**ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ**

**НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ**

**«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**

Факультет компьютерных наук

Департамент программной инженерии

|  |  |
| --- | --- |
| **СОГЛАСОВАНО**  Доцент факультета компьютерных наук базовой кафедры «Системное программирование» НИУ ВШЭ, канд. физ.-мат. наук  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Д.Ю. Турдаков  «\_\_» \_\_\_\_\_\_\_\_\_\_\_ 2017 г. | **УТВЕРЖДАЮ**  Академический руководитель образовательной программы  «Программная инженерия»  профессор департамента программной инженерии, канд. техн. наук  В.В. Шилов  « » 2017 г. |

**Отчет**

**по курсовой работе**

**Построение иерархии аспектов по пользовательским отзывам об электронных устройствах**

по направлению подготовки бакалавров 09.03.04 «Программная инженерия»

Выполнила:

студентка группы БПИ143 образовательной программы

09.03.04 «Программная инженерия»

Репина А.А.

Подпись, Дата

Москва, 2017

# Реферат

Отчет 28 c., 6 рис., 2 табл., 51 источн., 2 прил.

**Ключевые слова:** аспекты; пользовательские отзывы; иерархия; электронные устройства; семантическое расстояние; характеристики.

В отчете представлены результаты курсовой работы на тему “Построение иерархии аспектов по пользовательским отзывам об электронных устройствах”, выполненной на основе приказа Национального исследовательского университета "Высшая школа экономики" № 2.3-02/0812-01 от 08.12.2016.

**Объект исследования –** аспектная иерархия пользовательских отзывов.

**Предмет исследования –** технология построения иерархии аспектов по пользовательским отзывам об электронных устройствах.

**Цель исследования:** Данная работа представляет собой подход организации различных аспектов продукта, относящегося к категории электронные устройства, в иерархию на основе знаний о потребительских отзывах. Основываясь на производной иерархии (построенной вручную), создается иерархическая организация опросов потребителей по различным аспектам продукта и совокупным мнениям потребителей по этим аспектам. При такой организации пользователь может получить обзор потребительских мнений в максимально короткий срок.

**Научная новизна работы.** Достоверность научных результатов подтверждена результатами экспериментальных исследований с прототипом программы построения иерархии аспектов.

**Методы проведения исследования.** В качестве основного логико-теоретического метода проведения исследования использовалось моделирование: разрабатывалась программа, наделенная исключительно той функциональностью, которая необходима для решения задач настоящей НИР.

**Практическая значимость.** С точки зрения практического применения, разработанный инструмент может дополняться функциональностью и использоваться для решения более широких задач, выходящих за рамки данной курсовой работы.

**Результаты работы.** По итогам выполнения исследования все поставленные цели были достигнуты.

**Содержание**

[Реферат 2](#_Toc479110700)

[Определения 4](#_Toc479110701)

[Введение 5](#_Toc479110702)

[Основная часть 7](#_Toc479110703)

[1. Обзор и анализ источников 7](#_Toc479110704)

[2. Теоретическая часть 7](#_Toc479110705)

[3. Описание эксперимента, анализ и оценка полученных результатов. 7](#_Toc479110706)

[Заключение 8](#_Toc479110707)

[Список использованных источников 9](#_Toc479110708)

[ПРИЛОЖЕНИЕ А. Примеры входных файлов программы. 10](#_Toc479110709)

[ПРИЛОЖЕНИЕ Б. Примеры выходных файлов программы. 11](#_Toc479110710)

[ПРИЛОЖЕНИЕ В. Код программы. 12](#_Toc479110711)

# Определения

В настоящем отчете о НИР применяют следующие термины с соответствующими определениями:

**Парсер -** скрипт или программа, которые используются для сбора информации с сайтов

**PMI(pointwise mutual information )** - точечные взаимные сведения или точечная взаимная информация являются мерой ассоциации, используемой в теории информации и статистике. В отличие от MI, которая основывается на PMI, PMI относится к отдельным событиям, тогда как MI относится к среднему значению всех возможных событий.

