АЛГОРИТМ ИЗВЛЕЧЕНИЯ КЛЮЧЕВЫХ СЛОВ

Мазура И. А., Гуринович А. Б.
Кафедра информационных технологий автоматизированных систем, Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: irynamazura22@gmail.com, gurinovich@bsuir.by

Исследование алгоритмов извлечения ключевых слов включает в себя как исследование известных решений, так и поиск новых алгоритмов. Актуален анализ методов, позволяющих оптимизировать существующие алгоритмы.

Введение

Во многих документах и научных работах зачастую содержится большое количество текста, которое не несет существенной информации. Людям, у которых нет времени на прочтение полного текста, нужно ознакомиться с кратким содержанием.

Таким образом, появляется необходимость сокращать объем документа, выделяя наиболее значимую часть текста, называемую рефератом. Ручное реферирование является сложной и рутинной работой. Данная задача требует дополнительных сотрудников, поэтому целесообразно использовать системы автоматического извлечения ключевых слов из текста.

Наборы назначенных вручную или автоматически выделенных ключевых слов и словосочетаний из текста используются для формирования у пользователя общего представления о содержании текста. Необходимо различать два основных подхода к решению проблемы автоматизации выделения ключевых слов и словосочетаний и их извлечение. Главное отличие заключается в том, что первый подход позволяет выделять только те ключевые слова и словосочетания, которые содержатся в некотором предусмотренном словаре, а второй подход предполагает выбор ключевой информации непосредственно из текста.

В статьях ключевые слова формируют важный компонент, поскольку они обеспечивают краткое представления о содержании. Ключевые слова также играют важную роль в поиске статьи из информационно-поисковых систем и для поисковой оптимизации. Традиционные подходы к извлечению ключевых слов предполагают ручное присвоение ключевых слов на основе содержания статьи и суждений авторов. Это требует много времени и усилий, а также может быть неточным с точки зрения выбора соответствующих ключевых слов.

Так как человеческая речь иногда достаточно непредсказуема и не укладывается в четкие границы заранее прописанный правил, то данная задача должна иметь нетривиальное решение.

В работе представлена общая схема решения подобных задач. Конкретные реализации данной схемы могут быть абсолютно разными, однако условные основные шаги будут одинаковыми.

I. Ключевые слова и их функции

В общенаучном плане ключевые слова рассматриваются как наиболее важный, существенный для понимания смысловой элемент. Несмотря на разнообразие многосторонних подходов, которые уже были выведены, к толкованию ключевых слов, абсолютно объективным представляется то, что ключевое слово определяет содержание всего текста и является носителем его основного смысла. Учитывая это, следует признать, что ключевые слова – это слова, наиболее значимые и существенные для понимания всего содержания текста. Все основные признаки ключевых слов приведены на рисунке 1.

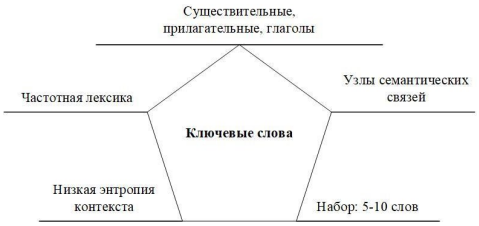

Рис. 1 – Основные признаки ключевых слов

Функции ключевых слов:

- ключевые слова формируют смысл текста и обеспечивают хранение его в памяти;
- ключевые слова служат созданию структурно-семантического единства текста, его целостности;
- ключевые слова задают функциональный стиль развертываемого текста (при развертывании сжатого текста с опорой на ключевые слова сохраняется стиль исходного текста);
- анализ ключевых слов позволяет извлечь целостность содержания, которое является относительно неизменным и независимым от интерпретатора;
- логически упорядоченные ключевые слова передают обобщенное содержание текста, указывают и ограничивают направление ассоциаций читателя [1].

