Министерство образования и науки Российской Федерации
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ
Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра дискретной математики и
информационных технологий

Информационные технологии Data mining в маркетинговых исследованиях

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 5 курса 521 группы
направления 09.03.01 «Информатика и вычислительная техника»
факультета компьютерных наук и информационных технологий
Калдузова Дмитрия Викторовича

Научный руководитель
к. ф.-м.н., д. э. н., профессор _____ Л. В. Кальянов
подпись, дата

Зав. кафедрой
к. ф.-м.н., доцент _____ Л.Б. Тяпаев
подпись, дата

Саратов 2017

Введение. Исследования в области маркетинга подразумевают сбор необработанных данных в большом количестве, в связи с чем возникает дополнительная потребность в использовании методов Data mining для последующей обработки полученных данных. Data mining можно перевести как добыча данных, раскопка данных, извлечение знаний, извлечение информации, интеллектуальный анализ данных, средства поиска закономерностей, анализ шаблонов, раскопка знаний в базах данных. Задачи, решаемые с помощью методов Data Mining: классификация, регрессия, кластеризация, ассоциация, последовательные шаблоны, анализ отклонений.

Целью бакалаврской работы является – Анализ и выявление зависимостей среди отзывов клиентов об отелях с помощью информационной технологий Data mining.

Для достижения цели были выделены следующие подзадачи:

- Выявление источников, содержащих доступную и подходящую информацию для анализа.
- Выявление способов доступа к подобной информации.
- Обоснование выбранного источника информации и загрузка необходимых данных для маркетингового исследования.
- Выбор технологии для выявления закономерностей в текстах.
- Реализация технологии в инструментальной системе RapidMiner.
- Расчёт вероятностей возникновения слов и связанной пары слов среди всего множества текстов, а также для каждого текста по отдельности.
- Реализация метода кластеризации.
- Представление полученных результатов в табличном и графическом виде.
- Анализ полученных результатов.

В первой главе рассмотрены поиск и добыча маркетинговой информации в WWW.

Во второй главе приведено обоснование выбранного источника данных.

В третьей главе приведено понятие Text mining, затрагивающее интеллектуальный анализ текстов.

В четвертой главе приведено описание использованного программного обеспечения, а также необходимого пакета расширений.

В пятой главе приведена практическая реализация метода ассоциации.

В шестой главе приведена практическая реализация метода кластеризации.

В седьмой главе рассмотрены результаты выполнения операторов. Выявлены закономерности с помощью метода ассоциации, рассчитаны Support и Confidence. Входные данные были распределены на пять кластеров с помощью метода кластеризации.

1 Поиск подходящей маркетинговой информации в WWW с возможностью выгрузки данных. В сети существуют открытые репозитории, поисковой запрос: «data set repository» способен предоставить их список. Ссылка на популярный каталог репозиториев доступна в приложении [3].

Данные также можно получить через интерфейс API. API - программный интерфейс приложения, интерфейс прикладного программирования. [4].

2 Обоснование выбранного источника информации и загрузка необходимых данных. Данные необходимые для анализа, были получены с популярного веб-портала о путешествиях – TripAdvisor.com в виде архива [5]. Архив содержал 1850 текстовых файлов, где каждый файл представлял из себя отель, а его содержимое - отзывы к этому отелю. На других репозиториях подходящие данные либо отсутствовали, либо были слишком сильно «урезаны».

3 Интеллектуальный анализ текстов как элемент маркетингового исследования. Интеллектуальный анализ текстов (Text mining) – направление в искусственном интеллекте, целью которого является получение информации из коллекций текстовых документов, основываясь на применении

эффективных в практическом плане методов машинного обучения и обработки естественного языка. «Интеллектуальный анализ текстов» перекликается с понятием «интеллектуальный анализ данных» (Data mining) [6].

