ТЕМАТИЧЕСКИЙ АНАЛИЗ И КВАЗИРЕФЕРИРОВАНИЕ ТЕКСТА С ИСПОЛЬЗОВАНИЕМ СКАНИРУЮЩИХ СТАТИСТИК *

THEMATIC ANALYSIS AND QUASI-ABSTRACTING OF TEXT USING SCAN STATISTICS

В.Д. Гусев

ИМ СО РАН, г. Новосибирск

gusev@math.nsc.ru

Л.А. Мирошниченко

ИМ СО РАН, г. Новосибирск

luba@math.nsc.ru

Н.В. Саломатина

ИМ СО РАН, г. Новосибирск

nataly@math.nsc.ru

Сканирующие статистики — эффективный и чувствительный инструмент для выявления отклонений от равномерности в распределении лексических единиц по тексту. Выделенные с их помощью кластеры (сгущения) отдельных типов единиц можно трактовать как сверхфразовые единства, определяющие тематику текста. Предлагается способ квазиреферирования, основанный на различных стратегиях оценивания информативности предложений внутри сверхфразовых единств и регулируемом выборе ограниченного числа предложений для реферата. Проводится сравнение с известными аналогами.

Введение

Задача автоматического индексирования и реферирования текстов не утрачивает своей актуальности, хотя первые работы в этом направлении появились свыше 50 лет назад. Новый импульс дало создание сети Интернет, в которой далеко не каждый документ снабжен авторским резюме и ключевыми словами. Заметную роль в технике автоматического индексирования и реферирования играют суперсинтаксические методы (см. обзор[1]). Обычно они предусматривают два этапа: 1) сегментацию текста на субтексты, называемые сверхфразовыми единствами (в общем случае они не совпадают с авторским разбиением текста на разделы и подразделы); 2) выявление в каждом субтексте наиболее информативных слов, словосочетаний или фраз.

Самый простой способ выделения сверхфразовых единств основан на анализе лексических повторов в тексте. Как правило, отбираются полнозначные слова (преимущественно существительные), фигурирующие в значительном количестве предложений текста. Особо важными считаются повторы, многократно отмеченные в одном фрагменте текста, что свидетельствует о его тематическом единстве [2, 3].

В настоящей работе предлагается *формальный критерий выделения сверхфразовых единств*, основанный на использовании *сканирующих статистик*. Фактически реализован эффективный и чувствительный инструмент для выявления аномальных отклонений от равномерности в распределении лексических единиц по тексту. Предполагается, что выделенные с помощью сканирующих статистик аномальные сгущения (кластеры) отдельных лексических единиц, трактуемые как сверхфразовые единства, достаточно адекватно отражают смысловое содержание соответствующего фрагмента текста.

Отличительной особенностью развиваемого подхода является то, что проводить предварительную фильтрацию малоинформативных лексических единиц (служебные и общие слова), в принципе, необязательно. Большая часть их устраняется автоматически, не пройдя критерий на аномальность. Другой особенностью является построение *«про-*

* Работа выполнена в рамках проекта № 03-06-80118, поддержанного грантом РФФИ.

филя кластеризуемости» лексических единиц в тексте и отбор фраз для квазиреферата в местах существенного нарастания и/или убывания этой функции. Такая стратегия присуща позиционным методам отбора значимых фрагментов в тексте [4], но они работают с задаваемой автором структурой текста, а не со сверхфразовыми единствами.

1. Схема выявления неслучайной кластеризации лексических единиц

Задача о выявлении неравномерностей позиционного распределения отдельных словоформ текста сводится к изучению различных схем расстановки точек на линии (каждую точку можно трактовать как место вхождения анализируемой словоформы в текст). Простейшая постановка выглядит следующим образом.

