УДК 004.4’6:655.535.54

МЕТОДЫ АВТОМАТИЧЕСКОГО ПОСТРОЕНИЯ РЕФЕРАТОВ НА ОСНОВЕ ЧАСТОТНОГО АНАЛИЗА ТЕКСТОВ

Ф.И. ТРЕТЬЯКОВ, Л.В. СЕРЕБРЯНАЯ

Белорусский государственный университет информатики и радиоэлектроники
П. Бровки, 6, Минск, 220013, Беларусь

Поступила в редакцию 31 октября 2013

Рассмотрены алгоритмы построения авторефератов на основе частотного анализа текстов. Выполнен сравнительный анализ алгоритмов, а также предложены способы улучшения качества автореферата. Реализован метод составления сводного реферата. Создан программный модуль для решения задачи автореферирования на основе разработанных методов.

Ключевые слова: частотный анализ, автореферат, закон Зипфа.

Введение

В условиях огромного и постоянно растущего объема обрабатываемой информации задача автореферирования является весьма актуальной, поскольку краткий смысловой аналог исходного текста позволяет облегчить и ускорить обработку информации. Частотное автореферирование – это способ создания авторефератов, использующий только количественные характеристики текстов (например, частоту встречаемости слов). Частотный автореферат состоит из некоторой совокупности предложений, извлеченных из текста, причем порядок предложений автореферата может быть изменен относительно исходного текста. В результате, прочитав автореферат, не всегда удается правильно понять его содержание.

Обычно частотные авторефераты составляются на основе частоты вхождения ключевых слов. Это термины, которые чаще других встречаются в тексте и используются для передачи его содержания. Поэтому в автореферат попадают предложения, включающие в себя наибольшее количества ключевых слов. Сводный автореферат – это текст, построенный из совокупности текстов-авторефератов [1]. Одной из задач автореферирования является проблема построения реферата по текстам схожего содержания. В этом случае в результирующем документе могут оказаться предложения с одинаковыми лексическими конструкциями (дублирование) [2]. Примерами таких подзадач может быть создание автореферата по работам одного автора или совокупности работ по рассматриваемой тематике. Несложным способом является построение автореферата по каждому отдельному тексту с последующим их объединением. В результате получится текст, в котором лексически все предложения считаются уникальными, но во многом идентичными с точки зрения смысла. Ключевым словам присваиваются весовые коэффициенты. Высока вероятность того, что слова, имеющие большие веса, попадут в авторефераты схожих тематик, а затем и в сводный автореферат. В результате получится текст, формально корректный, но содержащий повторяющиеся предложения. При этом оригинальная часть каждой работы может «выпасть» из результирующего текста, т.к. не будет иметь большие весовые коэффициенты ключевых слов. Следовательно, автореферат, построенный человеком, может существенно отличаться от текста, полученного автоматически с помощью программного средства [3]. Для оценки качества автоматически построенного частотного реферата предложено использовать сводный автореферат, построенный человеком.

Цель данной работы – оценить возможность исключения дублирования в авторефератах, построенных по методу квазиреферирования. Для этого рассматриваются различные методы автореферирования и выполняется их сравнительный анализ. В статье предложены методы частотного автореферирования на основе законов Зипфа. Создан программный модуль для решения задачи автореферирования на основе вышеуказанных методов. Подготовлены наборы тестовых данных. Разработан метод верификации результатов. Выбран оптимальный метод для каждого класса задач.

Алгоритм построения автореферата

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 сводный автореферат.

3. Разбить предложения на ключевые слова по закону Зипфа [4].

4. Назначить каждому ключевому слову коэффициент значимости *k*. Для этого можно применить разные методы. Например, использовать частоту вхождения ключевых слов (является численной метрикой).

5. Определить значимость каждого предложения путем определения для него численного значения *m*, являющегося суммой всех *k,* входящих в него, деленное на общее количество слов *w* в предложении *k*/*w*. Данный прием называют нормализацией предложений по длине. Ввиду этого условия резко возрастает вероятность появляения 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, т.е. устанавливать определенное ограничение на *w*. 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.

При этом следуют учитывать следующие характеристики текстов.

1. Сигнальне фразы. Это слова или словосочетания, которые априори содержат ключевую информацию, имеют максимальный вес и должны быть включены в итоговый реферат. Например, «вывод», «необходимо подчеркнуть». Их также называет маркерами. Для того, чтобы обнаружить такую фразу в тексте, используют специальные словари маркеров [5].

2. Вопросительные предложения имеют понижающий коэффициент.

Способы создания сводного автореферата

Рассмотрим наиболее часто используемые способы. Первый метод.

1. Для каждого текста создается свой автореферат.

2. Все тексты объединяются в один.

