Клячин В.А., Помельников Ю.В., Попов В.В., Штельмах Т.В.
О некоторых частотных характеристиках текста

*Волгоградский государственный университет
(Россия, Волгоград)*

doi: 10.18411/lj-12-2018-128

idsp: ljjournal-12-2018-128

Аннотация

В работе рассматриваются методы фильтрации некачественных текстов, опирающиеся на набор слов в тексте и на их частотность. Исследуется возможность применения частотных характеристик на основе закона Зипфа. Они позволяют выявить особенности словарного состава отдельных законченных текстов и их коллекций, но не дают возможности выявить некачественные тексты рассматриваемых типов: произвольная перетасовка слов, использование синонимайзеров. Предложены модификации этого закона. Однако с применением графовой модели текста, в которой учитывается нахождение слов в одном предложении, получена характеристика (медиана степеней вершин графа), позволяющая отделить бессмысленные тексты на основе перетасовки от осмысленных.

Ключевые слова: закон Зипфа, "мешок слов", графовая модель текста, фильтрация текстов.

Введение

Фильтрация наборов текстовой информации зачастую является необходимым этапом в решении задач анализа данных. В определенных ситуациях исходные данные могут быть представлены массивом текстов, для анализа которых необходимо структурированное их представление. В настоящей статье мы ориентируемся на две простые модели бесполезных текстов («мусора»). Первая модель «А» представляет список слов и знаков препинания, случайно выбранных из словаря. Вторая модель «В» представляет собой список слов и знаков препинания, случайно выбранных из словаря, но для каждого элемента текста задана вероятность выборки. Основная задача состоит в определении таких характеристик текста, используя которые возможно идентифицировать текст как «мусор» по первой или второй модели. Математически это означает построение функции, которая вычисляет значение, используемое для дискриминации текста. С целью решения указанной задачи, мы исследуем возможность применения модификаций законов Зипфа (G.K. Zipf). Несложно заметить, что модель «мешка слов» может быть применимой только для первой модели «мусора», т. к. в этой модели, в частности, все вычисляемые характеристики текста будут инвариантны при его перемешивании. Для решения идентификации «мусора» для второй модели мы применяем графовую модель.

1. Исследование на основе модели «мешка слов»

Рассматривается следующая модель текста. Если D — документ и w — слово из этого документа, то через $k(w)$ обозначается его частота вхождения, т.е. сколько раз это слово встретилось в документе. Вероятность $p(w)$ встретить слово w в произвольном месте документа равна $k(w)/Nw$, где Nw — общее число слов документа D . Предварительно можно провести лемматизацию. Все слова из документа D разбивают на группы: слова w_1 и w_2 попадают в одну группу, если $p(w_1)=p(w_2)$. Группы располагают в порядке убывания частот $p(w)$ входящих в них слов и нумеруют числами 1, 2, 3, Эти номера называют рангами частот слов. Таким образом, наиболее часто встречающиеся слова будут иметь ранг 1, следующие за ними — ранг 2 и т.д. Ранг слова w обозначается через $r(w)$.

Первый закон Ципфа (Zipf) можно сформулировать так [1, стр. 378]:

$$k(w)=A \cdot (r(w))^{-\gamma}, \quad (1)$$

где A и γ — некоторые постоянные величины. Параметр γ обычно близок к 1 и имеются веские причины, по которым для многих текстов удобно считать, что $\gamma=1$ [1, стр. 386]. Если равенство (1) разделить на число Nw всех слов в документе и положить $\gamma=1$, то получим равенство:

$$p(w) \cdot r(w)=C, \quad (2)$$

где $C=A/Nw$ — постоянная величина. Значение константы C зависит от языка и особенностей текста: стиля, объема словаря [2] и других параметров. Например, для русских текстов по электронике $C=0.145$ [3, стр. 105].

В наших экспериментальных исследованиях получены результаты, показывающие значительные отклонения от закона Ципфа в форме (2) как для отдельных текстов, так и для коллекций текстов..

Ясно, что, для всех слов w ранга r произведение $C(w)=p(w) \cdot r(w)$ одинаково и зависит только от величины ранга $r=r(w)$. Если предположить справедливость равенства (2) для всех слов документа D , то графиком функции $C(w)$ должна являться горизонтальная прямая. Однако в действительности графики выглядят иначе. На рисунках 1 и 2 приведены графики зависимости величины $C(w)$ от ранга r для 8 текстов художественных классических произведений на русском языке. В первом случае составлялись словари полных и кратких прилагательных, во втором — словари всех используемых слов. Рисунок 2 показывает, в частности, что увеличение объема текста приводит к уменьшению $C(w)$.

