2024-09-03

Интеллектуальный анализ Web-данных

Он же **web data mining** или просто **web mining**

Что вообще такое веб-данные

Данные в контексте веба (т.е. Интернет как совокупности ресурсов, предоставляющих человеко- и машиночитаемые данные) всегда в современном понимании являются **мультимодальными**.

Это и текст, и изображения, и аудио, и видео (которое, конечно, можно представить в виде набора изображений и аудио, но применительно к анализу видео все равно должно рассматриваться как специфический модус — у него есть понятие временного ряда, например)

Веб-данные, соответственно, хранятся и собираются с веб-ресурсов. Веб-ресурс в нашем приближении — любое хранилище данных (в широком смысле) с доступом в глобальные сети.

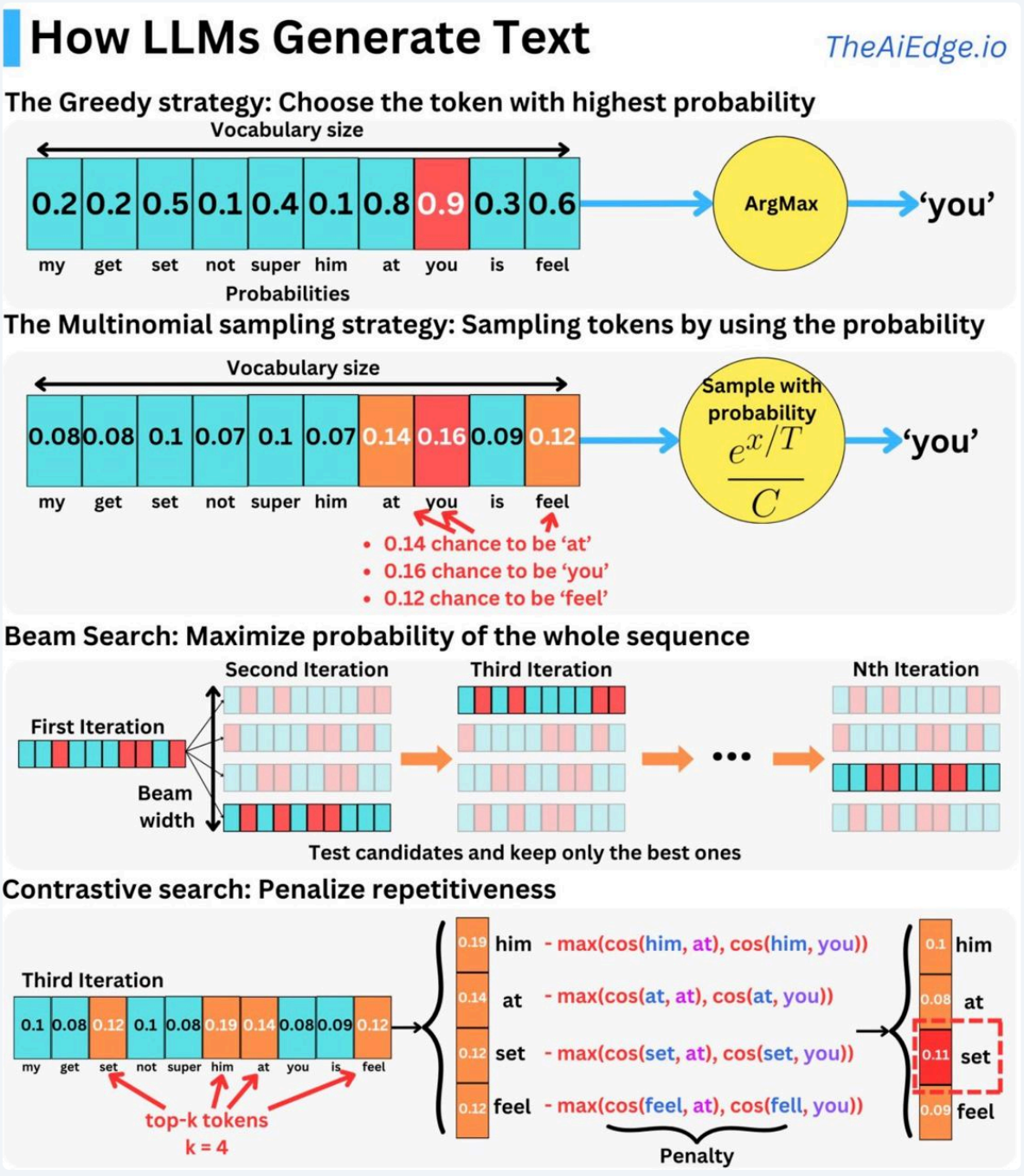
Помимо человекочитаемых данных (то, что можно назвать контентом — посты в социальных сетях, комментарии, статьи, музыкальные произведения, смешные картинки котов, собак и капибар, видео — образовательные, развлекательные, личные сообщения и т.п.), значительный вклад в общую совокупность веб-данных вносят метаданные, которые сопровождают контент и являются машиночитаемыми, а также средства обмена информацией (и данными) между программными и информационными системами — API и схожие средства коммуникации ИС–ИС.

