УДК 004.8

Рекомендательная система новостных статей на основе методов машинного обучения

Василегин В.И. (Университет ИТМО, Санкт-Петербург)
Научный руководитель – к.т.н., доцент, Шилин И.А.
(Университет ИТМО, Санкт-Петербург)

В работе рассматривается задача формирования персонализированного списка статей-рекомендаций для пользователя сайта habr.com (далее - Habr) на основе информации о статьях, описывается процесс подготовки датасета для тестирования моделей кластеризации и классификации, рассматриваются методы построения рекомендательных систем, описывается архитектура разработанной рекомендательной системы и результаты подсчета метрик для полученных рекомендаций.

Задача построения рекомендательных систем на данный момент актуальна для многих областей и является частью таких задач, как предложение товаров в интернет магазинах, ранжирование результатов выдачи в поисковых системах, поиск подходящего контента в музыкальных, видео сервисах и СМИ. В целом, рекомендательные системы в интернете применяются с целью персонализации контента и его автоматической адаптации под текущие нужды конкретного пользователя.

Любая рекомендательная система имеет дело с двумя видами сущностей: пользователь и объект. Пользователь - получатель рекомендации и источник данных о предпочтениях, а объект - в зависимости от предметной области это может быть товар, фильм, музыкальный трек, книга или статья [1]. В данной работе рассматривается рекомендательная система для новостных статей IT портала Habr. Для её создания были проанализированы существующие подходы к построению рекомендательных систем:

1. Фильтрация, основанная на контенте (content-based filtering).
   Методы фильтрации на основе содержания основаны на описании объекта и профиле предпочтений пользователя. Описанием объекта является конечное множество его дескрипторов, таких как ключевые слова, бинарные дескрипторы и т.д, а профиль предпочтений представляет собой взвешенный вектор дескрипторов объекта, в котором веса показывают важность каждого дескриптора для пользователя и его вклад в принятие конечного решения. Этот подход позволяет подобрать объекты, похожие на те, которые нравились пользователю ранее, и опирается на методы информационного поиска и машинного обучения [2].

2. Коллаборативная фильтрация (collaborative filtering).
   Метод коллаборативной фильтрации базируется на информации об истории поведения пользователей в системе. Например, могут использоваться данные о покупках или оценках. В этом случае для пользователя находятся похожие на него по истории пользователи, а список рекомендаций формируется на основе их отношения к объектам [2].

3. Гибридные рекомендательные системы (hybrid filtering).
   Гибридные подходы сочетают коллаборативную и контентную фильтрацию, повышая эффективность рекомендательных систем. Кроме того, гибридный подход может быть полезен, если применение коллаборативной фильтрации начинается при значительной разреженности данных (“холодный старт”) [3].

В данной работе были рассмотрены все три метода, но для реализации рекомендательной системы был выбран подход основанный на фильтрации контента, т.к. собрать датасет с историей пользователей на сайте habr.com не представляется возможным. Для оптимизации и уменьшения времени на формирование списка с рекомендациями был использован алгоритм иерархической кластеризации Birch, который позволил уменьшить количество статей-кандидатов.

Также в работе была решена проблема пользовательских тегов, количество которых на портале Habr превышает 100,000, при этом некоторые из тегов никак не характеризуют статью. В качестве решения был использован алгоритм мультилейбл-классификации основанный на подходе OneVsRest.

По итогам проделанной работы была разработана рекомендательная система, позволяющая сократить время пользователей на поиск интересующих их статей.

Список использованных источников:

1. Пономарев А. В. Обзор методов учета контекста в системах коллаборативной фильтрации //Информатика и автоматизация. – 2013. – №. 30. – С. 169-188.

2. Кутянин А. Р. Рекомендательные системы: обзор основных постановок и результатов //Интеллектуальные системы. Теория и приложения. – 2017. – Т. 21. – №. 4. – С. 18-30.

3. Wang X., Wang Y.: Improving content-based and hybrid music recommendation using deep learning. In: Proceeding of ACM MM, 2014.