Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 51-77

Рудая
Елена Игоревна

# Аналитика в электронной коммерции. Иерархическая кластеризация как средство анализа поведения клиентов

**АВТОРЕФЕРАТ**
на соискание степени магистра технических наук
по специальности 1-40 80 04 «Математическое моделирование, численные методы и комплексы программ»

Научный руководитель
Липницкий Валерий Антонович
доктор технических наук, профессор



Минск 2016

## КРАТКОЕ ВВЕДЕНИЕ

В мире огромными темпами растет количество пользователей Internet и, как следствие, количество потенциальных «электронных» покупателей.

Электронные магазины существенно уменьшают издержки производителя, сэкономив на содержании обычного магазина, расширяют рынки сбыта, так же дают безграничное множество возможностей для покупателя – приобретать любой товар в любое время в любой стране, в любом городе, в любое время суток, в любое время года. Это дает электронным магазинам неоспариваемое преимущество перед обычными магазинами. Таким образом следует сделать вывод, что имеет место быть развитие нового направления – анализ онлайн-розницы.

Информация о поведении посетителей на страницах сайта представляет особую ценность для владельца любого ресурса. Обычно можно только предполагать о возможных действиях клиента на сайте: что его заинтересовало или показалось непонятным, что осталось без внимания и что попало в его поле зрения, какие затруднения возникли у него, что заставило покинуть сайт и т.д.

Применение инструментов аналитики является перспективным направлением операционного управления бизнесом в самых разных сегментах рынка. Но в бизнесе, соприкасающемся с потребительским рынком, эти технологии дают наиболее заметный эффект в сравнении с другими отраслями. Особенно результативно использование бизнес-аналитики в онлайн-рознице. А именно очень важно знать поведение покупателя на сайте для того, чтобы понять правильно ли развивается сайт и удовлетворен ли пользователь тем, что нашел на сайте.

## ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

### Цель и задачи исследования

*Цель* диссертационной работы можно сформулировать как улучшение качества обслуживания, возможность привлечения новых, а также удержание старых клиентов за счет получения оперативных данных и проведения их анализа.

Для достижения поставленной цели необходимо решить следующие задачи:

– доказать важность аналитики в электронной торговле;

– изучить основные особенности систем анализа поведения пользователей на сайте;

– определить связь стратегического планирования и анализа пользовательского поведения;

– определить проблему, связанную с недостаточной персонализацией канала B2C.

– предоставить инструментарий, который позволит сконструировать систему для анализа поведения пользователей.

– разработать систему с помощью выбранной платформы

*Объектом* исследования являются статистические данные о посетителях Интернет ресурсов.

*Предметом* исследования является поведение посетителей сайта.

Основной *гипотезой*, положенной в основу диссертационной работы, является возможность извлечь нетривиальные и практически полезные знания из набора данных о посетителях различных интернет-магазинов. Эти знания впоследствии могут стать базой для принятия важных стратегических решений, а также персонализировать канал взаимодействия Интренет ресурса и посетителя.

### Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя В.А. Липницкого заключается в формулировке целей и задач исследования, содействии в подборе и анализе научной литературы по исследуемому вопросу, консультации и обсуждению возникавших проблем.

**Апробация результатов диссертации**

Основные положения диссертационной работы докладывались и обсуждались на Международной научно-практической конференции «Молодёжный форум: технические и математические науки» (Воронеж, Россия, 9-12 ноября 2015 года) и на 51-ой научной конференции аспирантов, магистрантов и студентов БГУИР (Минск, Республика Беларусь)

**Опубликованность результатов диссертации**

По теме диссертации опубликована 1 статья в сборнике трудов и материалов международной конференции.

**Структура и объем диссертации**

Диссертация состоит из введения, общей характеристики работы, трех глав, заключения, списка использованных источников, списка публикаций автора и приложения.

Общий объем работы составляет 84 страницы, которые включают 44 рисунка, 7 таблиц, список использованных источников из 16 наименований и 1 приложение на 7 страницах.

