Михалькевич Илья Сергеевич. Инструменты и методы анализа слабоструктурированных данных в оптимизации маркетинговых коммуникаций: диссертация ... кандидата Экономических наук: 08.00.13 / Михалькевич Илья Сергеевич;[Место защиты: ФГОБУ ВО Финансовый университет при Правительстве Российской Федерации], 2018.- 141 с.

**Введение к работе**

**Актуальность темы исследования.** Одним из ключевых факторов роста любой компании является повышение эффективности управления внутренними ресурсами. Одним из таких ресурсов является клиентская база компании, потенциал которой в значительной степени реализуется за счёт маркетинговых коммуникаций. Высокие стандарты качества, предъявляемые к клиентской базе, а также знания, полученные в результате её глубокого анализа, являются необходимыми условиями эффективного управления маркетинговыми коммуникациями.

В условиях наличия больших массивов разнородных данных у организации их анализ значительно затрудняется. Существующие механизмы интеграции систем не позволяют полностью учесть специфику бизнес-процессов организации: это чревато потерей важной информации, необходимой для принятия решений. Кроме этого, в силу особенностей процесса накопления данных назревает потребность в использовании специфических средств, методов анализа и моделирования.

Анализ опыта российских и зарубежных компаний показывает, что применение маркетинговых коммуникаций в бизнесе позволяет значительно повысить лояльность существующих клиентов, привлечь большое количество новых клиентов и, таким образом, увеличить продажи. Использование такого дорогостоящего инструмента маркетинговых коммуникаций, как телефонный звонок, требует оптимизации использования этого канала коммуникации.

Лизинг автомобилей является динамично развивающейся отраслью на  
российском рынке финансовых услуг, однако, применение маркетинговых

коммуникаций в лизинге имеет весьма ограниченный характер. Всё больше компаний, как российских, так и зарубежных понимают важность комплексного подхода к повышению качества клиентской базы и повышению эффективности маркетинговых коммуникаций. Поэтому разработка соответствующих инструментов и методов является актуальным направлением научного исследования.

**Степень разработанности темы исследования.** С целью совершенствования алгоритмов сопоставления информации о клиентах в работе изучены методы дедупликации и извлечения текста на основе нечёткого поиска, так как именно в них затрагиваются наиболее сложные проблемы работы с данными в значительной степени

4 субъективного характера. Фундаментальные исследования в данном научном

направлении освещены в работах Р. Хемминга, Ф. Дамерау и В.И. Левенштейна. В этих

работах предлагается ввести метрику для оценки расстояния между строковыми

последовательностями. Дальнейшие работы, расширяющие область применения

предложенных метрик, были посвящены, в основном, так называемым оффлайн

алгоритмам нечёткого поиска, таких как нечёткий поиск с индексацией и алгоритм

расширения выборки. В работах Л.М. Бойцова вводится понятие сигнатуры,

отражающей наличие тех или иных символов алфавита в строке, затем предлагается

алгоритм индексирования хеш-таблиц, построенных на сигнатуре. Работы Е. Укконена

посвящены нечёткому поиску методом N-грамм. Достоинством этого подхода является

простота реализации и хорошая производительность алгоритмов, недостатком – то, что

очень близкие друг к другу строковые последовательности могут оказаться

незамеченными. Помимо перечисленных методов существует множество интересных,

но малоэффективных вариантов, таких как, фонетические алгоритмы, один из которых

впервые предложен Р. Расселом, или адаптация префиксных деревьев к задачам

нечёткого поиска Т.Г. Меррета.

Из представленных метрических алгоритмов нечёткого поиска наиболее  
адекватные результаты показывают алгоритмы на основе метрики

Дамерау-Левенштейна. Однако, понятие метрики Дамерау-Левенштейна можно расширить с целью реализации тонкой настройки нечёткого поиска, а также поиска последовательностей заданного типа.

Оценку качества дедупликации данных часто проводят либо с точки зрения скорости работы алгоритмов дедупликации, либо путём непосредственного применения метрик строкового расстояния. Однако, остаётся нереализованной потребность в оценке качества дедупликации независимо от выбранных метрик и скорости работы алгоритмов.

Основы теории оптимизации поиска и сортировки данных разработаны в фундаментальных работах Д. Кнута, Э. Дейкстры и развиты в исследованиях Н. Вирта, Г. Гарсией-Молины, Дж. Ульмана, Дж. Видом. Вместе с тем почти не представлено работ по оптимизации поиска по составному ключу.

