АВТОМАТИЗИРОВАННАЯ СИСТЕМА ОЦЕНКИ ЕСТЕСТВЕННОСТИ ТЕКСТОВ

А.В. Юрасов, О.А. Дегтярёва

Самарский государственный аэрокосмический университет имени академика С.П. Королёва (национальный исследовательский университет) (СГАУ), Самара, Россия

В данной статье описываются результаты исследования естественности текстов, законы и алгоритмы, лежащие в основе работы разработанной системы. Данная статья должна дать представление об изменении количественных характеристик текстов после применения к ним первого закона Джорджа Ципфа.

Ключевые слова: естественность, законы Ципфа, предложение, реферат, слово, стемминг, статистика, текст.

Введение

Большинство текстов, с которыми мы имеем дело в повседневной жизни – естественные тексты. В качестве примера неестественных текстов можно привести тексты, созданные для поисковых роботов. Они создаются с целью повышения позиции сайта в поисковой выдаче. С такой «чёрной» оптимизацией поисковые системы ведут активную борьбу вплоть до исключения сайта из поискового индекса.

Законы, сформулированные Ципфом, описывают закономерности частотного распределения слов в тексте на любом естественном языке. Законы эмпирические – они не имеют строгого математического доказательства и основаны на статистическом анализе распределения слов в больших массивах текстов на разных языках. Тем не менее, статистически их верность не вызывает никаких сомнений.

1. Постановка задачи

В качестве объекта исследования выступает некоторый текст с большим содержание повторяющихся слов и рефераты, созданные на его основе. Цель исследования – изучение того, как изменяется содержание ключевых слов в текстах исходного реферата и реферата, полученного на основании статистики, вычисленной с использованием первого закона Ципфа.

По мнению авторов, содержание ключевых слов, определяемых пользователем для каждого отдельно взятого текста, должно увеличиться в целевом реферате по сравнению с исходным.

2. Первый закон Ципфа

Первый закон [1] связывает понятия ранга слова и частоты, где частота слова – это количество вхождения слова в текст, а ранг – номер слова в общем списке обнаруженных слов, упорядоченном по убыванию частоты. В любом тексте, написанном человеком, этот закон статистически верен. Статистически, а не математически – потому что для не-

больших текстов всегда возможны отклонения, но чем больше число слов в тексте, тем эти отклонения меньше.

Первый закон Ципфа (формула 1) говорит о том, что вероятность обнаружения любого слова, умноженная на его ранг – постоянная величина (C).

$$C = P * r, \quad (1)$$

где C – постоянная величина, r – ранг слова, P – вероятность обнаружения некоторого слова на в тексте.

Вероятность P определяется формулой 2.

$$P = \frac{f}{N}, \quad (2)$$

где P – вероятность, f – частота, N – общее число слов.

3. Алгоритм стемминга

Первый опубликованный стеммер был написан Джули Бет Ловинс в 1968 году. Позже стеммер был написан Мартином Портером и опубликован в 1980 году. Конкретная реализация данного алгоритма – стеммер [2].

Стемматизацией [3] обычно называется приближённый эвристический процесс, в ходе которого от слов отбрасываются окончания и суффиксы. Стемминг часто подразумевает удаление производных аффиксов. Аффикс – это морфема, которая присоединяется к корню и служит для образования новых слов.

Слова в русском языке могут быть очень короткими [2]. Также очень много частиц, союзов и т.п. Такие слова образуют группу естественных «стоп-слов». Это дает основание создать своеобразный фильтр, позволяющий исключить из рассмотрения данные слова. В данной работе в качестве фильтра использовалось ограничение по длине слова. Для этих слов не будут вычислены идеальные и рекомендуемые значения их повторения в текст. Также они не будут оказывать влияния на процесс формирования рефератов.

4. Работа системы

На первом этапе работы система выделяет слова, используя алгоритм стемминга, в тексте и вычисляет количество их повторений.

На втором этапе система на основании полученной статистики вхождения слов в текст вычисляет по формуле 3 идеальные и рекомендуемые значения количества повторений каждого слова в тексте. Согласно первому закону Ципфа график распределения слов в естественном тексте должен быть приближен к графику обратной функциональной зависимости.

$$y = \frac{a}{x} + b, \quad (3)$$

где x – ранг слова, y – частота слова.

Для нахождения коэффициентов a и b в работе использовался метод наименьших квадратов.

Если идеальное значение больше текущего количества повторений слова в тексте, то рекомендуемым значением будет округлённое идеальное до ближайшего меньшего целого. Если же меньше, то округлённое идеальное значение до ближайшего большего целого.