3. Выделяются *n* % от итогового текста в итоговый сводный автореферат.

Достоинства: самый простой в реализации метод, не требует больших временных затрат на выполнение, способен работать с большими текстами.

Недостатки: на выходе *n* % не будет достигнуто точно, т. к. работы могут сильно отличаться по объему. Существует высокая вероятность появления дубликатов, вызванная наличием в разных рефератах предложений, имеющих большой вес и схожее содержание. В результате все эти предложения-дубликаты попадут в итоговый автореферат.

Второй метод.

1. Все тексты объединяются в один.

2. По нему строится сводный автореферат.

Достоинства: точность *n* % удовлетворяется, скорость работы средняя, т. к. время обработки текста увеличивается с добавлением каждой новой структурной единицы.

Недостатки: дубликатов меньше, чем в первом методе. Скорость работы значительно ниже, чем в предыдущем методе, поскольку параллельные вычисления показывают более

высокую скорость работы на независимых друг от друга структурах данных [6].

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.

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

Недостатки: сложность реализации, низкая скорость работы.

Рассмотрим данный метод подробнее. Необходимо модифицировать алгоритм создания автореферата следующим образом.

После шага 5 необходимо выполнить дополнительные действия над предложениями – исключить дубликаты. Для этого необходимо ввести специальный коэффициент *E*, который будет означать порог дубликатов или степень похожести предложений. Есть некая функция *f (a, b) >E*. Она сравнивает ключевые слова из предложений *a* и *b*. Если они совпадают больше, чем на *E*, тогда предложения признаются дублирующими, и из них остается только то, у которого *m* больше. По результатам исследований оптимальным значением *E* является 0,7. Тогда предложения *a* и *b* признаются лексически идентичными, если совпадают больше, чем на 70 %. Данная функция вызывается для всех предложений, т. е. выполняется n^2 раз, где *n* – количество предложений. Функция *E* значительно замедляет скорость работы алгоритма. Для его ускорения предложено использовать параллельные вычисления [6].

Созданный программный модуль строит автореферат текстов сходной тематики, который наиболее полно включает тезисы каждого текста, содержит минимальное количество дублирующей информации и имеет размер, равный *n* % от общего размера всех текстов.

Оценка полученных результатов

Функции программного модуля верификации результатов связаны с оценкой качества сводного автореферата. Для этого использован перечень ключевых слов сводного автореферата, построенного автором. Основной метрикой качества является соотношение совпадения ключевых слов программного модуля и исходного сводного автореферата.

В качестве объектов исследования выберем некоторую совокупность текстов, разделенных на три группы.

1. Тексты различной тематики, не коррелирующие между собой.
2. Тексты сходной тематики, но разных авторов.
3. Тексты работ одного автора.

Исследования проводились для 100 текстов каждой категории. Размер текстов приблизительно около 15 000 символов. Каждая совокупность текстов содержит сводный автореферат, который будет использоваться для верификации результатов.

Для верификации результатов было разработано программное средство на платформе .NET Framework 4.5. Программное средство принимает на вход два автореферата и выводит на экран результаты в виде графической информации, показанной на рисунке.

Верификация результатов выполняется следующим образом.

1. Выделяются ключевые слова с их весовыми коэффициентами из двух авторефератов: исходного *A* и полученного *B* (построенного программой).

2. Строятся векторы весов для обоих авторефератов. Причем, если одно из ключевых слов отсутствует в другом векторе, оно добавляется туда с весом 0. Таким образом, векторы нормализуются по длине.

3. Для каждого элемента вектора *B* вычисляется его разница с соответствующим элементом вектора *A* по следующей формуле: $C_i = \frac{|A_i - B_i|}{A_i}$. Получается вектор различий *C*.

4. Находится среднеквадратическое отклонение вектора различий по формуле:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

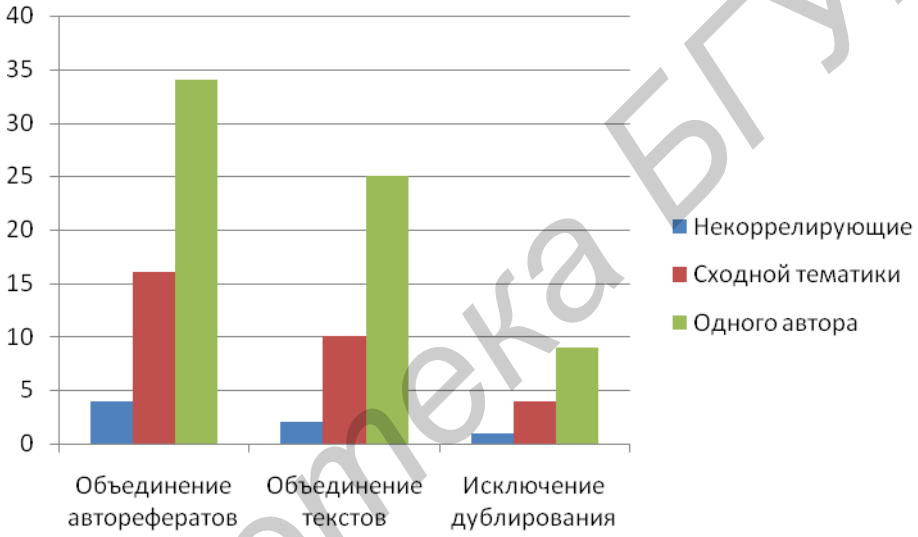
Таким образом, получается итоговая оценка отклонения полученного автореферата от исходного. Чем меньше отклонение, тем качественнее получился текст, тем лучше выбранный метод.

