Создать модель для поисковой системы в 50 раз быстрее

Учёными МФТИ и ВЦ РАН Андреем Кулунчаковым и Вадимом Стрижовым предложен метод автоматического построения ранжирующих моделей. Ранжирующие модели ищут документы в коллекциях согласно запросам пользователей. Новый метод значительно повышает скорость построения моделей. Полученные модели имеют интерпретируемый вид и состоят из небольшого числа элементарных функций. Результаты исследования опубликованы в журнале Expert Systems with Applications. Этот журнал стоит на первой позиции в рейтинге журналов по искусственному интеллекту по версии Google Scholar.

При поиске в коллекциях, содержащих миллионы документов, пользователь ожидает получить небольшой полезный список. Документы списка должны быть отранжированы согласно поисковому запросу. Остальные документы для пользователя являются информационным шумом. Цель пользователя – найти нужный документ, сформулировав запрос небольшой длины. Предложенный метод строит ранжирующие модели, позволяющие быстро достичь эту цель. Эти модели являются ядром современной поисковой системы.

Андрей Кулунчаков, студент Кафедры интеллектуальных систем МФТИ, комментирует: *«Постановка задачи предполагала использование только коллекций документов и поисковых запросов. Не допускалось использование никакой внешней информации о контексте, в котором выполнялся поиск. Такая задача имеет наиболее общий характер. Ранжирующие модели, предназначенные для быстрого и точного поиска информации, используются во множестве областей от спам-фильтров до колл-центров».*

Ранжирующая модель состоит из элементарных функций. Построить модель означает задать суперпозицию этих функций. Качество построенной модели оценивается критерием, который включает экспертные оценки адекватности отранжированного списка документов. Требуется найти модель наиболее высокого качества. Работа учёных была направлена на оптимизацию способа построения такой модели. Более качественная модель содержит большее число функций и построить ее сложнее.

Одним из способов построения моделей является генетическое программирование. Такое название оно получило из-за схожести с механизмом естественного отбора в природе. В ходе решения задачи строится множество промежуточных поколений моделей. Модели поколения в большей или меньшей степени похожи на искомую модель высокого качества. Алгоритм отсеивает модели низкого качества. Путем естественного отбора на основе оставшихся он создаёт более подходящие. Лучшие особи имеют большую вероятность быть включенными в следующее поколении. Сменяя множество поколений мы приближаемся к оптимальному решению.

К сожалению, так происходит в теории. На практике число моделей растет быстрее чем экспоненциально с ростом сложности модели. Для перебора моделей, состоящих из восьми функций, требуется не менее суток вычислений. При этом следует перебрать все варианты, которые могут доставить наилучшее решение. В предшествующих работах это достигалось методами полного перебора.

Именно проблему быстрого построения и выбора моделей решали Андрей Кулунчаков и Вадим Стрижов в рамках своего исследования. Они создали новый подход к порождению ранжирующих моделей для поиска документов в больших коллекциях. Исследователи решили проблему стагнации порождаемых моделей. Когда в сменяющих друг друга поколениях модели структурно похожи, и их скрещивание не даёт существенно новых результатов, происходит стагнация, или «застой». Тогда вероятность появления качественной модели существенно снижается. Для того, чтобы избежать стагнации, в поколение добавляются новые модели. Они создаются с целью повышения разнообразия.

Чтобы показать, что созданный метод получает модели, превосходящие по качеству современные альтернативы, авторы поставили численный эксперимент. Были использованы базы данных Национального института стандартов и технологий США, предназначенные для анализа и сравнения подобных систем. Они состояли из двух миллионов документов и двухсот тысяч запросов. Эксперимент показал, что полученные модели имеют более высокое качество ранжирования согласно принятому критерию MAP. Сам же метод позволяет получить модель высокого качества за существенно меньшее время.

Значимость работы российских ученых трудно переоценить: мы с вами будем тратить на существенно меньше времени при поиске информации. Открывая почту, мы будем видеть меньше спама, а в организациях станет проще обмениваться полезной информацией. И область применения этим не ограничивается. Ученые говорят, что это только начало, работа продолжается и нас ждут новые открытия.