**Context** - KL-расхождение между языковыми моделями.

**Lexical** - разница в длине слов между двумя аспектами.

**Syntactic** - средняя длина кратчайшего синтаксического пути между парами аспектов в дереве.

**Семантическое расстояние** – насколько два аспекта близки друг к другу, это возможно определить с помощью набора характеристик.

**Дерево иерархии** - расположение элементов системы в порядке подчиненности (от высшего к низшему).

**Аспект** - слово или набор слов, главным словом в которых является существительное.

**Идеальный аспект** – слово-характеристика, полученная из технических описаний товаров на сайте ulmart.ru

**KL-divergence** -  это неотрицательнозначный функционал, являющийся несимметричной мерой удаленности друг от друга двух вероятностных распределений.

# Введение

С быстро растущей электронной торговлей большинство розничных веб-сайтов побуждают потребителей писать обзоры, чтобы выразить свои взгляды на различные аспекты продуктов. Огромная коллекция отзывов потребителей теперь доступна в Интернете. Эти обзоры стали важным ресурсом для потребителей и бизнеса. Потребители обычно ищут информацию о качестве в онлайн-опросах потребителей перед покупкой продукта, в то время как многие компании используют онлайн опросы в качестве важного ресурса в их разработке продукта, маркетинге и управлении взаимоотношениями с клиентами. Однако обзоры дезорганизованы, что приводит к трудностям в навигации информации и приобретении знаний. Пользователю нецелесообразно изучать обзоры потребительских мнений по различным аспектам продукта из тысяч источников. Среди аспектов продукта также неэффективно для пользователя просматривать данные и мнения потребителей по определенному аспекту. Таким образом, существует острая необходимость в организации опросов потребителей, чтобы превратить обзоры в полезную структуру знаний. Поскольку иерархия может улучшить представление и доступность информации, то видится разумным организовать аспекты продукта в иерархии и, соответственно, создать иерархическую организацию опросов клиентов.

Чтобы автоматически получить иерархию аспектов из обзоров, можно было бы обратиться к традиционным методам генерации иерархии, которые сначала идентифицируют понятия из текста, а затем определяют отношения между родителем и ребенком. Тем не менее, основанные на шаблонах методы обычно страдают от несогласованности отношений «родитель-потомок» между понятиями, в то время как методы на основе кластеризации часто приводят к низкой точности. Таким образом, путем непосредственного использования этих методов для создания иерархии аспектов из обзоров потребителей полученная иерархия обычно неточна, что приводит к неудовлетворительной организации обзора. Данная работа предназначена для построения иерархии аспектов по пользовательским отзывам об электронных устройствах на основе производной иерархии, построенной вручную, что пользоволяет максимально быстро предоставить пользователю структурированные данные потребительских мнений. В широком смысле, с точки зрения актуальности и практического применения разработанный алгоритм может пополняться дополнительной функциональностью и использоваться для дальнейших, более глобальных исследований в данной области.

**Цель исследования:** Данная работа представляет собой подход организации различных аспектов продукта, относящегося к категории электронные устройства, в иерархию на основе знаний о потребительских отзывах. Основываясь на производной иерархии (построенной вручную), создается иерархическая организация опросов потребителей по различным аспектам продукта и совокупным мнениям потребителей по этим аспектам. При такой организации пользователь может получить обзор потребительских мнений в максимально короткий срок.

**Задачи исследования:**

* изучение наиболее эффективных и применимых на практике методов построения иерархии аспектов по пользовательским отзывам об электронных устройствах;
* разработка программы построения иерархии аспектов по пользовательским отзывам об электронных устройствах с использованием анализа данных и методов машинного обучения;
* проведение эксперимента, позволяющего определить точность построение иерархии аспектов, с помощью предложенного программного решения.

**Предметом исследования** в данной работе являетсятехнология построения иерархии аспектов по пользовательским отзывам об электронных устройствах.

