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

УДК 681.3(075.8)

Кирьянов
Егор Сергеевич

Разбиение текста на семантические страницы

АВТОРЕФЕРАТ
на соискание степени магистра технических наук
по специальности 1-40 80 02 – «Системный анализ, управление и обработка информации»

Научный руководитель

Герман Олег Витольдович
канд. технических наук, доцент

Минск, 2014

Библиотека БГУИР

КРАТКОЕ ВВЕДЕНИЕ

С развитием вычислительной техники и технологий в области Интернет-услуг, с увеличением числа хранящихся и обрабатываемых документов, с увеличением способов копирования и распространения данных - все большую значимость принимают сложные механизмы поиска ответов и другие современные информационные технологии.

Сложность современного информационного поиска постоянно растёт. В связи с этим, требования к эффективности алгоритмов обработки информации также увеличиваются. Одним из самых перспективных способов обработки информации могут быть семантические сети, которые предоставляют возможности по эффективному поиску и анализу данных как человеком так и программным обеспечением.

Однако применение различных механизмов поиска затрудняет наличие дублирующейся информации в виде однотипных, одинаковых, повторяющихся документов, что является следствием неграмотного использования ресурсов компьютерной техники.

Главным распространителем дублированной информации является Интернет. Более 30-40% документов в Интернете имеют дубликаты [1, 2], различного происхождения и степени схожести. Актуальность проблемы дублирования для поисковых систем определяется увеличением индексных баз за счет избыточной информации, что в свою очередь ухудшает качество ответа на запрос пользователя, увеличивает затраты на обслуживание и хранение данных, требует большую ресурсную подготовку.

На текущий момент существует ряд способов борьбы с избыточной информацией, например использование эвристических алгоритмов для устранения дублей из результатов поиска по документам. Однако не в каждой ситуации устранение дублей будет лучшим решением на запрос пользователя поисковой системы.

Одним из способов улучшения систем обработки документов является представление текста в виде семантической сети, в частности декомпозиция на смысловые блоки - семантические страницы.

Смысловые страницы представляют собой законченные по смыслу и логике фрагменты текста, освещающие некоторый круг понятий. Они необходимы для фиксации поисковой машины или пользователя на наиболее важных составляющих документа.

Задача разбиения на страницы сложна тем, что их границы заранее не известны, не четки, а круг основных понятий априори не определен.

Библиотека БГУИР

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Объектом исследования диссертационной работы являются текстовые базы знаний в виде документов.

Предметом исследования диссертационной работы является метод выделения смысловых страниц.

Цель диссертационной работы состоит в разработке методики и алгоритма разбиения текста на смысловые страницы, используя механизм поиска ключевых слов.

Задачи исследования:

- изучение применения семантических сетей в системах обработки текстов;
- рассмотрение методов анализа текстовых документов;
- разработка собственного оригинального метода разбиения текста на смысловые страницы.

Результатом применения разрабатываемого метода на текстовых базах знаний должны представлять собой законченные по смыслу и логике фрагменты текста, освещающие некоторый круг понятий.

Практическая ценность работы состоит в том, что предложенные методы позволяют повысить эффективность распознавания документов.

Полученные в диссертации результаты могут быть использованы для дальнейшего решения теоретических и практических задач в области обработки текстовых баз знаний с учетом применения различных усовершенствованных подходов, учитывающих современные способы анализа текста.

В первой главе рассматривается актуальная проблема информационного поиска, которая является большой междисциплинарной областью науки, стоящей на пересечении когнитивной психологии, информатики, информационного дизайна, лингвистики, семиотики и библиотечного дела.

Во второй главе раскрываются существующие теоретические методы анализа текстовых баз данных. Детально рассмотрены синтаксические, лексические и статистические методы. Приведены такие основные метрики подобия текстовых документов, как косинус угла между векторами, коэффициент перекрытия, Дайса и Джаккарда.

В третьей главе содержится описание работы оригинального алгоритма. Каждый подраздел отражает отдельный этап работы алгоритма. Алгоритм представлен такими этапами, как поиск ключевых слов в исходном текстовом документе, установление статистической взаимосвязи ключевых слов в тексте, кластеризация исходного текста на основе эвристического алгоритма.

Библиотека БГУИР

В четвертой главе разработанный алгоритм применен на практике, проведен анализ качества работы алгоритма. Приведены окончательные выводы по полученным результатам.

Библиотека БГУИР

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Диссертация состоит из введения, четырех разделов, заключения, библиографического списка, приложений.

Во введении магистерской диссертации обоснована актуальность темы по требованиям к эффективности алгоритмов обработки информации.

В первой главе рассматривается актуальная проблема информационного поиска, которая является большой междисциплинарной областью науки, стоящей на пересечении когнитивной психологии, информатики, информационного дизайна, лингвистики, семиотики и библиотечного дела.

Во второй главе раскрываются существующие теоретические методы анализа текстовых баз данных. Детально рассмотрены синтаксические, лексические и статистические методы. Приведены такие основные метрики подобия текстовых документов, как косинус угла между векторами, коэффициент перекрытия, Дайса и Джаккарда.

В третьей главе содержится описание работы оригинального алгоритма. Каждый подраздел отражает отдельный этап работы алгоритма. Алгоритм представлен такими этапами, как поиск ключевых слов в исходном текстовом документе, установление статистической взаимосвязи ключевых слов в тексте, кластеризация исходного текста на основе эвристического алгоритма.

В четвертой главе разработанный алгоритм применен на практике, проведен анализ качества работы алгоритма. Приведены окончательные выводы по полученным результатам.

В заключении приводятся результаты магистерской диссертации и строятся выводы об эффективности алгоритма разбиения текста на семантические страницы. Рассматриваются области применения разработанного алгоритма для дальнейшего использования.

Библиотека БГУИР

ЗАКЛЮЧЕНИЕ

Проведенные в рамках диссертации теоретические исследования позволили получить результаты, имеющие практическое и научное значение.

Был разработан алгоритм разбиения текста на семантические страницы. В ходе реализации алгоритма были проведены его испытания с использованием проверочного текста, который показал высокие результаты качества работы алгоритма. Был проведен анализ результатов и были рассмотрены пути улучшения работы алгоритма. Было изучено применения семантических сетей в системах обработки текстов, были рассмотрены методы анализа текстовых документов.

Результатом применения разрабатываемого метода на текстовых базах знаний получились законченные по смыслу и логике фрагменты текста, освещающие некоторый круг понятий.

Полученные в диссертации результаты могут быть использованы для дальнейшего решения теоретических и практических задач в области обработки текстовых баз знаний с учетом применения различных усовершенствованных подходов, учитывающих современные способы анализа текста.

Библиотека БГУИР

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1-А. Кирьянов, Е.С. Разбиение текста на семантические страницы / Е.С. Кирьянов, О.В. Герман // 50-я научная конференция аспирантов, магистрантов и студентов. Информационные технологии и управление: Тезисы докл. – Минск : БГУИР, 2014 – С.49.

Библиотека БГУИР