В данной работе представлен подход организации различных аспектов продукта в иерархию на основе знаний о потребительских отзывах. Основываясь на производной иерархии (построенной вручную), мы создаем иерархическую организацию 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.

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 в обучении онтологий, 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.

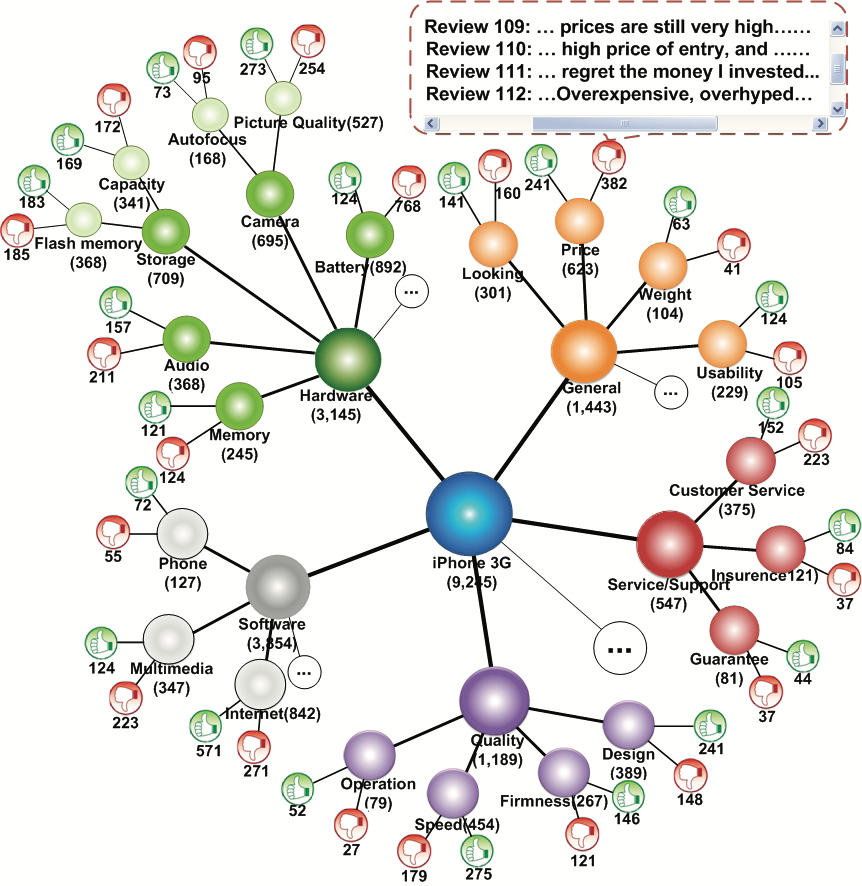


Рисунок 1 иллюстрирует пример иерархической организации обзора для продукта «iPhone 3G».

Благодаря такой организации пользователи могут легко ознакомиться с обзором аспектов продукта, а также удобно ориентироваться в обзорах и мнениях потребителей по любому аспекту.

**Подход**

Подход состоит из четырех компонентов, включая первоначальную сборку иерархии, идентификацию аспекта, семантическое дистанционное обучение и генерацию иерархии аспектов.

Описание используемых для вычисления семантического расстояния характеристик

1. **PMI**

PMI включает в себя 2 характеристики: для отзывов и для предложений.

Для вычисления обоих производится итерация по всем парам аспектов, далее производится поиск количества предложений, где есть оба аспекта из пары. Значение PMI считается как логарифм от количества предложений, где есть оба аспекта из пары, деленный на произведение их предвычисленного количества по-отдельности.

Разница в характеристиках заключается в корпусе отзывов/предложений, который передается для дальнейшей обработки в метод расчёта PMI.

1. **Contextual**

Данная характеристика состоит из двух значений: локального и глобального.

Для получения локального контекста для каждой пары аспектов были предварительно вычислены значения 4 окружающих каждый из аспектов слов в отзывах: двух справа и двух слева. Из данных предварительных значений была сформирована база данных, которая позволила повысить скорость дальнейших вычислений.

Для вычисления глобального контекста для каждого из аспектов в паре был сформирован корпус предложений, где он встречался. Далее, как для глобального, так и для локального контекста были загружены предвычисленные данные из баз данных в память, для получения результатов, поддающихся анализу, был произведен smoothing данных методом Add-1 Smoothing[[1]](#footnote-1).

После было осуществлено вычисления значения, именуемого kl divergence для локального и глобального контекстов, которое означает, насколько контексты аспектов в паре удалены друг от друга.

1. **Syntactic**

Для вычисления данной величины были предварительно сохранены в базу данных синтаксические деревья, полученные с помощью API ИСП РАН[[2]](#footnote-2), для всех предложений из отзывов. Далее для каждой из пар аспектов были найдены предложения, где встречаются они оба. Во всех таких предложениях с помощью вышеуказанной базы данных был вычислен кратчайший путь от одного аспекта к другому. Далее для каждой из пар был вычислен средний такой путь через деление суммы путей на их количество.

1. **Lexical**

Для вычисления данной величины были проитерированы все пары аспектов и вычислена разница между их длинами.

**Semantic distance**

Для вычисления семантической дистанции производится вычисления вектора w, размера 1Х5, представляющего собой набор характеристик (PMI, Lexical, Syntactic, Contextual) идеальных аспектов. Данный вектор вычисляется по формуле:

,

где f – вектор характеристик, η – константное значение, равное 0.4, I – единичная матрица, d – вектор, содержащий в себе все кратчайшие пути между идеальными аспектами в построенном вручную дереве. Далее для каждой пары аспектов семантическое расстояние равно:

,

где w - вектор, вычисленный на шаг раньше, а другие множители являются значениями характеристик для конкретной пары аспектов, которые получаются путем извлечения из баз данных.

**Построение иерархии**

Для построения иерархии аспектов было вручную построено дерево сначала для идеальных аспектов. Далее алгоритм для каждого из свободных листов дерева(родителей) выполняет поиск подходящих продолжений(детей) на основе того, насколько семантическая дистанция отличается от среднего значения дистанции для изначального дерева идеальных аспектов. Каждый новый ребенок запоминается в базу данных вместе со своим родителем.

1. https://courses.engr.illinois.edu/cs498jh/Slides/Lecture03.pdf [↑](#footnote-ref-1)
2. https://api.ispras.ru/ [↑](#footnote-ref-2)