Другим инструментом, позволяющим повысить эффективность маркетинговых коммуникаций, является анализ данных, машинное обучение и математическое

5 моделирование отклика на маркетинговые коммуникации. В области машинного

обучения разработано огромное количество различных методик, таких как методы

кластеризации (графовые: связных компонент, кратчайшего незамкнутого пути, ForEl;

статистические: EM-алгоритм, k-средних; иерархические), решающие деревья,

регрессия, нейронные сети. Большинство методов рассмотрены в работах

К.В. Воронцова, описаны их достоинства и недостатки. Различные методы визуализации

многомерных данных изучены в работах К. Пирсона, К. Карунена, М. Лоэва,

Т. Кохонена, А. Зиновьева и других учёных. Тем не менее, вопросы отображения

многомерных данных, имеющих признаки с высоким коэффициентом эксцесса требуют

отдельного рассмотрения.

Популярным методом в анализе маркетинговых коммуникаций является RFM-анализ (recency – давность, frequency – частота, monetary – затраченные деньги), теоретические основы которого развиты в работах П. Фадера, Б. Харди и К. Ли. Эти зарубежные исследования описывают наиболее общие характеристики в поведении клиентов. Вместе с тем существуют перспективы уточнения набора значимых признаков с точки зрения оптимизации маркетинговых коммуникаций в условиях современного российского бизнеса.

Существующие подходы к моделированию конверсии маркетинговых

коммуникаций часто не отражают специфики прикладных областей, где необходимо применять перечисленные методы. Это, в свою очередь, требует развития математических моделей отклика на маркетинговые коммуникации, учитывающих особенности клиентской базы лизинговой компании, а также методов нечёткого поиска информации в слабоструктурированных или неструктурированных текстовых данных.

**Объект** **исследования.** Объектом исследования является компания,

специализирующаяся на лизинге автомобилей для юридических лиц.

**Предмет исследования.** Предметом исследования является моделирование отклика на маркетинговые кампании, построение портрета клиента и его оценка, а также методы обработки данных, включая их очистку и стандартизацию.

**Область исследования.** Тема работы соответствует направлениям исследований, описанным в разделах 1.4. «Разработка и исследование моделей и математических методов анализа микроэкономических процессов и систем: отраслей народного хозяйства, фирм и предприятий, домашних хозяйств, рынков, механизмов

6  
формирования спроса и потребления, способов количественной оценки

предпринимательских рисков и обоснования инвестиционных решений»; 2.6. «Развитие

теоретических основ методологии и инструментария проектирования, разработки и

сопровождения информационных систем субъектов экономической деятельности:

методы формализованного представления предметной области, программные средства,

базы данных, корпоративные хранилища данных, базы знаний, коммуникационные

технологии» Паспорта научной специальности" 08.00.13 – «Математические и

инструментальные методы экономики» (экономические науки).

**Целью** исследования является разработка инструментов и методов анализа слабоструктурированных данных клиентской базы лизинговой компании и построение модели отклика клиентов на маркетинговые коммуникации.

Для достижения цели исследования были сформулированы и решены следующие **задачи**:

1. Применение системного подхода к таргетированию маркетинговых коммуникаций посредством обогащения данных о клиентах за счёт внутренних ресурсов клиентской базы.
2. Прогнозирование отклика клиентов на маркетинговые коммуникации, позволяющее значительно повысить лояльность клиентов и увеличить продажи.
3. Адаптация метода нечёткого поиска для обнаружения данных заданного типа в строковой последовательности.
4. Реализация алгоритма дедупликации данных на основе методов нечёткого сравнения символьных строк.
5. Разработка метода предварительного анализа признаков для включения в регрессионную модель отклика на маркетинговые коммуникации.
6. Выявление значимых признаков клиентов на основе анализа их влияния на результат маркетинговых коммуникаций с помощью разработанного метода.
7. Выявление значимых признаков клиентов с использованием нелинейных методов отображения данных с последующей кластеризацией.
8. Разработка алгоритма нормирования признаков для включения в регрессионную модель с учётом их распределения в выборке.
9. Построение регрессионной модели отклика на маркетинговые коммуникации с использованием полученных признаков.

