Метод поиска семантически близких веб-документов на основе графа переходов из поисковых систем

Белоусов Степан Леонидович

Студент

Кафедра вычислительной математики и программирования МАИ, Москва, Россия

E-mail: s.belousov@corp.mail.ru

Задача поиска похожих по смыслу документов в сети Интернет имеет достаточно высокое практическое значение на сегодняшний день. Например, она часто возникает при разработке различных форумов, блогов, новостных порталов и других сервисов, имеющих большой объем контента, оформленного в виде отдельных однотипных страниц (статей, постов, тредов). Важной составляющей удобной навигации по таким сервисам является наличие между их страницами ссылок, построенных по принципу семантической близости. Другими словами, если документ с сайта содержит несколько ссылок на другие документы того же сайта, имеющие схожую тематику, они нередко способны заинтересовать пользователя и оказаться ему полезны. Поскольку количество страниц на многих сервисах исчисляется тысячами и даже миллионами, встает вопрос об автоматическом построении подобных ссылок.

Для решения описанной задачи могут применяться различные подходы, основанные как на сопоставлении содержимого документов (статистический анализ текста), так и на структуре связей между ними (кластеризация графа ссылок) и на поведенческих данных (рекомендательные системы). В данной работе предлагается альтернативный метод решения указанной задачи, использующий информацию о переходах на страницы сайта из поисковых систем. В сравнении с перечисленными подходами он отличается простотой реализации при достаточно высоком качестве полученных результатов.

Основная идея предложенного метода заключается в том, что два документа, находящиеся в выдаче поисковика по одному запросу, достаточно похожи. Можно составить двудольный граф, в одной доле которого будут находиться веб-страницы, а в другой — запросы из поисковых систем. Ребро в таком графе означает присутствие документа в выдаче по запросу. Алгоритм заключается в обходе данного графа в ширину, начиная с того документа, для которого ищутся похожие варианты, и до достижения определенной глубины поиска. Ре-

зультатом являются посещенные вершины с другими документами. Для построения указанного графа достаточно иметь логи переходов пользователей на страницы сайта из поисковых систем.

Описанный метод дает высокую точность результатов, но его полнота бывает недостаточна. Поэтому в случаях, когда результатов найдено слишком мало, предлагается в качестве второго шага использовать его модификацию, при которой вместо целых запросов в графе хранятся термы, т.е. каждая вершина с запросом разбивается на отдельные вершины со всеми словами исходного запроса. Такая структура графа позволяет находить документы, к которым ведут запросы, имеющие много общих слов, что повышает полноту до приемлемого уровня. Стоит отметить, что если первый алгоритм, как правило, находит веб-страницы на очень близкие темы, то второй вносит в результаты больше разнообразия, хотя они по-прежнему остаются семантически похожими.

С учетом объемов контента на сайтах, размеры упомянутых графов часто оказываются очень велики, и для достижения хорошей производительности может потребоваться их распределенная предобработка, вплоть до полного предрасчета результатов по каждому документу. Впрочем, обходы графов при помощи распределенных вычислений реализуются достаточно легко, например, средствами парадигмы Map-Reduce [1].

После выполнения вышеописанных обходов необходимо отранжировать найденные результаты, после чего, возможно, оставить лишь несколько лучших. В целом, это отдельная самостоятельная задача, и здесь также могут быть использованы различные методы, в том числе с применением машинного обучения. Однако практические эксперименты показывают, что хорошего качества здесь зачастую можно достичь и при использовании простой линейной комбинации значений факторов с ручной настройкой весов. При этом важным фактором выступает расстояние, пройденное в графе. При сравнении же результатов, найденных по отдельным словам, главную роль играет редкость и специфичность набора слов, по которому они были найдены. Для ее оценки можно использовать подход, основанный на метрике IDF [2].

Предложенный метод был успешно применен на практике. Он был реализован и внедрен на портале Ответы@Mail.Ru для генерации блока «Похожие вопросы» взамен использовавшегося ранее решения, основанного на полнотекстовом поиске заголовка вопроса в базе документов портала. Новый метод показал лучшие результаты

как с точки зрения полноты (число показов блока), так и с точки зрения точности (CTR блока, т.е. его кликабельность). Графики качества системы представлены на рис. 1, момент перехода на новый алгоритм хорошо заметен. Покрытие (доля тех вопросов в потоке посещений, для которых были найдены похожие) повысилось с 48.3% до 93.4%, при этом CTR возрос с 5.9% до 14.2%. Вместе эти изменения эквивалентны увеличению общего числа переходов по «Похожим вопросам» в 4.6 раза.

Иллюстрации

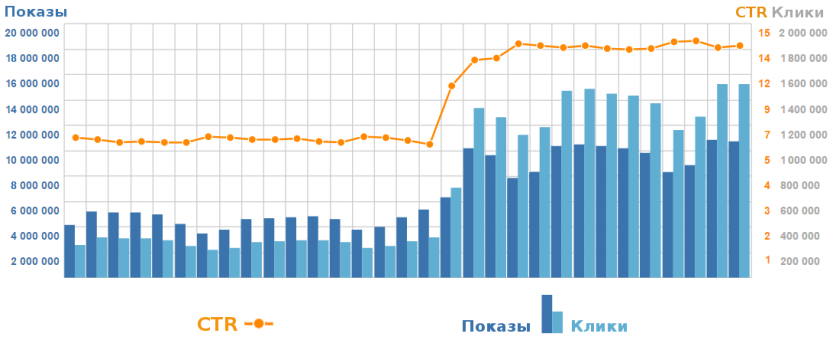

Рис. 1. Изменения показателей качества системы «Похожие вопросы» на портале Ответы@Mail.Ru при внедрении предложенного алгоритма.

В заключение хотелось бы выразить благодарность руководителю проекта Поиск@Mail.Ru, старшему преподавателю кафедры вычислительной математики и программирования МАИ Калинину А. Л. за интересную практическую задачу, в которой удалось применить описанный в работе алгоритм, а также за помощь в подготовке доклада.

Литература

1. Graph algorithms using Map-Reduce, IIIT, Hyderabad: `http://search.iiit.ac.in/cloud/presentations/6.pdf`
2. Manning C., Raghavan P., Schütze H. Introduction to Information Retrieval. — New York: Cambridge University Press, 2008. — P. 117.