## ОСНОВНОЕ СОДЕРЖАНИЕ

Во **введении** определена область и указаны основные направления исследования, показана актуальность темы диссертационной работы, дана краткая характеристика исследуемых вопросов, обозначена практическая ценность работы.

**Первая глава** посвящена исследованию стратегического планирования развития интернет-сервисов. Здесь будут представлены предметная и теоретическая база, а именно - что такое электронная коммерция и ее подраздел электронная торговля, освещены основные тенденции в аналитике для электронной торговли и ее связь с планированием. Также представлен метод иерархической кластеризации как один из инструментариев для анализа аудитории ресурса.

**Вторая глава** посвящена обзору технологий. Также изложено обоснование важности персонализации канала взаимодействия бизнеса и клиента и способ, как это можно реализовать.

**Третья глава** - это постановка задачи и непосредственно описание системы: модели представления системы, информационная модель, а также руководство пользователя.

Существуют различные системы, использующие коды, скрипты отслеживания, размещающиеся на страницах сайта и позволяющие накапливать информацию о клиенте, после чего она может быть проанализирована для получения дополнительных знаний, эффективным методом получения которых является проведение кластерного анализа. Наиболее часто в социальных задачах используются иерархические агломеративные методы. Эти методы просматривают матрицу сходства и последовательно объединяют схожие объекты. Алгоритм иерархической кластеризации строит иерархию групп, объединяя на каждом шаге две самые похожие группы. В начале каждая группа состоит из одного элемента. На каждой итерации вычисляются попарные расстояния между группами, и группы, оказавшиеся самыми близкими, объединяются в новую группу. Так повторяется до тех пор, пока не останется всего одна группа [1]. Эта процедура изображена на рисунке 1:

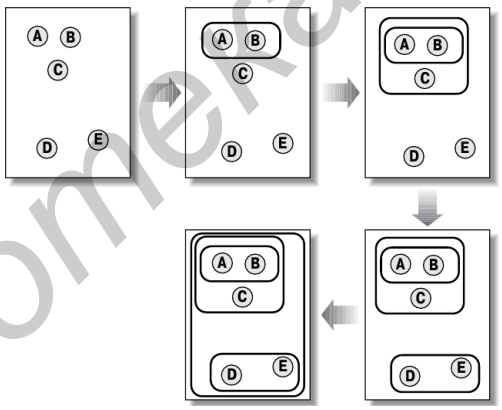


**Рисунок 1 – Иерархическая кластеризация в действии**

Для анализа информации о покупателях интернет-магазина был использован метод Уорда, особенностью которого является стремление к разбиению на кластеры схожего размера. Целевой функцией является сумма квадратов отклонений. Данный метод предполагает, что первоначально каждый кластер состоит из одного объекта. Сначала объединяются два ближайших кластера. Для них определяются средние значения каждого признака и рассчитывается сумма квадратов отклонений:

$$V_l = \sum_i \sum_j (x_{ij} - x_{jl})^2 \qquad (1)$$

где $l$ - номер кластера, $i$ - номер объекта ($i = 1,2, \ldots, n_l$), $n_l$ - количество объектов в $l$ - том кластере, $j$ - номер признака ($j = 1,2, \ldots, k$), $k$ - количество признаков, характеризующих каждый объект.

В дальнейшем объединяются те объекты или кластеры, которые дают наименьшее приращение величины $V_l$.

К числу характеристик клиентов, значимых для интернет - торговли на основе экспертных оценок были отнесены следующие:

– количество страниц, которые пользователь посетил за сеанс;

– была ли совершена покупка;

– совершенные покупки;

– тип посещения сайта (социальный, поисковой, реферальный, прямой);

– используемый девайс при посещении сайта;

– продолжительность сессии;

– возраст и пол клиента.

Для хранения этой информации использовано хранилище данных, созданное с использованием инструментов MySql 5.1 и OLAP сервера Mondrian 3.5.0.

