УДК 004.891

Семантический поиск научных публикаций на основе коллаборативной фильтрации

Костин В.В.

Вычислительный центр им. А.А. Дородницына РАН

В докладе рассмотрен подход подборки научных публикаций при семантическом поиске и подборе рекомендуемых научных публикаций для конкретного пользователя. Используется система оценок каждым пользователем публикации, с которой он работал. В частности, такие действия как добавление в свою электронную библиотеку, аннотирование, рекомендации оценку увеличивают. Применяется предварительная обработка данных – вычисляются базовые предикторы каждого пользователя, на их основе вычисляются базовые предикторы каждой публикации, на основе которых вычисляются базовые предикторы каждой оценки, данной пользователем публикации. Вычисление базовых предикторов пользователей производится на множестве научных публикаций, с которыми они работали, а базовые предикторы научных публикаций – на множестве работавших с ними пользователей. Для уменьшения возможной погрешности при вычислении предикторов вводится демпфирующий член. Базовые предикторы позволяют нивелировать особенности отдельных пользователей и привести их систему оценок к единой оценочной шкале.

Для поиска рекомендуемых научных публикаций используются коллаборативная фильтрация по пользователям и научным публикациям. Предварительно вычисляются близость пользователей и научных публикаций друг к другу на основе коэффициента Пирсона. На ограниченном множестве наиболее близких пользователей вычисляются предполагаемые относительные рейтинги научных публикаций, по которым они ранжируются и предоставляются пользователю. Относительные рейтинги нормализуются при помощи стандартного отклонения исследуемого пользователя и наиболее близкой к нему группы с добавлением базовых предикторов, чтобы соответствовать системе уже зафиксированных оценок у научных публикаций, с которыми работал пользователь. В случаях, когда пользователь и группа близких к нему пользователей системы использует только бинарную систему оценок (например, добавление/не добавление в свою библиотеку), данное вычисление оценок публикаций вырождается в ситуацию оценки близости ограниченного множества наиболее близких пользователей, нашедших данный научный труд для себя полезным.

При вычислении предполагаемых рейтингов рейтинг всех публикаций системы представляется вектором, а множество оценок, заданных пользователями научным публикациям – матрицей. Для оптимизации предварительных вычислений оптимальным видится использование сингулярного разложения данной матрицы с целью уменьшения пространственности вычислений.

Литература

1. *Ekstrand M. D., Riedl J. T., Konstan J. A.* Collaborative filtering recommender systems //Foundations and Trends in Human-Computer Interaction. – 2011. – Т. 4. – №. 2. – С. 81-173.
2. *Костин В.В*. Обзор семантических моделей, описывающих научные публикации и научно-исследовательскую деятельность. // Труды 16-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2014, Дубна, Россия, 13–16 октября 2014 г.
3. *Костин В.В*. К вопросу создания поддержки работы с научными публикациями // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2014. Т. 12, вып. 4. С. 32–37.
4. *Костин В.В*. Семантический поиск в системе поддержки работы с научными публикациями. // Труды 58-й научной конференции МФТИ 23–28 ноября 2015 года Москва–Долгопрудный–Жуковский МФТИ.