А.С. Проектирование базы знаний системы управления контентом сайта автомагазина на основе персонализации [Электронный ресурс] / А.С. Тюха, Р.Г. Асадуллаев. – международный научно-практический журнал «Теория и практика современной науки». Выпуск № 5(35) (МАЙ, 2018), режим доступа: http://modern-j.ru/matematika__informatika_i_inzheneriya__5_35_/, свободный.

53. Тюха А.С. Проектирование структуры CMS, реализующей функцию персонализации контента [Электронный ресурс] / А.С. Тюха, Р.Г. Асадуллаев. – международный научно-практический журнал «Форум молодых ученых». Выпуск № 6(22) (июнь, 2018), режим доступа: http://forum-nauka.ru/_6_22__iyun_2018/, свободный.

54. Тюха А.С. Разработка алгоритма подбора персонализированного контента сайта [Электронный ресурс] / А.С. Тюха, Р.Г. Асадуллаев. – международное научное издание «Мировая наука». Выпуск № 6(15) (июнь, 2018), режим доступа: http://science-j.com/_6_15__iyun_2018/, свободный.

55. Тюха А.С. Разработка алгоритма гиперсегментации пользователей интернет-ресурса [текст] / А.С. Тюха, Р.Г. Асадуллаев. – научно-методический журнал Academy. Выпуск № 6(33) Том 2, 2018, 9 – 12 с.

56. Тюха А.С. Разработка подхода персонализированного управления контентом сайта [Электронный ресурс] / А.С. Тюха, Р.Г. Асадуллаев. – международный научно-практический журнал «Теория и практика современной науки». Выпуск № 6(36) (июнь, 2018), режим доступа: http://modern-j.ru/matematika__informatika_i_inzheneriya__6_36_/, свободный.

57. Уиллиамс У. Т. Методы иерархической классификации. Статистические методы для ЭВМ Под ред. М. Б. Малютов. [текст]/ У.Т. Уиллиамс, Д.Н. Ланс. — М.: Наука, 2016. — 425 с.

58. Царев А.Г. Исследование однопараметрических индикаторов заинтересованности пользователей веб-сайта [текст]/ А.Г. Царев, Т.Н. Царева. – Инновации в условиях развития информационно-коммуникационных технологий: Материалы научно-практической конференции - М.: МИЭМ, 2009. – 4 с.

59. Царев А.Г. Массовая рекомендательная система для веб-сайтов на основе SAAS-технологии [текст]/ А.Г. Царев. – Труды II международной научно-практической интернет-конференции. Под ред. Г.К. Сафаралиева, А.Н. Андреева, В.А. Казакова. – Пенза: Издательство Пензенского филиала РГУИТП, 2010. – 5 с.

60. Царев А.Г. Метод персонализации веб-сайта на основе анализа постоянных и текущих потребностей конечного пользователя. Труды II международной научно-практической интернет-конференции [текст]/ А.Г. Царев. Под ред. Г.К. Сафаралиева, А.Н. Андреева, В.А. Казакова. – Пенза: Издательство Пензенского филиала РГУИТП, 2010. – 6 с.

61. Царев А.Г. Многокритериальная оптимизация в задаче вычисления релевантности страниц веб-сайта. Естественные и технические науки [текст]/ А.Г. Царев. – Естественные и технические науки №4 (48). – Москва: «Издательство «Спутник+», 2010. – 3 с.

62. Царев А.Г. Модель индикатора предпочтений конечного пользователя веб-сайта на основе многокритериальной комплексной оценки альтернатив [текст]/ А.Г. Царев. – Мониторинг. Наука и технологии. №3, 2010. – 7 с.

63. Царев А.Г. Модель персонализации сайта на основе анализа постоянных потребностей конечного пользователя [текст]/ А.Г. Царев, В.Г. Домрачев, И.В. Ракитянская. – Новые информационные технологии и менеджмент качества. Материалы международной научной конференции/Редкол.: А.Н. Тихонов (пред.) и др.; ФГУ ГНИЙ ИТТ «Информика». – М.: ООО «Арт-Флэш», 2010. – 5 с.

64. Царев А.Г. О сборе пользовательских данных в системе персонализации Интернет-магазина. [текст]/ А.Г. Царев. – Вестник Московского государственного университета леса - Лесной вестник, 2009. - №3(66). – 5 с.