АНАЛИЗ ТЕКСТОВЫХ ОТЗЫВОВ ПОЛЬЗОВАТЕЛЕЙ ДИСТАНЦИОННЫХ КУРСОВ НА ОСНОВЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

Малашков В.Б., Шульдова С.Г.

Белорусский государственный университет информатики и радиоэлектроники, г. Минск, Беларусь, malashkovv@gmail.com

Abstract. Keywords extraction algorithm PositionRank was described and considered as a way to extract useful information from user's reviews in massive open online courses. Results are observed over a dataset of real reviews from Coursera.

В современном мире наблюдается растущий спрос на дистанционное обучение. Это обусловлено, прежде всего, открытостью и гибкостью процесса обучения, а также необходимостью постоянной актуализации знаний для современного специалиста, особенно в сфере информационных технологий. Для самообразования в Интернете существует большое количество открытых образовательных платформ, предоставляющих интерактивные уроки, в ходе которых предлагается прослушать теоретический материал, а затем выполнить практические задания. Наиболее известные среди них: Coursera, основанная профессорами Стенфордского университета; MIT's Open Courseware – дистанционные курсы Массачусетского Технологического Института (MIT); Stepik – некоммерческая площадка, курсы для которой создают российские компании и университеты и др.

Каждый из подобных сервисов предлагает огромное количество курсов (например, MIT's Open Courseware более 2000), число их зарегистрированных пользователей измеряется миллионами, и один курс может иметь десятки тысяч отзывов. Поэтому сориентироваться при выборе курса достаточно сложно. Анализ содержания отзывов позволит осуществить слушателю обоснованный выбор курса, а также будет способствовать повышению качества содержания курсов как один из каналов обратной связи.

Обосновать выбор курса на основе анализа текстов отзывов обеспечит решение задачи «извлечение ключевых слов» (англ. key-word extraction). Суть задачи заключается в извлечении слов из документа, которые лучше всего описывают предмет, в нем обсуждаемый [1]. В данной ситуации документом является отзыв пользователя.

Из существующих методов решения данной задачи рассмотрены методы, которые относятся к алгоритмам машинного обучения без учителя, так как данный подход не требует первоначальной разметки данных.

Отзывы имеют небольшой размер: от одного предложения до целого абзаца. Для такого размера текста хороший результат показывают графовые (граф-ориентированные) алгоритмы [1], которые основаны на представлении текста в виде графа, где каждая вершина – кандидат на ключевое слово, отобранное по определенному признаку, а ребро представляет из себя отношение между кандидатами, которое может выражаться в виде появления кандидатов в окне заданного размера или по семантической близости [2].

В решаемой задаче желаемым результатом будет отбор таких ключевых фраз, которые характеризуют определенные аспекты дистанционного курса и имеют описательный характер. Например, «хороший преподаватель», «интересные задания», «сложный финальный проект» и т.п. Такие ключевые фразы как «преподаватель», «задание» или «проект» не будут информативны и полезны для конечного результата.

Вышеизложенным условиям из граф-ориентированных алгоритмов удовлетворяет алгоритм «PositionRank» [1], который состоит из следующих шагов:

1) отбор фраз-кандидатов с использованием алгоритмов для частеречной разметки (англ. part-of-speech tagging) [1]. Необходимо отобрать все кандидаты, где подряд расположены хотя бы одно прилагательное с существительным в конце. Данная задача и ее решение зависит от языка и на данный момент существуют решения как основанные на правилах конкретного языка, так и с использованием вероятностных моделей;

2) построение ориентированного графа, где каждое ребро – это появление обоих кандидатов в рамках окна заданного размера;

3) выполнение алгоритма «PageRank» [3], задача которого сводится к вычислению релевантности фраз-кандидатов. После того, как алгоритм отработал, происходит отбор фраз, начиная с самых высоких оценок релевантности.

Предложенный алгоритм был реализован и апробирован на отзывах онлайн-ресурса «Coursera» на английском языке. Примеры фраз, которые выделил алгоритм: «good course», «great introduction», «little content».

Данный метод можно улучшить за счет более точной настройки параметров алгоритма, а также за счет введения дополнительной обработки в виде поиска дубликатов, например, фразы «замечательный курс» и «отличный курс» по смыслу идентичны и должны быть семантически близки друг к другу.

Литература

1. A Review of Keyphrase Extraction [Электронный ресурс]. – Режим доступа: https://arxiv.org/pdf/1905.05044v2.pdf.

2. Ванюшкин А.С. Методы и алгоритмы извлечения ключевых слов / А.С. Ванюшкин, Л.А. Гращенко // Новые информационные технологии в автоматизированных системах. – 2016. – №19. – С.85-93.

3. PageRank Citation Ranking: Bringing Order to the Web [Электронный ресурс]. – Режим доступа: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.