Обнаружение нетривиальных данных включает в себя анализ извлекаемой информации посредством обработки текстовых файлов с последующим формированием новых фактов и гипотез. Интеллектуальный анализ может включать в себя категоризацию текстов, извлечение признаков, кластеризацию, анализ тенденций, объединение и визуализацию ассоциаций.

4 Необходимое программное обеспечение и дополнения. Для проведения анализа использовалось программное обеспечение RapidMiner [7], а также дополнительные расширения: Text Processing, Web Mining, Wordnet Extention.

5 Реализация метода ассоциации. Метод был реализован при помощи операторов: Process Documents from Files, Text to Nominal, Numerical to Binominal, FP-Growth, Create Association Rules. Подпроцесс «Process Documents from Files» включал в себя дополнительные операторы: Tokenize Non-letters, Tokenize Linguistic, Filter Stopwords, Filter Tokenz, Stem (Porter), Transform Cases. Подпроцесс использовал «вектор создания» TF-IDF. TF-IDF предоставляет больший вес словам с высокой частотой появления в пределах одного документа и с низкой частотой употреблений в других документах [8]. С помощью ограничения prune below percent, были исключены все слова, встречающиеся менее чем в 70% документов.

Оператор «Text to Norminal» позволил преобразовать входные текстовые данные в номинальные (категориальные данные).

Оператор «Numerical to Binominal» позволил преобразовать данные в биномиальные формы. Где каждая строка представляла документ, несколько столбцов представляли метаданные об этом документе, а остальные столбцы представляли найденные слова.

Одной из наиболее эффективных процедур поиска ассоциативных

правил является алгоритм Frequent Pattern-Growth (алгоритм FPG), Он позволяет не только избежать затратной процедуры генерации кандидатов, но и уменьшить необходимое число проходов БД до двух. С помощью ограничения min support = 0.7, был получен список частых наборов слов (наборов элементов), появлявшихся по меньшей мере не менее чем в 70% документов.

Оператор «Create Association Rules» позволил получить список частых наборов слов от оператора FP-Growth, а также вычислить правила, удовлетворяющие заданным ограничениям для выбранных критериев объединения. Правила ассоциации вычислялись в соответствии с критерием доверия(confidence), а также значением gain theta и значением laplace k. Использовались минимальные значения для этих 3 критериев: 0,8, 1,0 и 1,0 соответственно.

Правила ассоциации создавались путём анализа данных для частых шаблонов if / then с использованием критериев поддержки(support) и уверенности(confidence) для идентификации наиболее важных отношений.

Оператор «Create Association Rules» позволил выбрать частые наборы элементов из «FP-Growth» и создать правила ассоциации.

6 Реализация метода кластеризации. Для реализации метода потребовалось добавить операторы: «Process Documents from Files 2», «Select Attributes», «Clustering». Подроцесс «Process Documents from Files 2» включал в себя дополнительные операторы: Tokenize Non-letters, Tokenize Linguistic, Filter Stopwords, Filter Tokenz, Stem (Porter), Transform Cases, Generate n-Grams. У оператора «Generate n-Grams» было добавлено ограничение max length = 2, что позволило оператору генерировать N-граммы максимальной длинны 2. Подпроцесс «Process Documents from Files 2» использовал «вектор создания» Term Frequency. Term Frequency генерирует таблицу, где каждому уникальному слову из документа сопоставляет число, которое в свою очередь рассчитывается по формуле: (Количество слов «N» в документе / общее кол-во слов в документе). Было добавлено ограничение prune below percent, которое

позволило убрать все слова, которые появлялись менее чем в 20% документов.

Оператор «Select Attributes» позволил оставить только необходимые атрибуты (числа), все остальные данные были исключены. В настройках оператора «Select Attributes» были указаны следующие свойства: attribute filter type = value_type**,** value type = numeric.

Оператор «Clustering» (K-Means (fast)) позволил выполнить алгоритм кластеризации на числовом наборе входных данных. В свойствах оператора было указано количество кластеров (k) = 5, measure types = NumericalMeasure(EuclideanDistance), max runs = 10, max optimize steps = 10.

