Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.021:81.367.6

МАЗУРА
Ирина Александровна

Алгоритм извлечения ключевых слов

АВТОРЕФЕРАТ
на соискание степени магистра
по специальности 1-40 80 02
«Системный анализ, управление и обработка информации (по отраслям)»

Научный руководитель
Гуринович Алевтина Борисовна
Доцент кафедры ИТАС,
кандидат физико-математических наук

Минск, 2022

ВВЕДЕНИЕ

Многие документы содержат большое количество текста не несущего существенной информации. Например, новостные, научные и обзорные статьи. Для экономии времени необходимо ознакомиться с кратким содержанием публикации или издания.

Правильно сформированная краткая аннотация позволяет сделать вывод о необходимости подробного изучения полной версии текста.

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.

За ключевые слова и словосочетания чаще всего принимаются структурные единицы текста с наиболее важной информацией о содержании текста.

Различают два основных подхода к решению проблемы автоматизации выделения ключевых слов и словосочетаний (keyphrase assignment) и их извлечение (keyphrase extraction). Основное различие состоит в том, что первый метод позволят выделять только те ключевые слова и словосочетания, которые содержатся в некотором предусмотренном словаре, а второй метод предполагает выбор ключевой информации непосредственно из текста.

Основные функции ключевых слов:

– формируют важный компонент.

– играют важную роль для поисковой оптимизации из информационно-поисковых систем, библиографических баз данных.

– дают возможность классифицировать статью по соответствующему предмету, дисциплине или области исследования.

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.

Существует большое количество методов и алгоритмов для извлечения ключевых слов из текста. Методы отличаются процессом реализации. Но цель у них одна: извлечение достаточного количества слов из текста, которые имеют наибольший вес в данном контексте.

Так как человеческая речь и методы изложения информации достаточно непредсказуемы и не укладываются в четкие границы заранее прописанных правил, то данная задача должна иметь нетривиальное решение.

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

Предложенный в диссертационном исследовании алгоритм является универсальным. Перспективой дальнейшего улучшения данного алгоритма может стать расширяемость для других языков.

При расширяемости алгоритма в случаях языков, отличных от английского, необходимо учитывать особенности конкретного языка. Например, в русском языке у существительных имеется шесть падежей, что заметно осложняет алгоритм подсчета слов ввиду необходимости учитывать разные падежи одного и того же существительного как одно слово. Данная проблема может быть успешно решена использованием алгоритма стемминга, рассмотренного в данной работе.

Например, для японского и китайского языков необходима реализация разбиения текста на фрагменты и работы с этими фрагментами как со словами, т.к. один иероглиф означает один слог и данные иероглифы необходимо, в первую очередь, разбить на группы, обозначающие отдельные слова.

Объект исследования диссертационной работы: алгоритмы и методы решения задач извлечения ключевых слов.

Цель диссертационной работы: оптимизация алгоритмов извлечения ключевых слов.

На защиту выносится модифицированный алгоритм извлечения ключевых слов.

Результаты исследовательской работы докладывались на следующих конференциях: Информационные технологии и системы 2020 (ИТС 2020); 57-ая научная конференции аспирантов, магистрантов и студентов; Информационные технологии и системы 2021 (ИТС 2021).

Диссертационная работа выполнена полностью самостоятельно.

Результаты исследовательской работы являются оригинальными.

Работа проверена в системе «Антиплагиат». Процент оригинальности соответствует норме, установленной кафедрой информационных технологий автоматизированных систем.

Цитирования обозначены ссылками на публикации, указанные в «Списке использованных источников».

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Объект исследования – ключевые слова. Предмет исследования – извлечение ключевых слов.

Целью диссертации является разработка новых или модернизация существующих алгоритмов извлечения ключевых слов.

Задачей исследования является изучение основных существующих алгоритмов извлечения ключевых слов, определение основных методов улучшения алгоритмов, нахождение нового подхода к извлечению ключевых слов.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Общий объем магистерской диссертации составляет 51 страниц, включая 9 таблиц, 19 рисунков, библиографический список из 17 наименований, одно приложение.

Магистерская диссертация посвящена разработке новых и модернизации существующих алгоритмов извлечения ключевых слов.

Введение содержит описание проблемы, рассматриваемой в работе, дается обоснование актуальности темы магистерской диссертации.

В первой главе проведен обзор предметной области, обзор существующих алгоритмов извлечения ключевых слов, поставлена задача для диссертации.

Во второй главе рассмотрены способы оценивания качества алгоритма, проведен анализ работы алгоритмов.

В третьей главе выполнены проектирование и разработка алгоритма, представлены результаты работы оптимизированного алгоритма извлечения ключевых слов.

В четвертой главе проведен анализ результатов исследования.

В заключении кратко изложены полученные результаты.

ЗАКЛЮЧЕНИЕ

В магистерской диссертации были исследованы алгоритмы и методы выделения ключевых слов из текста. Был разработан алгоритм, показывающий большую эффективность по сравнению с изученными алгоритмами.

Проведен анализ предметной области, обобщены достоинства и недостатки существующих алгоритмов.

Основная цель исследования – алгоритм, обеспечивающий получение ключевых слов из текста с высокой точностью. Анализ работы ансамбля алгоритмов показал увеличение значения метрик по сравнению с показателями метрик работы отдельных алгоритмов.

Были рассмотрены вопросы построения выборок для статистического анализа качества работы алгоритма. Были выбраны надежные метрики, по которым можно сделать вывод об эффективности алгоритмов. Была предложена общая схема алгоритма для извлечения ключевых слов из текста.

Предложенный в диссертационном исследовании алгоритм является универсальным и перспективным.

В ближайшей перспективе предложены следующие аспекты для усовершенствования:

– улучшение (оптимизация) данного алгоритма;

– использование методов машинного обучения;

– применение алгоритма на различных языках программирования.

Также в ходе работы было разработано программное средство, для апробации и реализации алгоритма.

Примеры применения предложенного модифицированного алгоритма:

– для оптимизации хранения больших объемов текста;

– для классификации текстов на определенные категории;

– для ускорения обработки большого количества текстов.

По результатам статистического анализа работы модифицированный алгоритм имеет наибольшее количество правильно извлеченных ключевых слов.

Результаты исследовательской работы докладывались на следующих конференциях: Информационные технологии и системы 2020 (ИТС 2020); 57-ая научная конференции аспирантов, магистрантов и студентов; Информационные технологии и системы 2021 (ИТС 2021).

Проведена апробация модернизированного алгоритма.

Результаты исследования являются эффективным методом извлечения ключевых слов, что позволило успешно внедрить его в поисковую систему обучающего веб-приложения, о чем свидетельствует акт о внедрении, предоставленный EPAM Systems.

Цели диссертационного исследования полностью достигнуты.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

[1] И.А. Мазура, А.Б. Гуринович Алгоритм извлечения ключевых слов // Информационные технологии и системы 2020 (ИТС 2020) = Information Technologies and Systems 2020 (ITS 2020) : материалы международной научной конференции, Минск, 2020.

[2] И.А. Мазура Анализ текста методом извлечения ключевых слов // Информационные технологии и управление : материалы 57-ой научной конференции аспирантов, магистрантов и студентов по направлению 2, Минск, 2021.

[3] И.А. Мазура, А.Б. Гуринович Извлечение ключевых слов: графоорентированный подход // Информационные технологии и системы 2021 (ИТС 2021) = Information Technologies and Systems 2021 (ITS 2021) : материалы международной научной конференции, Минск, 2021.