Рисунок 1. Графики функции $C=p \cdot r$ для 8 файлов, словарь прилагательных

Рисунок 2. Графики функции $C=p \cdot r$ для 8 файлов, словарь любых слов

Нами анализировались также коллекция текстов на русском языке из 60 художественных произведений объемом примерно 36Мб с различными разбиениями на составляющие (18 и 19 век; российские и зарубежные авторы) и различными ограничениями на тип слов (части речи, длина слов, стоп-слова). Документов, подтверждающих равенство (2), не обнаружено.

Нами предложены три различных показателя отклонения документа D от закона Ципфа в форме (2), из которых наиболее эффективным оказался такой показатель:

$$Norm0 = \max\{|C(w) - MC| : w \in D\}, \quad (3)$$

где MC — среднее арифметическое величин $C(w)$ по всем словам w из документа D . Так, для всех существительных из романа Ф.М. Достоевского "Братья Карамазовы" получена по формуле (3) величина $Norm0=0.11558$.

Рассмотрены также два альтернативных способа ведения шкалы рангов слов, при которых уменьшается общее число рангов и, главное, уменьшаются отклонения от закона Ципфа. В первом случае выбираем положительное число m и разбиваем множество неотрицательных действительных чисел на полуинтервалы $[0, m)$, $[m, 2m)$, $[2m, 3m)$, $[3m, 4m)$, Во втором случае выбираем действительное число $\lambda > 1$ и разбиваем то же множество чисел на полуинтервалы $[0, \lambda)$, $[\lambda, \lambda^2)$, $[\lambda^2, \lambda^3)$ и т.д.

Считаем, что слова w_1 и w_2 имеют один ранг, если их частоты $k(w_1)$ и $k(w_2)$ попадают в один полуинтервал. При $\lambda=1.02$ для упомянутого выше романа Ф.М. Достоевского $Norm0=0.10098$. Полученное отклонение меньше по сравнению с отклонением, найденным при традиционном способе определения ранга.

3. Исследование на основе графовой модели текста

Для построения графовой модели вводится граф $G=(V,E)$, в котором множество вершин образуется множеством всех слов текста, а ребрами отмечаем факт нахождения пары слов в одном предложении. При этом каждому ребру приписывается его вес b — количество предложений текста, содержащих эту пару слов. Некоторые свойства такого графа были изучены в работе [4]. Отметим, что при перемешивании слов текста его граф меняет структуру. В соответствии с моделью «В» текста «мусора» при указанном перемешивании мы получаем тексты, которые имеют то же распределение частот слов, что и исходный текст. Поэтому, мы ставим задачу поиска характеристики графа текста, которая была бы не инвариантной при таком перемешивании. Среди имеющихся характеристик графа мы рассмотрим:

- средний вес ребер графа,
- медианное значение весов ребер графа,
- среднее значение степеней вершин графа,
- медианное значение степеней вершин графа.

В качестве текстов для исследования возьмем тексты художественных произведений с ограничением на размер файла. Будем рассматривать все файлы из заданного каталога и его подкаталогов с ограничением на размер 60Кб. Проведенные эксперименты показывают, что медианное значение степеней вершин графа текста может быть использована для идентификации «мусора». При этом схема идентификации может быть следующей. Для исходного текста вычисляются степени вершин графа. Затем вычисляется их медианное значение. Далее исходный текст подвергается многократному перемешиванию и после каждой такой операции вычисляются значения медианы степеней вершин нового графа. После чего, берется среднее значение вычисленных медиан и сравнивается с медианой степеней исходного текста. Если полученное значение будет больше, то текст предполагается реальным, в противном случае — «мусор». Результаты проведенных экспериментов представлены на рисунке 3.

Рисунок 3. Результаты экспериментов

На этом рисунке по оси абсцисс отложен номер текста в корпусе, по оси ординат — медианное значение степеней вершин графа текста. Синим цветом отмечен график для исходных текстов, а красным — для текстов, полученных случайным

перемешиванием исходного. Таким образом, последнее значение является надежным параметром с помощью которого можно разделить осмысленные тексты от «мусора».

Исследование профинансировано Российским фондом фундаментальных исследований и правительством Волгоградской области, грант № 18-412-340007

1. Арапов М. В. Системный анализ лексической структуры текстов /М.В. Арапов// Системные исследования. Ежегодник.— М.: Наука, 1981.— С.372–403.
 2. Бакулина М. П. Использование закона Ципфа для сжатия текстов. Дискретный анализ и исследование операций. Сер.2. Том 14, 2007. №2. С. 3-13
 3. Пиотровский Р. Г. Текст, машина, человек. Л.: Наука, 1975.
 4. Григорьева Е. Г., Клячин В. А., Помельников Ю. В. и Попов В. В. Алгоритм выделения ключевых слов на основе графовой модели лингвистического корпуса // Вестник Волгоградского государственного университета. Серия 2: Языкознание. 2017. Т. 16. № 2. С. 58-67.
-