**Методы исследования:**

* изучение монографических публикаций и статей;
* сравнительный анализ;
* машинное обучение;
* анализ данных.

**Новизна исследования** и **достоверность исследования** определяются следующим:

Идея построения аспектной иерархии на основе пользовательских отзывов не нова для англоязычного мира, однако русскоязычный вариант алгоритма не нашел отражения в обнаруженных источниках, поэтому в силу малого количества исследований в данной области есть основания утверждать, что подобное исследование в русскоязычном формате проводится впервые. Во время проведения экспериментов ошибка может возникнуть на любом из этапов в связи с неидеальностью входных данных, которыми являются отзывы рядовых пользователей интернета. Приняв во внимание данное обстоятельство, результаты исследования в целом считаются достоверными.

**Теоретическая значимость** работы обусловлена тем, что не существует русскоязычных аналогов программ, которые строят подобные иерархии на основе пользовательских отзыва, не смотря на всю актуальность. Таким образом, исследование в данной области позволит улучшить сложившуюся ситуации.

**Практическая ценность** работы заключается в дальнейшем возможном использовании разработанного алгоритма в области анализа текстов.

# Основная часть

## Обзор и анализ источников

Для проведения исследования использовался логико-теоретический метод моделирование: была разработана программа (см. Рис.1), в которой представлено графическое изображение двумерного поля, на котором все поведение Муравья симулируется, а также все изменения в конфигурации поля отображаются динамически.

Рис. 1 – скриншот программы

Проанализировав найденные источники по данной области, можно сделать вывод, что почти все программы разработаны исключительно для ознакомления пользователя с поведением Муравья и, по сути, не несут исследовательский характер. Таким образом, для решения задачи появилась необходимость в разработке собственного алгоритма поиска длины предпериода, не опираясь на ранее придуманные алгоритмы.

## Теоретическая часть

При разработке инструмента для проведения исследования одной из основных задач являлась создание эффективного по времени алгоритма поиска длины предпериода Муравья. В результате был разработан алгоритм с вычислительной сложностью O(n), где n – количество выполненных Муравьем шагов на момент вызова алгоритма. Эффективность такого алгоритма обоснована тем, что для подсчета длины предпериода необходимо, как минимум, сделать один проход по всем шагам Муравья (которые бинарно закодированы, так как он может только повернуть налево или направо, прежде чем сделать шаг вперед), который как раз-таки и имеет вычислительную сложность O(n). Описание алгоритма смотреть в Приложении 1 настоящего отчета.

## Описание эксперимента, анализ и оценка полученных результатов.

# Заключение

Были изучены самые распространенные анонимные сети и методы деанонимизации пользователей в этих сетях, была построена классификация методов деанонимизации на основе ресурсов, использующихся атаками, была построена структура интеллектуальной системы деанонимизации пользователей.

Была построена программа для деанонимизации пользователей Tor, был проведен эксперимент, подтверждающий деанонимизацию пользователей Tor и позволяющий определить точность деанонимизации.

Показано, что атакующий без супер-современных методов машинного обучения может использовать website fingerprinting атаку. Если атакующий имеет достаточно опыта и технической компетентности, то он может построить такую систему и использовать ее для деанонимизации пользователей. Более того, предложенное решение будет работать лучше всего, если атакующий будет перехватывать Wi-Fi траффик или какую-либо другую локальную сеть, потому что для него будет проще найти трафик, относящийся к Tor и собрать его.

По окончании исследования было получено два теоретически значимых результата: − в среднем, длина предпериода больше при запусках на полях со случайным разбросом черных клеток, чем при самосопряженных запусках; − частота разброса черных клеток на поле прямо пропорционально коррелирует со средней длиной предпериода. Во время проведения исследования была разработана не имеющая аналогов в свободном доступе программа, являющаяся незаменимым инструментом для проведения экспериментов настоящей НИР. С помощью нее были достигнуты все поставленные цели исследования, а также повысился навык программирования у исполнителя. При дальнейшем исследовании в данной области необходимо определить условия, при которых Муравей Лэнгтона приобретает зацикленный характер и начинает строить «магистраль», а также доказать, что при любой изначальной конфигурации поля он способен динамически создать себе эти условия или все- таки найти такую раскраску, когда Муравей после запуска будет бесконечно двигаться хаотичным образом.