На рисунке 1 представлен график полученных распределений слов в тексте. Красная линия отображает исходное распределение «ранг-частота» анализируемого документа. Зелёная линия показывает идеальные значения количества повторений слов в тексте. С точки зрения первого закона Ципфа. Это нецелые значения. Синея же линия отображает целые значения, используемые для формирования второго реферата.

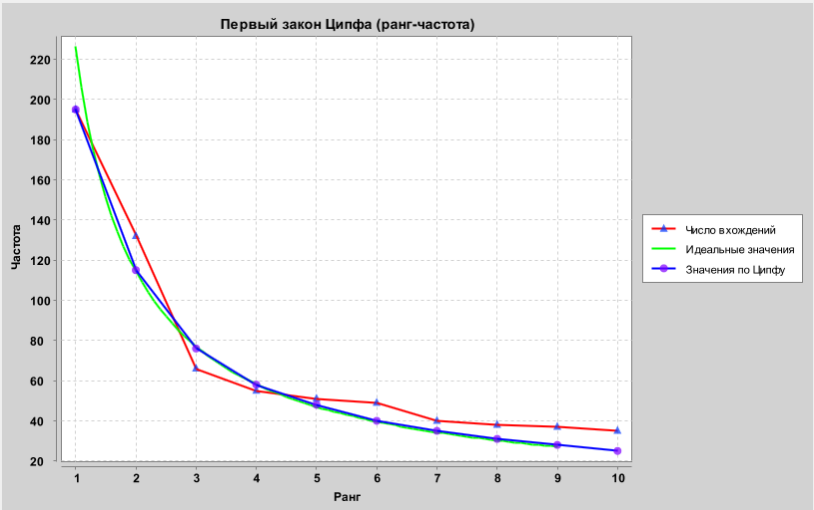

Рис.1. Распределение «Ранг-частота»

После этого выполняется построение двух рефератов: первого – на основе начального количества вхождений слов в текст, второго – на основе рекомендуемых значений количества повторений слов. Подробно алгоритм формирования рефератов описан в [4]. В первом и во втором случае создание реферата выполняется путём выборки заданного количества предложений с наибольшим весом в порядке встречаемости в тексте. Вес предложения рассчитывается как сумма количества повторений его слов.

Далее пользователем выбираются ключевые слова, характерные для анализируемого текста. Для этих слов выполняется подсчёт их вхождения полученные рефераты. После этого статистика выводится пользователю. Пример полученной статистики представлен на рисунке 2.

Красный столбец отображает содержание определенного слова в исходном реферате. Жёлтый столбец – в целевом реферате, сформированном на новых значениях количества повторений слова в тексте.

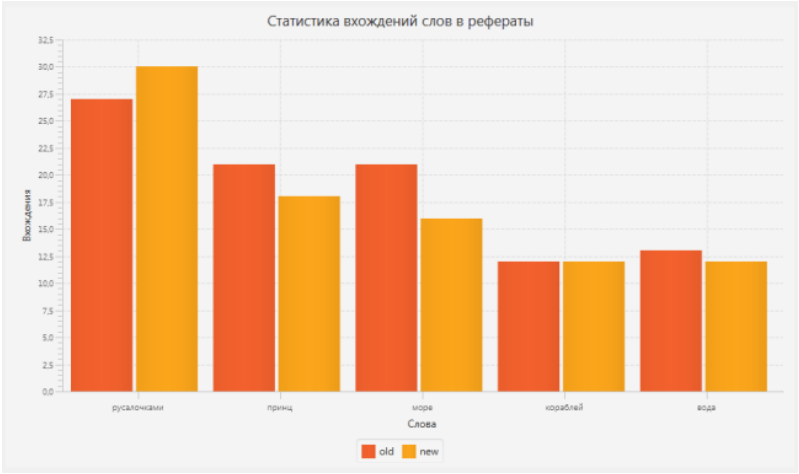

Рис.2. Статистика вхождения ключевых слов в рефераты

5. Результаты

В качестве объектов исследования естественности текста были выбраны сказки. Сказки были выбраны потому, что в них несмотря на относительно небольшой размер текста часто повторяются ключевые слова. Они специфичны для каждой отдельно взятой сказки. К ним можно отнести: Иван, Царь, Кощей, Василиса, принцесса и т.п. По предположениям, это позволило бы наглядно продемонстрировать изменение содержания реферата после приведения слов в тексте к естественному распределению.

Зависимость естественности текста от длины игнорируемых слов

В таблице 1 представлены результаты исследования зависимости естественности текста от длины игнорируемых слов в ходе работы системы. Следует помнить тот простой факт, что короткие слова встречаются в тексте чаще. Из таблицы клад в естественность текста, который вносят часто повторяющиеся короткие слова.

