Метод оценки надежности исправления поисковых запросов

Ильвохин Дмитрий Евгеньевич

Студент

Факультет прикладной математики и физики МАИ, Москва, Россия

E-mail: ilvokhin.d@gmail.com

Для поиска информации в интернете используются поисковые системы, которые формируют страницу результатов по запросу, введенному пользователем. Однако, около 10% запросов к поисковой системе содержат ошибки. Как правило, поисковая выдача по запросам с ошибками является нерелевантной. Для борьбы с подобными выдачами используются системы исправления ошибок в поисковых запросах, которые генерируют предполагаемого кандидата для исправления пользовательского запроса. Имея кандидата для исправления запроса, системе исправления ошибок нужно решить документы по какому запросу (оригинальному или исправленному) показать пользователю.

В работе описывается метод оценки надежности исправления поисковых запросов, который для произвольной пары *(запрос, исправление)* позволяет определить является ли исправление, предложенное системой, надежным для запроса, введенного пользователем.

Метод основан на построении и использовании модели машинного обучения для решения задачи бинарной классификации. Для построения модели машинного обучения использовался алгоритм градиентного бустинга [1] над небрежными деревьями (англ. gradient boosting oblivious trees [2]).

Метод должен использоваться в поисковой системе исправления ошибок, благодаря этому появляется возможность использовать некоторые полезные признаки, уже расчитанные этой системой. Но в то же время, накладываются существенные ограничения на сложность используемой модели машинного обучения, так как система исправления ошибок работает под высокой нагрузкой и должна возвращать ответ за очень короткий промежуток времени. Необходимо подобрать баланс между качеством модели машинного обучения и скоростью предсказания.

В результате разработанный метод был использован в существующей системе исправления ошибок Поиска Mail.Ru. Точность оценки надежности исправлений всей системы увеличилась на 8.8% при небольшом снижении полноты на 2.9%. В ходе разработки качество

модели машинного обучения было улучшено на 3.8% по сбалансированной F-мере по сравнению с базовой версией модели. Качество системы исправления ошибок в поисковых запросах улучшено на 2%.

Литература

1. Friedman J. Greedy function approximation: a gradient boosting machine // Annals of statistics, 2001, P. 1189–1232.
2. Gulin A. Matrixnet // Technical report, `http://www.ashmanov.com/arc/searchconf2010/08gulin-searchconf2010.ppt`, 2010, P. 17.