Кластеризация позволила разделить множество объектов на кластеры (группы). Объекты внутри кластера "похожи" друг на друга и отличаться от объектов, вошедших в другие кластеры.

7 Анализ полученных результатов. «Process Documents from Files» в качестве результата вернул таблицу, где каждому найденному слову из столбца (Word) сопоставил суммарное количество слов (Total Occurences), а также столбец с количеством слов в одном документе (Document Occurences).

Оператор «Numerical to Binominal» в качестве ответа вернул таблицу с 1850 строками. Каждая строка представляла документ, несколько столбцов представляли мета-данные об этом документе, а остальные столбцы представляли слова. На пересечении строк и столбцов встречались значения (true/false), каждое значение поясняло содержится ли «слово» в документе (true) или нет (false).

Оператор «Create Association Rules» в качестве ответа вернул таблицу со сформированными правилами ассоциации. По одной ассоциации на каждую строку. С левой стороны присутствовала возможность отсортиртировать результаты по нужному слову. Для премера было выбрано слово «perfect». Слова «buffet» и «perfect» имели поддержку(Support) равную 0.717, это означало, что пара слов встречалась вместе примерно в 71.7% документов. Значение в столбце доверие (Confidence) у них было равным 0.918, это

означало, что в документах, где появляется связанное слово со словом «buffet» была вероятность 91.8% встретить рядом и слово «perfect». Также информация была представлена в виде графа. На графе замечалась связь слов с объявленными правилами «Rule» (Support/Confidence).

Второй оператор «Process Documents from Files 2», содержащий внутри себя (в отличии от первого) дополнительный оператор «Generate n-Grams» в качестве ответа вернул таблицу. Полученные результаты от двух операторов «Process Documents from Files» очень сильно похожи. Результаты второго оператора, в отличии от первого, дополнительно включали в себя ещё и связанные пары слов, например, пара слов absolut_love означает, что в исходных текстах слова absolut и love достаточно часто встречаются вместе, всего было найдено 1304 пары слов absolut_love в 577 отелях.

Оператор «Select Attributes» в качестве ответа вернул таблицу состоящую из 1850 строк. Каждая строка ссылалась на файл с отзывами об отеле, несколько столбцов представляли мета-данные об этом отеле, а остальные столбцы представляли слова и пары слов, полученные от оператора «Process Documents from Files». На пересечении столбцов и строк присутствовали различные числовые значения в промежутке от 0 до 1. Каждое число поясняло, как часто встречалось «слово» среди всех отзывов относящихся к одному отелю. Например, у 1 отеля из списка для слова «abl» было сопоставлено число 0.017, следовательно слово «abl» с вероятностью в 1,7% могло встретиться среди всех отзывов, а вот пара слов «abl_get» уже не встречалась в отзывах к первому отелю, т.к. значение в ячейке было равно 0. Следует отметить, что оператор производил округления для всех чисел и оставлял 3 знака после запятой. В связи с чем значение 0,0001 могло округлиться до 0.

Результатом оператора «Clustering» (K-Means (fast)) являлся набор из 5 кластеров, где каждый кластер содержал внутри себя примерно «похожие» друг на друга отели. Размеры кластеров достаточно сильно отличались.

Результат из вкладки «Description»: Cluster 0: 517 items; Cluster 1: 147 items; Cluster 2: 132 items; Cluster 3: 650 items; Cluster 4: 404 items; Total number of items: 1850.

Результаты также можно представить и в виде графа.

Для наглядности кластеры были отсортированы. Ниже приведена статистика для каждого кластера. Внутри каждого кластера слова отсортированы по убыванию их популярности среди текстов ограниченных пределами одного кластера.

