МЕТОДЫ СОЗДАНИЯ АВТОРЕФЕРАТА РАБОТ ОДНОГО АВТОРА

Третьяков Ф. И., Серебряная Л. В.
Кафедра программного обеспечения информационных технологий, Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: Fiodor.Tretyakov@gmail.com, l_silver@mail.ru

Задача извлечения смысла из текста является одной из важнейших при работе с текстовой информацией. Автореферирование составляет подкласс таких задач. Одной из узкоспециализированных задач названной области является создание автореферата по статьям одного автора. В работе представлены методы построения автореферата публикаций одного автора, указаны достоинства и недостатки методов.

Введение

Актуальность задачи автореферирования весьма высока, поскольку с учетом постоянно растущего объема обрабатываемой информации все чаще требуется заменить ее кратким аналогом.

Одной задач автореферирования является построение реферата по текстам схожего содержания. В этом случае в результирующем тексте могут оказаться предложения с одинаковыми лексическими конструкциями (дублирование) [1]. Примером такой задачи может быть создание автореферата по работам одного автора или некоторой совокупности работ по рассматриваемой тематике. Например, при написании некоторой статьи или научной работы часто приходится изучать большое количество статей по заданной тематике.

Несложным решением будет создание автореферата по каждому отдельному тексту с последующим их объединением. В результате получится текст, в котором лексически все предложения окажутся уникальными, но часто будут совпадать по смыслу. Это является проблемой, так как большинство современных средств построения квазиавторефератов создадут такое дублирование [2].

Цель данной работы – оценить возможность исключения дублирования в авторефератах, построенных по методу квазиреферирования.

Объект исследования – задача квазиреферирования.

Предмет исследования – решение задачи квазиреферирования с помощью методов, направленных на исключение дублирования.

В работе изучены, сравнены и созданы методы автореферирования с использованием квазиреферирования и закона Зипфа. Также создан программный модуль для решения задачи автореферирования на основе вышеуказанных методов. Для него подготовлены различные тестовые данные, с его помощью проведены экспериментальные исследования и на их основе сделаны выводы о достоинствах и недостатках каждого метода. Выбраны оптимальные методы для разных классов задач. Сделаны выводы о возможности исключения дублирования исходя из заданных условий.

I. Алгоритм построения автореферата и общие сведения об авторефератах

1. Определить выходной размер автореферата. Как правило, задается пользователем в начале генерации в размере n %.
2. Разбить текст на предложения. Каждое предложение с некоторой вероятностью p попадет в итоговый автореферат.
3. Выделить в предложениях ключевые слова по закону Зипфа.
4. Назначить каждому ключевому слову коэффициент значимости k. Например, используя частоту вхождения ключевых слов, которая является численной метрикой.
5. Определить значимость каждого предложения путем назначения ему численного значения m, являющегося суммой всех входящих в него k, поделенной на общее количество слов w в предложении, - k/w. Данный прием называют нормализацией предложений по длине. Ввиду этого условия резко возрастает вероятность появления в тексте коротких предложений. Потому как вероятность содержания в тексте предложения с высоким m, но низким w весьма велика. Поэтому также необходима функция, которая будет отсеивать короткие предложения, иными словами, устанавливать на w определенное ограничение. Вес предложения пропорционален весу входящих в него слов.
6. Выбрать n % предложений из текста с наибольшим m.
7. Отсортировать их в исходном порядке для каждого текста, а тексты отсортировать по наборам ключевых слов [3].

II. Способы создания сводного автореферата

Автореферат, построенный частотными методами, – это набор предложений, которые содержат наиболее часто повторяющиеся слова в тексте. Другими словами, это краткое содержание текста. Сводный автореферат – это текст, построенный из совокупности рефератов.

Одним из наиболее простых способов его построения является подход, в котором берется текст в размере n % от исходного размера каждого текста, а затем полученные авторефераты объединяются в один текст – сводный автореферат. Шаги этого процесса:

1. Для каждого текста создается свой автореферат.
2. Все тексты объединяются в один.
3. Берется n % от итогового текста и получается сводный автореферат.

К достоинствам этого метода можно отнести его простоту, отсутствие больших временных затрат на выполнение. Метод работает с большими текстами.

Недостатками подхода считаются высокая вероятность появления дублирования, возможность потери уникальных особенностей каждой работы. Кроме того, предложения будут сгруппированы в рамках одного текста по смыслу, но сами тексты будут идти в произвольном порядке.

Существует метод, который сначала объединяет все тексты в один, затем выделяет в нем ключевые слова, а потом находит наиболее подходящие предложения. Шаги алгоритма:

1. Все тексты объединяются в один.
2. По нему уже строится сводный автореферат.

По скорости работы этот метод можно охарактеризовать как средний. Заданный размер, равный n %, достигается точно.

Главным недостатком считается наличие дублирования в тексте.

Один из сложных для реализации методов направлен на исключение дублирования. Метод состоит в том, чтобы по ключевым словам достичь с определенной точностью уникальности каждого предложения.

Рассмотрим данный метод подробнее.

Необходимо модифицировать первый из вышеуказанных алгоритмов создания автореферата следующим образом. После шага 5 необходимо выполнить дополнительные дейсвтия над предложениями, а именно исключить дубликаты. Для этого необходимо ввести коэффициент E, который будет означать порог дубликатов или степень похожести предложений. Также есть функция $f\ (\ a,\ b) > E$. Функция просто сравнивает ключевые слова из предложений a и b, и если они совпадают больше, чем на E, тогда предложения признаются дублирующими и остается из них только то, у которого больше m. На основании проведенных исследований в примере выбирается $E = 0{,}7$. Это означает, что предложения a и b признаются идентичными по смыслу, если совпадают больше, чем на 70 %. Данная функция вызывается для всех предложений n^2 раз, где n - количество предложений.

Достоинства метода: точность n % удовлетворяется, дублирование сведено к минимуму. Существует возможность сгруппировать предложения по смыслу.

Недостатки: сложность реализации, низкая скорость работы.

III. Выводы

Частотное автореферирование – это наиболее простой и быстрый способ получения текста, отражающего содержание исходного текста.

Существует много методов частотного автореферирования. Все они имеют свои достоинства и недостатки. Для всех частотных методов выполняется следующее правило. Чем полнее и осмысленнее реферат отражает содержание исходного текста, тем соответствующий ему метод сложнее в реализации и медленнее работает.

Общими недостатками частотных методов являются следующие. Итоговый текст получается слабо связанным, может не содержать главной идеи исходного текста, трудночитаем, допускает дублирование предложений Указанные недостатки частотных методов лучше всего исправлять с помощью семантических алгоритмов. Однако это приведет к усложнению алгоритмов и увеличению времени их работы.

1. Толстых, В. К. Text Mining: Глубинный анализ текста / В. К. Толстых.– Санкт-Петербург: Питер, 2008.– 132 с.
2. Браславский, П. Ю. Избранные прикладные задачи информатики / П. Ю. Браславский.– Москва: Вильямс, 2005.– 168 с.
3. Третьяков, Ф. И. Распараллеливание алгоритмов классификации и кластеризации данных / Ф. И. Третьяков, Л. В. Серебряная // Вестник БГУ.– Серия 1.– Минск, 2013.– С. 105–108.