Многомодальные тематические модели на гиперграфах

Жариков Илья Николаевич

Студент

Факультет управления и прикладной математики МФТИ, Москва, Россия

E-mail: zharikov.i.n@yandex.ru

Методы тематического моделирования развиваются на протяжении последних 20 лет для решения задач статистического анализа текстовых коллекций, включая информационный поиск, классификацию, категоризацию, сегментацию текстов [1]. И общим для многих подходов является предположение о существовании у объектов векторных описаний, которые соответствуют тематическим интересам людей, то есть описывают семантику этих объектов. В роли объектов могут выступать текстовые документы, слова или ключевые фразы, пользователи, рекламные объявления, товары или услуги и т. д. Семантика объектов является скрытой (латентной), но косвенно проявляется в транзакционных данных. Примерами транзакций являются взаимосвязи или взаимодействия объектов: пользователь создал (прочитал, рейтинговал, лайкнул) документ, сделал запрос, кликнул рекламное объявление, прокомментировал сообщение другого пользователя, слово встретилось в документе, в объявлении, в запросе пользователя, и т. д. Выявление латентных семантических описаний объектов и составляет суть *тематического моделирования* (topic modeling). Знание этих описаний позволяет решать множество задач анализа данных: смысловой информационный поиск, навигация в больших текстовых коллекциях, обнаружение и отслеживание темы в потоках новостей, таргетирование рекламы и т. д.

В классических задачах тематического моделирования участвуют только две модальности: документы и слова, для которых строятся тематические профили — дискретные распределения вероятностей на множестве латентных тем Для решения данной задачи используются вычислительные методы низкоранговых матричных разложений [2]. Аналогичные методы применяются в рекомендательных системах и коллаборативной фильтрации, с тем отличием, что в роли документов и слов выступают пользователи и предметы. Многомодальные тематические модели [4] позволяют описывать взаимосвязи между объектами разных модальностей.

Социальные сети дают пример ещё более сложной структуры мультимодальных транзакционных данных [3, 5]. Важную информацию несёт не только текст сообщения, но и сопровождающие его метаданные, включая отсчёт времени, автора сообщения, его географическое положение, социально-демографические данные и т. д. При этом появляются взаимосвязи не только между парами объектов разных модальностей, но и между тройками. Таким образом, актуальной проблемой является обобщение методов мультимодального тематического моделирования для анализа транзакционных данных о взаимодействиях между объектами различных модальностей, включая парные, тройные или более сложные взаимодействия.

Адекватной математической моделью для представления мультимодальных транзакционных данных является гиперграф. Вершинами гиперграфа являются объекты различных модальностей, причем с каждой вершиной связан неизвестный латентный тематический профиль. Наблюдаются транзакции между объектами, которые описываются рёбрами гиперграфа. Задача заключается в том, чтобы по этим данным восстановить тематические профили объектов. В данной работе развиваются мультимодальные и гиперграфовые методы вероятностного тематического моделирования, позволяющие восстанавливать латентные тематические профили вершин гиперграфа по наблюдаемой выборке его гиперрёбер (транзакций).

Литература

1. Blei D. M. Probabilistic topic models // Communications of the ACM. 2012. Vol. 55, № 4. P. 77–84.
2. Blei D. M., Ng A. Y., Jordan M. I. Latent dirichlet allocation // Journal of machine Learning research. 2003. Vol. 3, № 1. P. 993–1022.
3. Mei Q. et al. Topic modeling with network regularization // Proceedings of the 17th international conference on World Wide Web, Beijing, China, 2008, P. 101–110.
4. Vorontsov K. et al. Non-Bayesian additive regularization for multimodal topic modeling of large collections // Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications. Melbourne, Australia, 2015, P. 29–37.
5. Yin H. et al. Modeling location-based user rating profiles for personalized recommendation // ACM Transactions on Knowledge Discovery from Data (TKDD). 2015. Vol. 9, № 3, P. 19.