Кластерный и интеллектуальный анализ текстовой информации. Основные понятия и проблемы

К. А. Изофатов

Саратовский государственный технический университет

Доклад посвящен проблеме кластерного анализа текстовой информации. Рассмотрены основные направления исследований, в основе которых лежит семантический анализ текста и методы решения задач кластеризации.

Широкое применение информационных систем ведет к росту объемов информации и повышает необходимость в использовании аналитических систем вместо человеческих ресурсов для извлечения знаний из накопленной информации, делая актуальной задачу разработки специализированных методик и программных инструментов.

Для исследования структурированных массивов информации используется метод анализа фактографических данных, в котором выделены шесть различных задач, такие как: классификация, регрессия, кластеризация, выявление ассоциаций, выявление последовательностей, прогнозирование [1].

В настоящее время существует множество методов, с помощью которых решаются задачи классификации и кластеризации текстов. На их основе реализовано несколько систем, использующих семантическую обработку текстов. Например, система KONSPEKT, Интернет-порталы "Инновационное развитие регионов" (проект) и «Новотека», поисковая система SHOE.

Кластерный анализ занимает одно из центральных мест среди методов анализа данных и представляет собой совокупность методов, подходов и процедур, разработанных для решения проблемы формирования однородных классов в произвольной проблемной области. Зачастую проблемная область представляет собой огромный массив текстовой информации, что делает

невозможным его кластеризацию с помощью экспертов. Помимо этого, экспертная разбивка текстов на кластеры может быть субъективной и отражать лишь мнение конкретного эксперта [2].

В общем случае задача кластеризации текста распадается на две:

- техническая задача его преобразования в некоторую матричную, векторную или любую другу модель;
- математическая задача его кластеризации.

Можно выделить следующие задачи кластеризации:

- понимание данных путём выявления кластерной структуры. Разбиение выборки на группы схожих объектов позволяет упростить дальнейшую обработку данных и принятие решений, применяя к каждому кластеру свой метод анализа (стратегия «разделяй и властвуй»);
- сжатие данных. Если исходная выборка избыточна, можно сократить её, оставив по одному типичному представителю от каждого кластера;
- обнаружение новизны. Выделяются нетипичные объекты, которые не удаётся присоединить ни к одному из кластеров [4].

Решение задачи кластеризации выдвигает ряд следующих требований к алгоритму:

- отсутствие обучающей выборки;
- применимость сильно сгруппированных данных.

Перейдем к определению интеллектуального анализа данных и рассмотрению его задач. Интеллектуальный анализ данных – область знаний, относящаяся к обработке данных, изучающая поиск и описание скрытых, нетривиальных и практически полезных закономерностей.

К задачам интеллектуального анализа данных относится множество направлений, такие как поиск документов в локальных и глобальных сетях, сортировка и классификация документов, автоматическое аннотирование и реферирование, системы автоматического контроля, вопросно-ответные

системы, диалоговые системы, обучающие и обучающиеся системы, модификация и пополнение баз знаний, экспертные системы и машинный перевод. Однако в данной статье мы рассмотрим лишь некоторые из них.

Методы интеллектуального анализа данных опираются на математический аппарат классической теории множеств, теории нечетких множеств, математической статистики, нейронных сетей, а также разнообразные эмпирические методики. Алгоритмическое решение формализованной задачи интеллектуального анализа данных связано с задачами поиска экстремума целевой функции и вида целевой зависимости [3].

Перед применением какого-либо алгоритма набор текстовых документов надо преобразовать в более удобный вид. Существует две наиболее распространеных модели представления текста, это триграмная модель и модель-термин.

Известно большое число методов кластеризации, которые делятся на иерархические и неиерархические, среди которых наибольшее распространение получили методы разбиения. Наиболее известными методами кластеризации являются: EM-алгоритм, статистические алгоритмы кластеризации (K-средних), графовые алгоритмы кластеризации, алгоритмы семейства FOREL, иерархическая кластеризация или таксономия, нейронная сеть Кохонена, ансамбль кластеризаторов, алгоритмы семейства KRAB и алгоритмы, основанные на методе просеивания.

Список использованных источников

1. Нейский, И. М. Методика адаптивной кластеризации фактографических данных на базе Fuzzy C-means и MST / И. М. Нейский [Электронный ресурс]. - Режим доступа: http://www.philippovich.ru/Persons/Neyskiy/Avtoreferat_Neiskiy.pdf
2. Корунова, Н. В. Кластеризация документов проектного репозитария на основе нейронной сети кохонена / Н. В. Корунова [Электронный ресурс]. - Режим доступа: http://nsmv2008.ulstu.ru/docs/klasterzacij_dokumentov.pdf

3. Елизаров, С. И. Разработка и Исследование методов и алгоритмов кластеризации для систем анализа данных / С. И. Елизаров [Электронный ресурс]. - Режим доступа: http://www.eltech.ru/education/aspir/SIElizarov.doc
4. Кластерный анализ [Электронный ресурс]. - Режим доступа: http://ru.wikipedia.org/wiki/Кластерный_анализ