II. Алгоритмы извлечения ключевых слов

1. TF-IDF
   Для того чтобы оценить повторяемость слов в тексте, недостаточно просто подсчета слов,полученного от обычного счетчика. Это объясняется тем, что большое количества определенных общих слов в корпусе [2]. Это можно преодолеть при помощи векторизатора TF-IDF, который наказывает слова, повторяющиеся несколько раз в тексте. TF-IDF – это оценка частоты слов, которые выделяют слова более важные для контекста, а не те слова, которые просто часто появляются в документе.
2. RAKE
   Rake (Rapid Automatic Keyword Extraction) – сравнительно эффективный алгоритм, который применяется для одиночных документов. Он может быть использован для анализа документов из различных предметных областей. Кроме того, этот алгоритм никак не зависит от структуры документа. Данный алгоритм основывается на том, что в ключевых словах редко встречаются шумовые слова и знаки препинания. Нужно также отметить, что такие слова как правило считаются не информативными и не участвуют в индексации в системах информационного поиска. Для работы алгоритма потребуется список таких слов, которые называются стоп-словами [3].
3. TextRank
   TextRank – это алгоритм, основанный на PageRank, который часто используется для извлечения ключевых слов и суммаризации текста. PageRank – это алгоритм, используемый для расчета веса веб-страниц [4]. Например, можно представить все веб-страницы в виде большого направленного графа. В графе каждый узел – это веб-страница. Если веб-страница A имеет ссылку на веб-страницу E, то эта связь может быть представлена в виде направленного ребра от A до E.
4. Ансамбль методов
   Можно использовать сразу несколько алгоритмов, а их результаты обрабатывать в соответствии с разработанными дополнительными алгоритмами.
   Метод ансамбля будет состоять из следующих этапов:
   - получение списка ключевых слов-кандидатов от каждого автоматического экстрактора ключевых слов;
   - фильтрация списка ключевых слов-кандидатов;
   - объединение и пересчет баллов ключевых слов-кандидатов;
   - применение динамической пороговой функции для извлечения ключевых слов.

   Общая схема для ансамбля двух алгоритмов представлена на рисунке 2.

Рис. 2 – Общая схема ансамбля двух алгоритмов

III. Заключение

Понимание текста имеет своей целью проникновение в смысл текста. Смысл текста формируется на основе свернутой, явно невыраженной смысловой структуры, выводимой в соответствии с объемом лингвистических, энциклопедических, фоновых знаний человека, получающего эту информацию, его опыта и принадлежности к той или иной культурной общности. Потому предполагается множественность трактовок, возникающих в процессе восприятия и интерпретации. Эта структура представляет собой сжатое смысловое ядро текста, извлеченное из его ключевых слов, их ассоциативных связей и импликаций, которое должно быть сохранено при всех изменениях содержания интерпретаторами.

Существенной проблемой является и то, какие именно единицы текста являются решающими для понимания всего текста. Значение отдельных элементов текста для выражения общего смысла неодинаково, и наряду с центральными имеются также и второстепенные элементы текста. При изменении центральных элементов смысл текста меняется. Если изменить второстепенные единицы текста, то общий смысл может сохраниться. Актуальным на данный момент является вопрос о составлении объективной методики извлечения ключевых слов.

1. Методы выделения ключевых слов при реферировании научного текста [Электронный ресурс]. – Режим доступа: https://vestnik.tspu.edu.ru/files/vestnik/PDF/articles/moskvitina_t._n._45_50_8_197_2018.pdf.
2. TF-IDF [Электронный ресурс]. - Режим доступа: https://ru.wikipedia.org/wiki/TF-IDF.
3. Automatic Keyword Extraction from Individual Documents [Электронный ресурс]. – Режим доступа: https://www.researchgate.net/publication/227988510 _Automatic_Keyword_Extraction_-d-from_Individual_Documents.
4. Understand TextRank for Keyword Extraction by Python [Электронный ресурс]. – Режим доступа: https://towardsdatascience.com/textrank-for-keyword-extraction-by-python-c0bae21bcec0.