Секция «Вычислительная математика и кибернетика»

О методе предварительной кластеризации текстовых коллекций большого объема на основе онтологии.

Карпов Илья Андреевич
Аспирант
Московский государственный открытый университет, Факультет информатики и радиоэлектроники, Москва, Россия
E-mail: karilan@yandex.ru

В настоящее время актуальной является задача кластеризации и классификации больших массивов текстовых документов. При этом ряд задач текстового анализа не требует реализации кластеризации в классическом ее понимании – группировки объектов по произвольному признаку, имея цель сгруппировать документы по заранее не заданной тематической близости или общему смысловому контексту. Задачи такого типа не требуют полного сопоставления всех элементов массива, но предполагают наличие у системы знаний о предметной области или онтологии.

В качестве онтологии предлагается использовать семантичесую сеть, построенную на основе информации Wikipedia: узлами сети являются статьи энциклопедии, а ребрами – связи между близкими статьями. Критерием близости двух статей может выступать TF/IDF мера. Подобная сеть имеет хорошее покрытие ключевых терминов документа и является регулярно пополняемой.

В ходе обработки документа по каждому абзацу строится постсинтаксический граф [5], узлам которого соответствуют все фигурирующие в предложении сущности – наименования предметов и лиц, действий и признаков. Связь между двумя узлами такого графа описывает синтаксическое подчинение слов в предложении, нагруженное семантической характеристикой [5]. В качестве вершин используются термины, вошедшие в онтологию, или их синонимы, лексический материал, не совпавший с единицами онтологии, отбрасывается [1]. Подход разбиения документа на абзацы и их последующий независимый анализ позволяют упростить лингвистический анализ текста без потерь точности [3].

Полученные графы вектора накладываются на семантическую сеть по принципу, подобному муравьиным алгоритмам. Узлы сети совпадают с узлами графа и имеют накопительную функцию. Связи сети делятся на два типа – полученные при построении онтологии (статические) и добавленные при наложении графа на семантическую сеть (динамические). Вес статических связей вычисляется при построении онтологии и не меняется в ходе работы. Вес динамических связей имеет накопительную функцию и изменяется обратно пропорционально увеличению количества документов в системе. Каждое новое вхождение семантического отношения увеличивает значение функции.

В результате каждому абзацу ставится в соответствие несколько вершин сети, образующих одну или несколько связанных областей. Критерием формирования нового кластера является достижение группой связанных вершин порогового значения накопительной функции. Для построения пороговой функции использовался метод, описанный в [2].

Описанный метод позволяет проводить предварительную кластеризацию текстовых коллекций большого объема, реализуя чтение “крупным взглядом”. Вычислительная

сложность метода *O(n*log(m))* , где n – число документов в коллекции, m - число узлов семантической сети. Первичное построение и хранение семантической сети требует значительных ресурсов что делает метод оправданным только для коллекций большего объема.

Литература

1. Grineva M., Grinev M., Lizorkin D. Extracting Key Terms From Noisy and Multi-theme Documents // IW3C2. 2009

2. Kleinberg J.M. Bursty and hierarchical structure in streams // Data Mining and Knowledge Discovery. 2003

3. Зевайкин А. Н. Об одном подходе к кластеризации текстовых сообщений с разбиением на абзацы // Информационные технологии 2005. №5. с.16

4. Леонтьева Н.Н. Автоматическое понимание текстов – системы, модели, ресурсы. ACADEMIA. 2006

5. Сокирко А. В. Семантические словари в автоматической обработке текста : По материалам системы ДИАЛИНГ // диссертация к.т.н. Москва. 2001.