Attribute	cluster_0 ↓
hotel	0.496
room	0.328
rate	0.222
stai	0.207
date	0.205
content	0.201
author	0.201
rate_author	0.198
great	0.143
locat	0.134
staff	0.120
good	0.113
night	0.094
breakfast	0.092
walk	0.082

Attribute	cluster_1 ↓
room	0.248
hotel	0.219
rate	0.200
date	0.189
content	0.186
author	0.186
rate_author	0.183
beach	0.182
stai	0.181
resort	0.165
pool	0.158
great	0.154
dai	0.130
good	0.124
time	0.121

Рисунок 1 – Список самых популярных слов в 0 и 1 кластере.

Attribute	cluster_2 ↓
hotel	0.319
rate	0.304
date	0.295
content	0.293
author	0.293
rate_author	0.290
full	0.194
full_date	0.189
showreview	0.189
showreview_full	0.189
room	0.181
lass	0.160
lass_content	0.160
author_lass	0.160
stai	0.105

Attribute	cluster_3 ↓
hotel	0.397
room	0.385
stai	0.246
rate	0.221
date	0.200
author	0.195
content	0.195
rate_author	0.192
great	0.142
night	0.113
locat	0.107
staff	0.103
good	0.098
nice	0.097
clean	0.083

Рисунок 2 – Список самых популярных слов во 2 и 3 кластере.

Attribute	cluster_4
hotel	0.419
room	0.301
stai	0.179
rate	0.267
date	0.254
author	0.250
content	0.250
rate_author	0.246
great	0.118
night	0.090
locat	0.119
staff	0.098
good	0.111
nice	0.070
clean	0.075

Рисунок 3 – Список самых популярных слов в 4 кластере.

Заключение. В ходе выполнения бакалаврской работы были решены все поставленные задачи, удовлетворяющее требованиям первоначальной цели. Проведён обзор современной индустриальной среды RapidMiner для Text и Data mining. Были обнаружены закономерности с помощью методов ассоциации. Рассчитана вероятность возникновения слов и связанной пары слов среди всего множества отелей, а также для каждого отеля по отдельности. С помощью методов кластеризации все отели были распределены по кластерам, где каждый кластер содержал внутри себя похожие отели, но сильно отличающиеся от отелей из других кластеров. Все полученные результаты были представлены в графическом и табличном формате.

Исходный код для RapidMiner, полученный в результате выполнения бакалаврской работы, а также используемый каталог отзывов к отелям находятся на диске, который приложен к работе.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Современная технология обработки информационных данных Data Mining [Электронный источник]: [Сайт] URL: http://mirznanii.com/a/114723/sovremennaya-tekhnologiya-obrabotki-informatsionnykh-dannykh-data-mining (дата обращения 5.06.2017)

2 Актуальность и характерные особенности применения технологии Data Mining для решения корпоративных задач [Электронный источник]: [Сайт] URL: http://www.swsys.ru/index.php?page=article&id=297 (дата обращения 5.06.2017)

3 Recommended Data Repositories [Электронный источник]: [Сайт] URL: https://www.nature.com/sdata/policies/repositories (дата обращения 6.06.2017)

4 Wikipedia [Электронный источник]: [Сайт] URL: https://ru.wikipedia.org/wiki/API (дата обращения 6.06.2017)

5 Data Set [Электронный ресурс]: [Сайт] URL: http://sifaka.cs.uiuc.edu/~wang296/Data/ (дата обращения 6.06.2017)

6 Wikipedia [Электронный источник]: [Сайт] URL: https://ru.wikipedia.org/wiki/Интеллектуальный_анализ_текста (дата обращения 9.06.2017)

7 Data Science Platform | RapidMiner [Электронный ресурс]: [Сайт] URL: https://rapidminer.com/ (дата обращения 9.06.2017)

8 Wikipedia [Электронный источник]: [Сайт] URL: https://ru.wikipedia.org/wiki/TF-IDF (дата обращения 9.06.2017)

9 BaseGroupLabs – Технологии анализа данных [Электронный источник]: [Сайт] URL: https://basegroup.ru/community/articles/fpg (дата обращения 9.06.2017)