Если текст изобилует предложениями-дубликатами, то в нем увеличиваются веса ключевых слов этих предложений. В результате в тексте присутствует два-три ключевых слова с огромными весами, а все остальные – с низкими. Поэтому итоговое среднеквадратическое отклонение получается больше.

Для текстов с различной тематикой априори больше всего подходит первый либо второй метод. Способ устранения дублирования не очень подходит для данного вида текстов, т.к. вероятность самостоятельного появления дубликатов в нем весьма мала.

Тексты сходной тематики, но разных авторов могут содержать дубликаты, поскольку разные авторы часто ссылаются на одни и те же источники, и отклонение здесь может быть велико. В этом случае метод исключения дублирования подходит лучше всего. Он позволяет получить автореферат практически без дубликатов (отклонение в пределах 2–4 %).

Для текстов одного автора первые два метода подходят плохо, расхождение может составить до 35 %. Метод исключения дублирования имеет расхождение в 6–10 %, что является хорошим показателем для квазиреферирования.



Оценка результатов устранения дублирования: на оси абсцисс отложены исследуемые методы и виды тестовых текстов; на оси ординат отображается процент отклонения набора ключевых слов автоматического реферата от созданного автором

Заключение

Частотное автореферирование – это наиболее простой и быстрый способ создания авторефератов. С помощью частотного реферирования можно получить краткое содержание текста за минимальное время.

Существует много методов частотного автореферирования. Все они имеют свои достоинства и недостатки. Метод исключения дублирования демонстрирует неплохие показатели по скорости работы на больших объемах данных (от 100 текстов 15000 символов в каждом) с помощью распределенных вычислений [6] и показывает малый процент отклонения (1–10 %) по ключевым словам от идеального сводного автореферата.

Недостатки частотного реферирования обусловлены самим способом реферирования. Чтобы их устранить, необходимо вводить понятие семантики и строить итоговый текст, анализируя его содержание. Иными словами, использовать семантические методы. Однако это будет гораздо медленнее и сложнее. Для построения авторефератов методами со сложными алгоритмами (например, метод исключения дублирования) хорошо могут быть использованы распределенные вычисления (распараллеливание на уровне выполнения с помощью потоков) [6]. Кроме того, закон Зипфа, словари маркеров или позиционный поиск улучшают работу автореферирования.

AUTOMATIC CONSTRUCTION METHODS OF ABSTRACTS ON THE BASIS OF THE TEXTS FREQUENCY ANALYSIS

F.I. TRETYAKOV, L.V. SEREBRYANAYA

Abstract

Algorithms of author's abstracts construction on the basis of the texts frequency analysis are considered. Their comparative analysis of algorithms is made, and also ways of quality improvement of the author's abstract are offered. The algorithm of drawing up of the summary abstract is realized. The program module for automatic construction of abstracts on the basis of the developed methods is created.

Список литературы

1. TextMining. Глубинный анализ текста. Из цикла лекций «Современные Internet-технологии» для студентов 5-го курса кафедры Компьютерных технологий физического факультета Донецкого национального университета. ДонНУ, кафедра КТ, проф. В.К. Толстых.
2. *Паклин Н.Б., Орешков В.И.* Бизнес-аналитика: от данных к знаниям. СПб, 2009.
3. *Захаров В.П.* Информационные системы (документальный поиск). СПб, 2002.
4. *Серебряная Л.В., Третьяков Ф.И.* // Матер. VI междунар. науч.-практ. конф. «Актуальные вопросы методики преподавания математики и информатики». Биробиджан, 20 апреля 2011. С. 175–181.
5. *Серебряная Л.В., Чебаков С.В.* // Информатизация образования. 2011. № 2. С. 52–61.
6. *Третьяков Ф.И., Серебряная Л.В.* // Вест. БГУ. Сер. 1. 2013. № 2. С. 105–108.