7  
10. Реализация полученных методов и моделей в программном комплексе и

внедрение в производственный процесс.

**Методология и методы исследования.** Основу работы составили теоретические и методические разработки по анализу данных в сфере маркетинга, математической статистики, эконометрического моделирования, теории информации и компьютерной лингвистики, теории баз данных, проектирования информационных систем.

В ходе исследования применялись методы системного анализа, экспертных оценок, статистического и сравнительного анализа, а также табличные и графические приёмы визуализации данных.

**Информационной базой исследования** послужили данные российских и зарубежных аналитических агентств, информация из открытых источников о внедрении систем управления клиентской базой и оптимизации маркетинговых коммуникаций, научные труды российских и зарубежных исследователей, данные российской компании «Европлан», специализирующейся на лизинге автомобилей и спецтехники.

**Научная новизна исследования** заключается в модификации и развитии методов работы с клиентской базой компании, построении новой модели отклика клиентов на маркетинговые коммуникации. Основные результаты, содержащие элементы научной новизны:

1. Предложено и обосновано использование преобразования метрики Дамерау-Левенштейна для поиска дублирующихся текстовых записей и вычленения текстовых данных заданного типа.
2. Предложены методы оценки качества дедупликации и извлечения данных.
3. Предложено и обосновано использование модификации составного ключа с пустыми значениями в базе данных для возможности индексированного поиска новых объектов.
4. Предложен и обоснован метод нормализации многомерных признаков для корректировки отображения многомерных данных на нелинейное многообразие, вложенное в пространство большей размерности.
5. Выявлены новые значимые факторы, позволяющие уточнить оценку вероятности отклика клиента на маркетинговые коммуникации.
6. Построена регрессионная модель для оценки вероятности отклика клиента на маркетинговые коммуникации с использованием выявленных факторов.

8 **Положения, выносимые на защиту**:

1. Предложено и обосновано преобразование метрики Дамерау-Левенштейна, что расширяет возможности данной метрики при поиске дублирующихся записей и вычленении текстовых данных заданного типа (С. 59-62).
2. Разработаны методы оценки качества дедупликации и извлечения данных о клиентах на основе дополнительной информации, имеющейся в базе данных, а также полученной в результате эксперимента (С. 63-65).
3. Предложена модификация составного ключа с пустыми значениями в базе данных и обосновано её использование при индексированном поиске новых объектов (С. 69-72).
4. Предложен метод сегментации клиентской базы с помощью отображения многомерных данных на нелинейное многообразие, вложенное в пространство большей размерности, с дальнейшей кластеризацией этих данных (С. 83-85).
5. На основе анализа результатов взаимодействия компании с клиентами выявлены новые значимые характеристики клиентов, позволяющие уточнить оценку вероятности их отклика на маркетинговые коммуникации (С. 74-76, 78-79, 82, 86).
6. Построена регрессионная модель оценки вероятности отклика клиента на маркетинговые коммуникации, учитывающая влияние выявленных значимых характеристик клиента (С. 87-96).
7. Разработан программный комплекс автоматизации исходящих маркетинговых коммуникаций, позволивший существенно увеличить доход лизинговой компании (С. 97-122).

**Теоретическая значимость** представленных в работе результатов состоит в развитии методов оптимизации маркетинговых коммуникаций, математических моделей отклика клиентов на маркетинговые коммуникации и инструментов обработки клиентской базы. Основные положения и выводы диссертации дополняют существующие методы анализа слабоструктурированных данных и прогнозирования поведения клиентов. Материалы и обобщения, полученные в диссертации, могут служить теоретической основой для дальнейшего развития областей исследования, связанных с управлением клиентской базой компании и оптимизацией маркетинговых коммуникаций.

9 **Практическая значимость исследования.** Разработанные в диссертации методы

и модели ориентированы на широкое использование в крупных и средних организациях,

имеющих значительную клиентскую базу. Применение разработанных методов и

моделей обеспечивает экономический эффект, выражающийся в виде увеличения

продаж, повышения лояльности клиентов и наращивания клиентского портфеля.