Так как данные имеют смешанный характер (количественные, порядковые и дихотомические) для формирования матрицы сходства требуется применение коэффициента Гауэра:

$$S_{ij} = \frac{\sum_{k=1}^{p} S_{ijk}}{\sum_{k=1}^{p} W_{ijk}} \qquad (2)$$

где $W_{ijk}$ - весовая переменная, принимающая значение единицы, если сравнение объектов по признаку $k$ следует учитывать, и ноль - в противном случае; $S_{ijk}$ - «вклад» в сходство объектов, зависящий от того, учитывается ли признак при сравнении объектов $i$ и $j$.

После построения матрицы сходства используется Open Source фреймворк MultiDendrograms для построения дендрограммы – графа, представляющего результаты иерархической кластеризации:

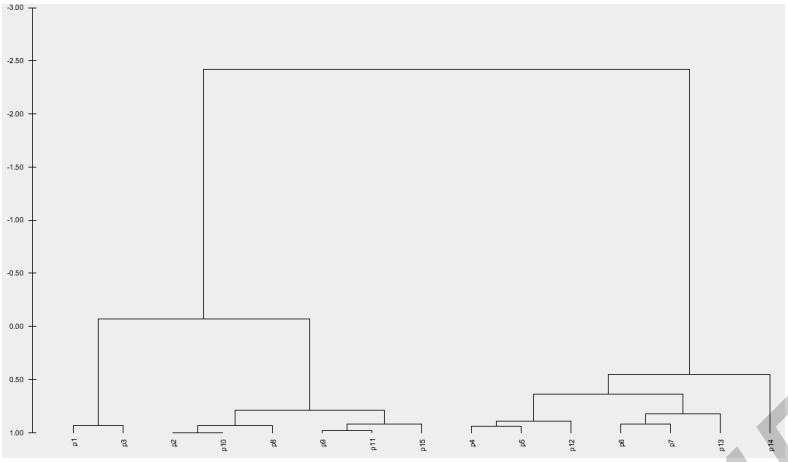


**Рисунок 2 – Дендрограмма, полученная в результате анализа характеристик 15-ти клиентов**

## ЗАКЛЮЧЕНИЕ

### Основные научные результаты диссертации

Разработана распределенная система, работающая с базой данных и хранилищем данных и реализующая следующие возможности:

– анализ пользовательского поведения на Интернет ресурсе;

– построение графиков по этим данным;

– конструирование специальных правил;

– осуществление иерархической кластеризации клиентов по набору признаков.

Все задачи, поставленные перед началом моделирования, выполнены:

– Изучены основные особенности систем анализа поведения пользователей на сайте.

– В качестве математического аппарата была выбрана иерархическая кластеризация.

– Разработана информационная модель системы, состоящая из базы данных, которая хранит информацию о настройках аналитика, и хранилища данных.

– Был выбран оптимальный инструментарий для разработки.

– Разработан программный продукт, позволяющий предоставить информацию о том, какие именно действия совершены потенциальным клиентом на сайте, какие товары были просмотрены, какова длительность сессии взаимодействия ресурса и клиента, и т. д.

**Рекомендации по практическому использованию результатов**

1. Полученные результаты формируют теоретическую и практическую базу для дальнейшего исследования в области изучения алгоритмов иерархических кластеризаций, а также дальнейшего развития программного продукта.

2. Разработанные методы и алгоритмы анализа могут применяться в компаниях, которые занимаются электронной торговлей либо планируют внедрить ее в будущем.

3. Результаты работы могут использоваться при подготовке бизнес-аналитиков для моделирования и сбора информации на основе предоставленной выборки о посетителях Интернет ресурсов.

## СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Рудая, Е.И. Метод иерархической кластеризации как средство анализа поведения клиентов интернет-магазинов / Е.И. Рудая // Международная научно-практическая конференция "Молодёжный форум: технические и математические науки – 2015. – № 12. – с. 158.