Совокупность данных может создавать определенный след без доступа к значимому содержимому — т.н. digital fingerprinting («цифровые отпечатки пальцев»).

Можно выделить несколько обширных категорий веб-данных:

- открытые веб-данные, публичные веб-данные и т.п. — публично доступные данные, извлеченные из «открытых» веб-источников: новостных сайтов, публичных блогов, досок объявлений, форумов, агрегаторов вакансий, рецензий и т.п., вопросно-ответных сервисов («Ответы Мейл.ру»), энциклопедий (Википедия), видеохостингов, хостинги с публично доступными изображениями (Imgur) и т.п., публичные профили социальных сетей и сайтов с user-generated content (Пикабу, DTF, Reddit и т.п.); гиперссылки на другие ресурсы, цитаты ресурсов друг друга и т.п.

Лирическое отступление о том, почему качество модели падает, если их учить не на новых данных от людей, а на выводе других моделей:



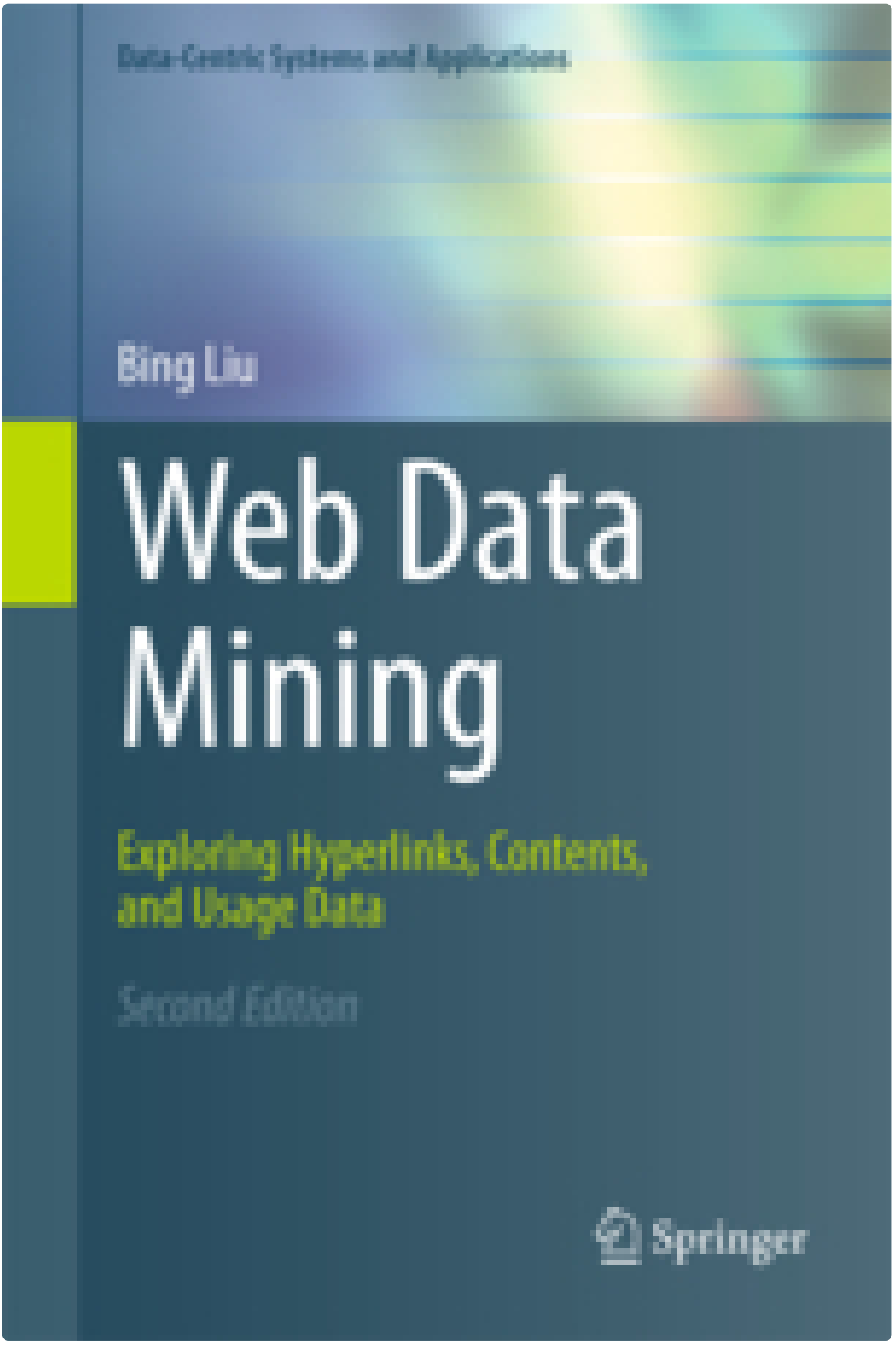
- данные из «dark web» — данные, размещаемые на веб-ресурсах, расположенных в скрытых или защищенных сетях. Как следствие, такие данные могут включать в себя информацию из нелегальных источников или полученную нелегальным путем (в результате утечек, например), в том числе это могут быть чувствительные персональные данные — номера документов (паспорта, ИНН, СНИЛС и т.п., ID в других странах и т.д.), номера банковских карт (и сопутствующую информацию, например, CVV-коды), адреса и иные чувствительные для частной жизни сведения