Самостоятельную практическую значимость имеют:

1. Повышение эффективности маркетинговых коммуникаций за счёт увеличения конверсии при сохранении среднего чека, а, следовательно, увеличение прибыли и окупаемости инвестиций в маркетинговые коммуникации.
2. Предотвращение оттока клиентов в результате соблюдения контактной политики.
3. Получение представления об основных факторах (времени, прошедшего с последней покупки, количестве и стоимости покупок, источнике обращения клиента, и др.) и характере их влияния на конверсию в маркетинговых коммуникациях лизинговой компании.
4. Внедрение программного комплекса, позволившего:

повысить скорость формирования списка клиентов для проведения маркетинговых коммуникаций за счёт очистки и структуризации данных;

устранить неопределённость, обеспечить полноту, точность и согласованность данных в управленческой отчётности благодаря преобразованию исходных данных адаптированными методами.

5. Сокращение трудовых затрат на создание управленческой отчётности о  
результатах маркетинговых коммуникаций.

Для достижения полученных практических результатов были применены:

Адаптированный метод нечёткого поиска для обнаружения данных заданного типа в строковой последовательности.

Алгоритм нормирования признаков для включения в регрессионную модель с учётом их распределения в выборке.

Метод оценки качества дедупликации и извлечения данных.

Регрессионная модель для оценки вероятности отклика клиента на маркетинговые коммуникации, учитывающая влияние выявленных значимых характеристик клиента.

10 Результаты исследования нашли практическое применение в маркетинговой

деятельности лизинговой компании ПАО «Европлан» и используются в учебном

процессе ФГОБУ ВО «Финансовый Университет при Правительстве Российской

Федерации» в преподавании дисциплины «Технологии интеллектуального анализа

данных».

**Степень достоверности, апробация и внедрение результатов исследования.**

Достоверность полученных результатов была подтверждена большим фактическим материалом, результаты исследования согласуются с фундаментальными положениями экономической теории. Методика проведения расчётов соответствует критериям, предъявляемым к научному подходу, и позволяет получить объективные результаты. Разработка программных средств велась с использованием современных платформ и языков программирования.

Результаты исследования обсуждались и получили положительные отзывы на  
межвузовских и международных научно-практических конференциях: на межвузовском  
круглом столе «Молодые учёные о проблемах отечественной науки» (Москва,  
Финансовый университет, 21 апреля 2014 г.), на научной конференции «Научные  
достижения молодых исследователей» (Москва, Финансовый университет, 29 марта  
2014 г.), на IV международном конкурсе научных работ аспирантов и студентов  
(Москва, Финансовый университет, 28 апреля 2015 г.), на московской

научно-практической конференции «Студенческая наука» (Москва, Финансовый университет, 30 ноября 2015 г.), на V международном конкурсе научных работ аспирантов и студентов (Москва, Финансовый университет, 4 апреля 2016 г.), на международной научно-практической конференции «Актуальные проблемы развития современной науки и образования» (Москва, Научное издательство «Ар-Консалт», 30 апреля 2016 г.).

Материалы диссертации используются в практической деятельности Управления  
исследований ПАО «Европлан». По материалам исследования внедрен

программно-аппаратный комплекс, предназначенный для решения задач маркетинговых коммуникаций, в том числе: объединения данных о клиенте компании из внутренних систем и внешних источников в единую сущность, централизованного хранения данных об истории взаимодействия с клиентом, внедрения математических моделей поведения клиента для повышения эффективности коммуникаций, формирования аналитической

11 отчетности. Выводы и основные положения диссертации дают эффект в виде

увеличения конверсии маркетинговых коммуникаций с 6,6 до 12,3% и, таким образом,

получения дополнительной прибыли в размере 9 000 тыс. руб. ежегодно.

Материалы диссертации используются в учебном процессе кафедрой «Прикладная информатика» ФГОБУ ВО «Финансовый Университет при Правительстве Российской Федерации» в преподавании учебной дисциплины «Технологии интеллектуального анализа данных».

Результаты внедрения подтверждены соответствующими документами.

**Публикации по теме диссертации.** По теме диссертации опубликовано 6 работ общим объемом 3,39 п.л. (весь объем авторский), в том числе 4 работы авторским объемом 2,5 п.л. опубликованы в рецензируемых научных изданиях, определенных ВАК при Минобрнауки России.

**Структура диссертации** определена целью, задачами и логикой исследования и состоит из введения, пяти глав, заключения, списка литературы из 161 источника и 1 приложения. Работа изложена на 141 странице и содержит 52 рисунка, 26 таблиц, 80 формул.