УДК 004.021

Правдоподобие популярности поста

Замятин Е. И., Университет ИТМО, г. Санкт-Петербург
Научный руководитель – Фильченков А. А., к.т.н., доц. каф. КТ Университета ИТМО

Введение

На сегодняшний день социальные сети являются неотъемлемой частью жизни современного общества. Одной из ключевых составляющих практически любой социальной сети является контентная экосистема. В рамках этой экосистемы социальная сеть выступает площадкой, которая соединяет создателей контента – авторов, и потребителей контента – пользователей.

На данный момент существует огромная конкуренция на рынке социальных сетей и контентных площадок. Каждая компания борется за внимание и время пользователей. В условиях такой конкуренции, огромным преимуществом контентной площадки может быть рекомендательная система, которая будет понимать интересы пользователя и соединять его с релевантными ему авторами. Таким образом, в текущих реалиях, рекомендательные системы являются неотъемлемой частью социальных сетей.

Естественной и распространённой проблемой для такого рода систем является проблема холодного старта. У этой проблемы есть две стороны. Первая – холодные пользователи. Когда на площадке появляется новый пользователь, про него еще нету никакой информации и рекомендательная система не может предложить ему релевантный контент. Однако очень важно заинтересовать пользователя, чтобы он остался на площадке, а не ушел на другую. Вторая проблема – холодный контент. Это может быть как новый автор, который только пришел на платформу, так и новый пост. В случае холодного контента есть возможность определить тематику, к которой он относится на основании контентных признаков, таких как текст или картинка, и исходя из этого понять, каким «горячим» пользователям этот контент мог бы быть интересен. Однако в случае с контентом возникает другая проблема – обычно на «зрелой» платформе генерируемого контента в разы больше, чем суммарное число просмотров, которые могут сделать пользователи за какой-то промежуток времени. Поэтому нам важно уметь определять качественный контент и давать ему большее число показов. К сожалению, понятие «качество» контента трудно формализуемо. Однако есть способ понять, что контент «хороший» или «плохой» по пользовательским взаимодействиям. Вот тут и встает проблема холодного контента, когда мы не можем понять его качество из-за отсутствия показов.

Цель работы

Целью данной работы является создание системы, которая будет находить популярный контент на платформе учитывая его качество. С помощью этой системы планируется улучшить пользовательский опыт новых пользователей, а также решить проблемы переизбытка контента, давая приоритет более качественному контенту.

Базовые положения исследования

В ходе данной работы планируется разработать методологию определения качества и популярности контента. На эту тему проводилось не так много исследований, так что прогресс в решения подобного рода задач остановился на весьма простых техниках основанных на подсчете явных взаимодействий и применении различных статистических приемов. Довольно распространенной проблемой контентых площадок является наличие кликбейтного контента, а также накруток, что делает неприменимым простые методы.

алгоритм машинного обучения, которой будет способен по данным контентным признакам поста и пользовательским взаимодействия определять его уровень качества и популярности.

Предварительные результаты

На данный момент проведено исследование существующих наработок и собран набор данных для обучения, состоящий из контентных признаков постов и пользовательских взаимодействий. Помимо этого, была разработана методология определения популярного и при том качественного контента, при том не пессимизирующая менее популярные типы контента. Сути методологии заключается в следующем: по контентым данным о посте обучается модель машинного обучения, которая предсказывает пользовательские взаимодействия, а именно ctr'ы поста по различным типам действий. По полученной модели определяется правдоподобие того, что пост в действительности набирает данное число взаимодействий.

Основные результаты

Было проведено исследование эффективности предложенного метода в сравнении с базовой моделью популярности. По результатам экспериментов на новых пользователях явно видно, что предложенный метод превосходит базовый по многим показателям.

Список литературы

1. Towards Detecting Anomalous User Behavior in Online Social Networks, Bimal Viswanath et. al. 2014
2. User-Interactions on Reddit, Maria Glenski et. al. 2017;
3. Popularity and Quality in Social News Aggregators: A Study of Reddit and Hacker News, Greg Stoddard 2015;
4. The Impact of Crowds on News Engagement: A Reddit Case Study, Benjamin D. Horne et. al. 2017;
5. PopRank: Ranking pages' impact and users' engagement on Facebook, Andrea Zaccaria et. al. 2018;