Методы и средства тематического поиска в коллекциях научных текстов

Савостин Петр Алексеевич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: petersavostin@gmail.com

В последние несколько лет, многие пользователи, находящиеся в состоянии поиска научных материалов, испытывают трудности, которые могут возникать по многим проблемам. Например, информационная перегрузка или особенности реализации поисковой машины, могут быть причиной выдачи нерелевантной информации. Основная цель данной работы заключается в создании системы научного поиска и мониторинга, которая позволяла бы решить данную проблему. Существуют следующие подходы к поиску научной информации: поиск по ключевым словам, поиск по образцу, поиск по теме. Каждый подход влечет за собой использование соответствующего метода.

Поиск по ключевым словам осуществляется по поисковому запросу, заданному пользователем. В данном подходе документ рассматривается, как неупорядоченное множество слов. Каждому слову приписывается вес, который можно вычислить с помощью весовых функций (TF-IDF, BM-25), зависящих, в большинстве случаев, от частоты появления слова в документе, количества слов в документе [1]. Далее каждый документ и запрос представляются в виде вектора слов. Заключающим этапом является вычисление степени схожести между документами и запросом с помощью специальных мер (косинусная мера, коэффициент Дайса).

В поиске по образцу пользователь может задать запрос несколькими способами, например, фрагмент текста, список литературы или целый документ. В зависимости от типа запроса можно осуществлять поиск разными методами. Если был задан список литературы, то можно искать документы с похожим списком, а также можно просто искать документы из списка. В случае использования целого документа в качестве запроса, можно воспользоваться методами классификации.

Разбиение документов на тематические группы предполагает использование методов кластеризации и латентно-семантического анализа. Очевидно, что близость того или иного конкретного документа информационным потребностям пользователя зависит от содержания, в рамках которого происходит поиск [2]. Описание со-

держания данного документа или его части является достаточно нетривиальной задачей, для решения которой используют тематико-ориентированные методы поиска. Основная цель таких методов — выявить тематическую принадлежность документов. В большинстве случаев эти методы основываются на достаточно простых предположениях: словарный запас и частота употребления слов зависят от тематики, в документе может присутствовать несколько тематик, но сама по себе задача тематического поиска достаточно нетривиальна. Латентно-семантический анализ позволяет выделить скрытые зависимости в множестве документов. Пусть есть коллекция документов, которая представляется в виде сопоставления слов из словаря коллекции количеству совпадений в определенном документе, назовем эту матрицу «документы и слова». Метод ЛСА подразумевает разложение этой матрицы на произведение двух матриц: «темы и слова», «документы и темы». Стоит заметить, что так как простой латентно-семантический анализ, по сути, не решает задачу корректно из-за того, что количество разложений матрицы «документы и слова» бесконечно много, вместо классического подхода будет использоваться метод аддитивной регуляризации тематических моделей, который позволит путем наложения комбинации ограничений добиться качественного решения задачи.

В работе предлагается комплексный метод решения задачи поиска научной информации на основе алгоритмов классификации, кластеризации, латентно-семантического анализа с комбинацией регуляризаторов [3].

Литература

1. Cummins. R T. The Evolution and Analysis of Term-Weighting Schemes in Information Retrieval, National University of Ireland, Galway, 2008.
2. Некрестьянов И.С. Тематико - ориентированные методы информационного поиска: Диссертационная работа к.т.н.: 05.13.11 / Санкт-Петербургский государственный университет - СПб., 2000.
3. Воронцов К.В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014. Т. 455, №3. С 268-278.