

УДК 004.852

ПОСТРОЕНИЕ ГРАФА СВЯЗЕЙ МЕЖДУ МЕТРИКАМИ ПО МНОЖЕСТВУ ЭКСПЕРИМЕНТОВ

Хусаинов А.Р. (Университет ИТМО)

Научный руководитель – доцент ФИТиП Сметанников И.Б.
(Университет ИТМО)

Современные крупные сервисы регулярно проводят А/В-тестирования, чтобы подобрать оптимальную конфигурацию, которая увеличивала бы важные для них метрики. Знание о связях между этими метриками позволило бы значительно уменьшить количество необходимых экспериментов. Построение графа связей между этими метриками является одним из наиболее наглядных способов выразить существующие зависимости.

Введение.

Существует набор данных от компании «ВКонтакте» с информацией о проведённых А/В-тестированиях. Данные представлены как изменение метрик в тестовой группе относительно контрольной группы в каждом конкретном эксперименте. Также представленные изменения содержат такую дополнительную информацию как p -value (нулевая гипотеза – тестовая и контрольная группа совпадает), по значению которой можно установить статистическую значимость изменения. Множество метрик также разделено на два подмножества: «короткие» и «длинные» метрики. Первое подмножество состоит из метрик, на которые компания может предсказуемо влиять (число комментариев или просмотров от каждого пользователя), но которые не являются основными для компании. Второе подмножество состоит из метрик, которые хочется оптимизировать. Такой является, например, среднее время, проведённое в приложении.

Задача состоит в том, чтобы найти взаимосвязи между данными и короткими метриками и понять, каким образом следует влиять на короткие метрики, чтобы оптимизировать значение длинных. На текущий момент решение похожей задачи в А/В-тестировании описала в своём исследовании компания «Facebook», где они предлагают использовать метод Instrumental Variable. Менее похожую задачу, целью которой было создание искусственной более чувствительной метрики для уменьшения необходимого числа экспериментов, решала компания «Яндекс». Однако для повторения такого решения нужны дополнительные данные, которых на данном этапе нет. Упрощение же предлагаемой модели не привело к желаемым результатам.

Основная часть.

Для решения задачи было решено использовать модели каузальных графов. Им является ориентированный граф, ребро в котором означает, что вершина, в которую входит ребро, зависит от вершины, из которой оно исходит. Задача построения графа, который отражает существующие зависимости, является NP-полной. Однако существуют эвристические алгоритмы, строящие субоптимальные графы по набору данных. Также частный случай искомого графа – дерево, можно эффективно построить на основе алгоритма поиска минимального остовного дерева. Но такой частный случай сильно ограничивает структуру графа, к тому же он позволяет построить только неориентированный граф.

Ещё одной сложностью построения желаемого графа является то, что все алгоритмы умеют находить графы лишь с точностью до I-эквивалентности. Иными словами, ориентацию рёбер в таких графах нужно будет устанавливать с помощью новых экспериментов, хотя существуют подходы, которые позволяют минимизировать их число.

Затем, имея граф зависимостей, можно будет оценить функции зависимостей с помощью, например, метода Instrumental Variable. Это необходимо чтобы понимать, как именно необходимо будет изменить короткие метрики, чтобы оптимизировать длинные.

Выводы.

Полученный в результате исследования граф связей позволит значительно сократить количество экспериментов, целью которых является оптимизация длинных метрик с помощью коротких.

Хусаинов А.Р. (автор)

Подпись

Сметанников И.Б. (научный руководитель)

Подпись