# Список использованных источников

1. Сайт материалов университета Иллинойс [Электронный ресурс]. URL: https://courses.engr.illinois.edu/cs498jh/Slides/Lecture03.pdf (дата обращения: 20.12.2016);
2. Сайт API Института системного программирования Российской академии наук [Электронный ресурс]. URL: https://api.ispras.ru/ (дата обращения: 22.11.2016);
3. С. Ким, Д. Цанг, Ж. Чен, Э. Оу, Ш. Лиу, “A Hierarchical Aspect-Sentiment Model for Online Reviews”, Департамент компьютерных наук, Корея, 2010;
4. Дж. Ю, Ж. Джа, М. Венг, К. Венг, Т. Чуа, “Domain-Assisted Product Aspect Hierarchy Generation: Towards Hierarchical Organization of Unstructured Consumer Reviews”, 2011;
5. В. Проноза, Е.В. Ягунова, “Аспектный анализ отзывов о ресторанах для рекомендательных систем е-туризма”, Санкт-Петербургский государственный университет, Санкт-Петербург, 2010;
6. А.А. Бреслав, А.П. Лукьянова, М.А. Коротков, “Построение иерархии классов по текстовым описаниям“, Санкт-Петербургский государственный политехнический университет, Санкт-Петербург, 2011;
7. Сайт документации языка Python [Электронный ресурс]. URL: https://docs.python.org/3/ (дата обращения: 20.01.2017).

# ПРИЛОЖЕНИЕ А. Примеры входных файлов программы.

# ПРИЛОЖЕНИЕ Б. Примеры выходных файлов программы.

# ПРИЛОЖЕНИЕ В. Код программы.

Программа состоит из 19 основных классов и 35 баз данных на языке Python. Текст программы на исходном языке находится в директории документация на носителе информации типа компакт-диск в связи с большим количеством строк кода.

Описание и функциональное назначение классов и интерфейсов

|  |  |
| --- | --- |
| Класс или интерфейс | Назначение |
| Aspects | Отвечает за поиск аспектов в отзывах пользователей и морфологический разбор предложений |
| Context | Отвечает за расчет характеристик Context Global и Context Local для пар аспектов |
| DB | Отвечает за создание и взаимодействие со всеми базами данных |
| FrequentAspects | Отвечает за поиск 1000 наиболее популярных аспектов |
| HierarchyBuilder | Отвечает за построение иерархии аспектов по пользовательским отзывам о электронных устройствах |
| IdealAspectsDB | Отвечает за заполнение базы данных с идеальными аспектами |
| Lexical | Отвечает за расчет характеристики Lexical для пар аспектов |
| Main | Является точкой управления программой. Отвечает за запуск работы всех классов |
| OneClassSVM | Отвечает за One Class SVM тренировку и обучение |
| PMI | Отвечает за расчет характеристик PMI Review и PMI Sentence для пар аспектов |
| SemanticDistanceLearning | Отвечает за вычисление семантического расстояния |
| RandomForest | Отвечает за вычисление семантического расстояния методами машинного обучения |
| Sentence | Отвечает за разделение пользовательских отзывов на предложения |
| Splitter | Отвечает за очистку отзывов и предложений от мусора |
| Syntactic | Отвечает за расчет характеристики Syntactic для пар аспектов |
| Unnecessary | Отвечает за удаление ненужных частей предложений |
| CategoryNames | Отвечает за создание списка категорий сайта ulmart.ru |
| DataBase\_Ulmart | Отвечает за формирование базы отзывов |
| Parser | Класс-парсер сайта ulmart.ru |