Секция «Вычислительная математика и кибернетика»

Влияние объема данных на использование инкрементных методов в коллаборативной фильтрации

Полежаева Елена Андреевна
Студент
Московский государственный университет имени М.В. Ломоносова, Факультет вычислительной математики и кибернетики, Москва, Россия
E-mail: lena_polejaeva@mail.ru

Методы коллаборативной фильтрации (CF) становятся все более актуальными, так как они используются в рекомендующих системах и системах управления взаимоотношениями с клиентами (CRM) для автоматического формирования персональных предложений. Основная идея CF состоит в том, что схожие клиенты интересуются схожими объектами. Объектами могут быть сайты, товары, услуги и т.д. Такое предположение позволяет определить клиентов со схожими предпочтениями (коллаборация) и спрогнозировать величину интереса (рейтинг, количество посещений сайта и т.д.) для ранее не оцененных ими объектов. В последнее время важно уметь обрабатывать информацию в реальном масштабе времени. Так как речь идет о миллионах клиентов и объектов, важность приобретают инкрементные методы, позволяющие быстро пересчитывать оценки сходства при появлении новых клиентов, объектов и при изменении значений в ячейках исходной матрицы. Исходными данными является разреженная матрица Y, строки которой соответствуют n клиентам, столбцы — d объектам. Каждая заполненная ячейка матрицы содержит информацию об использовании данным клиентом данного объекта. Задача состоит в том, чтобы для произвольного клиента спрогнозировать оценки предпочтительности объектов по всем незаполненным ячейкам в строке матрицы Y.

Предлагается метод, основанный на использовании инкрементного сингулярного разложения (ISVD) [1] и обобщенного алгоритма обучения Хебба (GHA) [2]. При этом для построения начального SVD используется GHA, т.к. GHA заполняет пропуски в Y так, что применение SVD для Y становится возможным. При добавлении нового клиента GHA не требует полной перенастройки алгоритма [3], что очень важно при работе с большими данными. Совместное применение ISVD и GHA дает неплохие результаты в условиях постоянных обновлений в матрице Y.

Также вводится функционал, позволяющий учесть информацию о том, что в ячейках Y находятся порядковые значения (например, рейтинги). Подбираются параметры для уменьшения суммы квадратов ошибок между значениями, полученными моделью, и реальными данными. При этом ограничений на сами параметры не накладывается.

В результате экспериментов выявлено, что возможно добавлять клиентов, не модифицируя профили (сжатые описания) объектов. При увеличении объема данных уменьшается среднеквадратичная ошибка (RMSE – Root Mean Square Error) и увеличивается скорость сходимости алгоритма (при изменении от 2000 до 10000 известных рейтингов число итераций и RMSE меняются от 1000 итераций и RMSE, равного 2, до 220 итераций и RMSE, равного 0,3, соответственно). Таким образом предложенный метод может быть полезен в современных динамических приложениях CF.

Литература

1. M. Brand. Fast Low-rank modifications of the thin singular value decomposition // Linear Algebra and Its Applications, Vol. 415, Issue 1, Pp. 20–30, 2006.

2. G. Gorrell. Generalized Hebbian Algorithm for Incremental Singular Value Decomposition in Natural Language Processing.// Proceedings of EACL, 2006.

3. G. Takacs, I. Pilaszy, B. Nemeth and D. Tikk. Scalable Collaborative Filtering Approaches for Large Recommender Systems.// The Journal of Machine Learning Research, Vol. 10, Pp. 623–656, 2009.

Слова благодарности

Исследования поддержаны грантом РФФИ 10-07-00609-а.