- данные из «глубинной» (термин условный) сети — данные, которые не индексируются поисковыми движками напрямую. Это месседжеры (Telegram, Viber, IRC), которые де-факто поддерживают свою собственную сеть поверх Интернета, работая по своим протоколам (например, MTProto у Telegram), в нее входят только клиенты и серверы указанного сервиса; сюда же можно отнести данные, размещенные в защищенных участках веб-ресурсов (за паролями и иными средствами ограничения доступа — как пример, закрытая часть pstu.ru)

Для чего веб-данные и их интеллектуальный анализ применяются

1. Анализ медиа — какие тренды в массовом сознании транслируются через медиа-ресурсы
2. Анализ рисков — поиск конкурентов, оценка риска нарушения логистических цепочек, отслеживание нарушений интеллектуальной собственности и т.п.
3. Финансовый анализ
4. Научный анализ (проверка гипотез на больших выборках, например, получение массивов архивных данных для анализа и т.п., а также краудсорсинг — например, сбор и анализ публично доступных данных с любительских телескопов)
5. Машинное обучение — сбор данных для подготовки датасетов для обучения моделей сверхвысоких размерностей (большие языковые модели, генеративные модели изображений, аудио, видео, большие мультимодальные модели; а также более специализированные модели, требующие для повышения точности значительное число источников — например, модели предсказания погоды)

Что и где почитать по теме

- Bing Liu https://www.cs.uic.edu/~liub/



- https://e.lanbook.com/book/108129

- https://e.lanbook.com/book/348086

- Сапрыкин, О. Н. Интеллектуальный анализ данных : учебное пособие https://e.lanbook.com/book/188906.