И. А. Мазура

АНАЛИЗ ТЕКСТА МЕТОДОМ ИЗВЛЕЧЕНИЯ КЛЮЧЕВЫХ СЛОВ

Исследование анализа текста методом извлечения ключевых слов включает обзор известных решений, и поиск новых алгоритмов. Актуален анализ методов, оптимизирующий существующие алгоритмы.

Введение

В условиях постоянно увеличивающегося потока информации и появления все новых текстовых массивов, на первый план выходит проблема систематизации текстов в информационном пространстве и оптимизация их поиска.

Понимание текста имеет своей целью извлечение смысла текста. Предполагается множественность трактовок, возникающих в процессе восприятия и интерпретации. Значение отдельных элементов текста для выражения общего смысла неодинаково, и наряду с центральными элементами имеются также и второстепенные элементы текста. Актуальным на данный момент является вопрос о составлении объективной методики выделения ключевых слов из текста.

I. Общие сведения о задаче извлечения ключевых слов из текста

Извлечение ключевых слов (также известное как обнаружение ключевых слов или анализ ключевых слов) – это метод анализа текста, который состоит из автоматического извлечения наиболее важных слов и выражений в тексте. Это помогает суммировать содержание текста и распознать основные темы, которые обсуждаются. При изучении терминов, представляющих наиболее релевантную информацию, содержащуюся в документе, используется различная терминология: ключевые фразы, ключевые термины или просто ключевые слова.

II. Методы извлечения ключевых слов из текста

Методы извлечения ключевых слов из текста условно разделяются на две категории:

– назначение ключевых слов;

– извлечение ключевых слов.

Оба эти метода сводятся к одной и той же проблеме – выбора лучшего ключевого слова. При назначении ключевые слова выбираются из контролируемого словаря терминов или предопределенной таксономии, а документы классифицируются по классам в соответствии с их содержанием. Извлечение ключевых слов обогащает документ ключевыми словами, которые явно упоминаются в тексте. Слова анализируются с целью выявления наиболее репрезентативных из них, обычно исследуя два свойства источника.

Чаще всего при извлечении ключевых слов не используется заранее определенный тезаурус для определения ключевых слов [1]. Существующие подходы автоматического извлечения ключевых слов могут быть разделены на:

– простые статистические подходы;

– лингвистические подходы;

– подходы машинного обучения и т. д.

Общая схема применения алгоритма представлена на рисунке 1.

Рис. 1 – Общая схема алгоритма извлечения ключевых слов

Выводы

С ростом количества доступных цифровых документов, автоматическое извлечение ключевых слов становится сутью исследований. Традиционно, эта задача решается для работы с отдельными документами. Этот алгоритм можно использовать для более сложных вопросов. С увеличением потока данных построение эффективной модели представления текста становится еще более актуальным и требовательным одновременно.

Методы автоматического извлечения ключевых слов из текста должны решить проблемы масштабируемости и разреженности. Исследуемые новые решения, в первую очередь, решают эти задачи.

1. Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов/ Н. А. Астраханцев// Труды Института системного программирования РАН. – 2014. – №4. – С. 155-165.

Мазура Ирина Александровна, магистрант кафедры информационных технологий автоматизированных систем БГУИР, irynamazura22@gmail.com.

Научный руководитель: Гуринович Алевтина Борисовна, доцент кафедры вычислительных методов и программирования БГУИР, gurinovich@bsuir.by.