МОДУЛЬ ПРЕДВАРИТЕЛЬНОГО АНАЛИЗА ТЕКСТОВ

Дудник А.В., Евсина Н.А., Клевцова Е.В.
Національний технічний університет «Харківський політехнічний інститут», вул. Кирпичова 2, Харків, Україна, 61002

При решении сложных современных задач управления нередко приходится выполнять вспомогательные подзадачи, такие как распознавание образов или анализ текстов. В данной работе рассматривается алгоритм, применяемый при кластеризации/классификации текстов. Особенностью такого рода задач является расхождение между методами анализа и объектом анализа. Компьютерные методы анализа базируются на формальной логике, в то время как тексты во многом представляют собой живую человеческую речь, с её метафорами, иронией, иносказательностью и пр. По этой причине от непосредственного анализа смысла текста отказываются и прибегают к обработке формальных признаков. Например, анализируют набор слов, входящих в текст, игнорируя смысловую связь между ними. Каждый текст порождает своё множество слов. Чем ближе тексты по тематике, тем сильнее множества пересекаются, и наоборот. Однако, при анализе больший текстов, такой метод чрезвычайно требователен к памяти и вычислительным возможностям системы.

Сократить количество необходимых для анализа слов, и тем самым уменьшить требования к памяти системы позволяет применение алгоритмов латентно-семантического анализа (LSA). В данной работе получили дальнейшее развитие идеи, изложенные в [1]. Основой рассматриваемого алгоритма является сингулярное разложение матрицы коэффициентов вхождения слов в различные тексты, т.н. матрица слов/документов.

Предварительно был составлен корпус текстов. Для этого было отобрано 10 коротких текстов от 500 до 1000 слов. Тексты отбирались по принадлежности к 5 темам (история, медицина, фауна, мода, астрономия), т.е. каждая из тем представлена двумя текстами. На предварительном этапе все тексты были трансформированы в множества слов. При этом вспомогательные слова (предлоги, союзы и т.д.) были удалены, а основные слова были отмечены процентом вхождений в каждый текст. Это позволило сократить обрабатываемые массивы слов на 45–50% (табл. 1).

Процент наиболее повторяющихся слов колеблется от 1,4 до 4,82. Дальнейшее развитие этого наблюдения позволило построить модуль предварительной фильтрации для выбраковки текстов, составленных генераторами слов.

Особый интерес представляет процедура удаления вспомогательных слов. В данном случае был составлен массив из этих слов по принципу орфографического словаря, с учётом разных форм. Подобный массив допускает расширение и удобен для редактирования.

При построении LSA-алгоритмов рекомендуется исключать из рассмотрения слова, встречающиеся в тексте не более 1-го раза. Следует отметить, что слова, повторяющиеся в тексте 2 раза и более, составляют от 10 до 40% оставшихся после предварительной обработки слов. Таким образом, количество анализируемых слов по сравнению с исходным может быть уменьшено в разы.

Таблица 1 – Результаты первичной обработки текстов

Заголовок текста	**Количество слов**		**Процент наиболее часто повторяющегося слова**
	исходное	**после обработки**	
Первая мировая война	871	466	2,36
Международные отношения в начале XX века	839	482	1,87
Атеросклероз сосудов: что это такое, причины и стадии развития	560	261	3,83
Ишемическая болезнь сердца	759	384	3,39
Животный мир Африки	869	456	1,88
Млекопитающие Африки	849	391	1,40

Далее формируется матрица слов/текстов, каждый столбец которой соответствует тексту, а строка — слову. В каждый из элементов массива записывается процентное значение вхождения слова в каждый из текстов. Полученная матрица подвергается сингулярному разложению, в результате которого получаем 3 матрицы, которые принято интерпретировать следующим образом: темы, слова/темы, темы/документы. Величина сингулярного числа в матрице тем позволяет отобрать темы, наиболее выраженные в корпусе текстов. Так же по мере соответствия сингулярным значениям можно осуществить кластеризацию документов.

Исследования были выполнены в среде MATLAB. Дальнейшее направление исследований видится в организации искусственной нейронной сети для кластеризации текстов.

Список литературы

1. Алгоритм LSA для поиска похожих документов. // [Электронный ресурс] –URL: https://netpeak.net/ru/blog/algoritm-lsa-dlya-poiska-pohozhih-dokumentov/

2. Анализ данных и процессов: учебное пособие / А.А. Барсегян, М.С. Куприянов, И.И. Холод и др. – СПб.: БХВ-Петербург, 2009. – 512 с.