

РАНЖИРОВАНИЕ С ПОМОЩЬЮ ПОВЕДЕНЧЕСКОЙ ИНФОРМАЦИИ

Викулин Всеволод Александрович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: va.vikulin@yandex.ru

Задача ранжирования заключается в сортировке набора документов из коллекции по убыванию релевантности к запросу пользователя.

Всю информацию, которая позволяет поисковой машине выполнить задачу ранжирования можно разделить на два больших класса: текстовая информация и поведенческая информация.

Под текстовой информацией понимаются признаки, оценивающие текстовую близость документа и запроса.

Поведенческие признаки оценивают документ по историческим поисковым сессиям. Простейшим примером поведенческого признака является отношение количества кликов в документ по запросу, деленное на общее число поисковых сессий по запросу.

По последним данным около 40% поисковых запросов уникальны, то есть в истории поисковой машины нет информации о том, на какие документы кликали люди, задавшие эти запросы. Для них нельзя сделать поведенческие признаки. Существует огромное число запросов, которые хотя и задавались ранее пользователями, но лишь небольшое число раз. Для них тяжело делать поведенческие признаки, так как данные очень шумные.

Для борьбы с этой проблемой обычно применяют сглаживание с помощью методов кластеризации. Самым популярным методом является случайное блуждание по двудольному графу [1–2]. Большим недостатком данного метода является невозможность его применения для новых запросов, которых не было в истории.

В данной работе предлагается метод, позволяющий для нового запроса найти похожие для него запросы, но которые уже часто задавались пользователями. Это позволяет оценить поведенческие признаки даже для тех запросов, которые раньше ни разу не задавались. Метод заключается в построении модели, которая по текстам пары запросов определяет степень их похожести друг на друга.

В рамках данной работы была сформирована тестовая выборка, позволяющая оценивать качество модели, предложен метод, позволяющий находить похожие запросы для формирования обучаю-

щей выборки, проведены эксперименты для различных моделей. Рассматривались как модели обучения без учителя, так и с учителем. Было продемонстрировано, что предлагаемые модели на тестовых данных показывают качество примерно на 20% выше, чем классические модели текстового поиска. Данные для экспериментов предоставлены отделом поиска компании «Mail.Ru Group».

Литература

1. Bilenko M., White R. Mining the search trails of surfing crowds: identifying relevant websites from user activity // Proceedings of the 17th international conference on World Wide Web, Beijing, China, 2008, P. 51–60.
2. Gao J., Yuan W., Li X., Deng K., Nie J-Y. Smoothing clickthrough data for web search ranking // Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, Boston, USA, 2009, P. 355–362.