Табл. 1. Зависимость естественности текста от длины игнорируемых слов

Длина игнорируемых слов, символ	Естественность, %
2	73,1
3	74,4
4	76,8

При увеличении длины игнорируемого слова на 1 символ происходит увеличение естественности текста в среднем на 1.8 процента. Естественность текста вычислялась как процент слов от всех слов текста, для которых рекомендуемые значения количества повторений совпадали с исходными значениями повторения в тексте.

Зависимость количества ключевых слов в рефератах от их размера рефератов для малых текстов

На рисунке 3 представлена зависимость количества ключевых слов в исходных и новых рефератах от размера рефератов. Из рисунка видно, что для малых текстов с количеством слов примерно до 20000 предположение об увеличении их содержания во втором реферате не подтверждается. Изменения носят довольно случайный характер. Во многом этот факт зависит от способа составления реферата. На вес предложения продолжают оказывать влияние часто употребимые слова и их формы такие как: быть, стать, идти и т.д.

Но можно отметить увеличение содержания ключевых слов при росте размера реферата в больших текстах: «Приключения Чиполлино» и «Золотой ключик или приключения Буратино» (рисунок 4). Во многом это происходит за счёт первой тройки наиболее часто встречающихся ключевых слов.

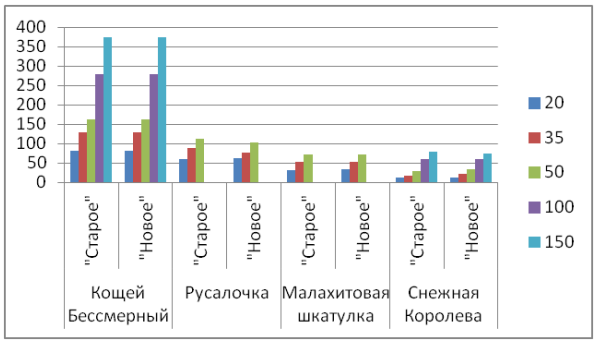

Рис.3. Количество ключевых слов в рефератах в зависимости от размера рефератов для малых текстов

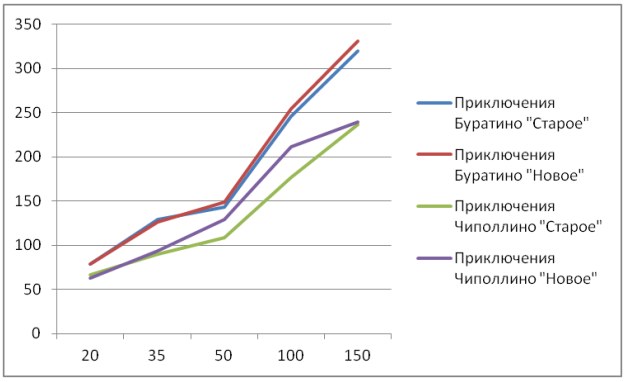

Рис.4. Количество ключевых слов в рефератах в зависимости от размера рефератов для больших текстов

6. Заключение

В статье были рассмотрен первый закон Ципфа о распределении слов естественных текстах, алгоритм стемминга, применяемый для выделения слов. Были приведены результат исследований зависимости изменения содержания ключевых слов в рефератах, построенных на основании рекомендуемых значений повторений слов в тексте. Из них можно сделать вывод о том, что применение первого закона Ципфа к небольшим текстам не ведёт к увеличению содержания ключевых слов в его новом реферате. Целесообразно применять данный закон к более большим текстам. Также следует продолжать работу по уменьшению влияния неинформативных слов на содержание реферата. Например, исключение часто повторяющихся глаголов и других частей речи. Алгоритм стемминга обладает возможность поиска частей речи по специфичным для них окончаниям.

Литература

1. Законы Ципфа – вводная статья [Электронный ресурс]. – URL: http://webpavilion.ru/статьи/закон-ципфа-вводная (дата обращения 18.02.2016).
2. Russian stemming algorithm [Электронный ресурс]. – URL: http://snowball.tartarus.org/algorithms/russian/stemmer.html (дата обращения 18.02.2016).
3. Стемминг [Электронный ресурс]. – URL: http://ru.wikipedia .org/ (дата обращения 18.02.2016).
4. Трусов В. Построение тезаурусов, тематических классификаций и рубрикаторов для поиска информации в распределенных информационных системах [Электронный ресурс]. – URL: http://www. aselibrary.ru/digital_resources /journal/irr/irr2725/irr27253027/irr272530273030/irr2725302730303032 (дата обращения 18.02.2016).