Пусть $x_1, x_2, ..., x_M$ будет произвольный набор точек из единичного интервала (0, 1]. Требуется проверить гипотезу о равномерности (H_0) против альтернативы (H_1), связанной с тем или иным типом отклонения от равномерности (кластеризация, сверхравномерное распределение, наличие «запретных» областей, изолированных точек и т.п.). Для случая кластеризации эффективное решение основано на использовании сканирующей статистики $n(d)$, фиксирующей максимальное число точек n, попавших в интервал длины d при всевозможных расположениях этого интервала внутри единичного отрезка [5]. Статистика названа сканирующей, поскольку вычисление ее ведется путем подсчета числа точек, попавших в окно ширины d, скользящее вдоль отрезка.

Из чисто алгоритмических соображений вместо статистики $n(d)$ мы используем связанную с ней статистику $d(n)$, фиксирующую длину минимального интервала, содержащего ровно n точек ($2 \le n \le M$). Распределение $d(n)$ при нулевой гипотезе получено в [6]. Поскольку табулирование распределения этой статистики в широком диапазоне значений n и M представляется достаточно трудоемким, для оценки значимости отклонения вычисленной на конкретном тексте статистики $d(n)$ от значения, постулируемого гипотезой H_0 (равномерность), целесообразно прибегнуть к имитационному моделированию [7]. Схема выявления позиционных аномалий в распределении лексических единиц тогда выглядит следующим образом.

1). Проводим нормализацию словоформ текста и подсчитываем частоту встречаемости каждой словоформы в нормализованном тексте.

2). Пусть x — произвольная нормализованная словоформа, $F(x)$ — число ее вхождений в текст, n — фиксированное число последовательных вхождений x в текст ($2 \le n \le F(x)$), $d(n)$ — длина минимального фрагмента текста, содержащего n вхождений цепочки x. Для дальнейшего анализа отбираем словоформы со значением $F(x) \ge F_{пор}$, где $F_{пор}$ — пороговое значение частоты, зависящее от длины текста N, исчисляемой в словоформах.

3). Для каждого из отобранных слов проводим перебор по всем допустимым значениям n ($F_{пор} \le n \le F(x)$). Для фиксированного n:

а) вычисляем значение $d(n)$ в анализируемом тексте;

б) с помощью имитационного моделирования оцениваем распределение этой статистики при гипотезе H_0. Для этого путем многократного перемешивания слов в исходном тексте формируем m его рандомизированных аналогов с равномерным распределением слова x по тексту (приемлемыми являются значения $m \ge 100$). По полученной подборке вычисляем оценки минимального, максимального и среднего значения статистики $d(n)$ (соответственно, $S_{\min}$, $S_{\max}$ и $\overline{S}$), а также среднеквадратичное отклонение s.

4). Сравниваем наблюдаемое на исходном тексте значение статистики $d(n) = S_{набл}$ с оценками, полученными в имитационном эксперименте. Считаем, что аномальное (неслучайное) отклонение от равномерности типа «кластеризация» имеет место, если:

$$(S_{набл} \le S_{\min}) \,\&\, (S_{набл} \le \overline{S} - 3s). \qquad (*)$$

Обычно правое условие в критерии (*) работает более жестко, чем левое, т.е. отсеивает больше претендентов на «аномальность», однако бывают случаи, когда выполняется правое условие, но не выполняется левое.

5). Значимость выделенного кластера удобно характеризовать безразмерной величиной $\delta(x) = u(x)/v(x)$, где $u(x) = N/F(x)$ — среднее расстояние между вхождениями слова x в текст, а $v(x) = d(n)/n$ — среднее внутрикластерное расстояние между вхождениями x. Показатель $\delta(x)$ следует использовать при не слишком малых n (например, $n > 5$), чтобы не придавать излишний вес тандемным вхождениям лексической единицы, вероятность которых при неучете знаков пунктуации довольно велика.

С помощью показателя $\delta(x)$ может быть осуществлена дополнительная фильтрация слов, демонстрирующих аномальную кластеризацию. При этом следует иметь в виду, что кластеры с относительно высоким $\delta(x)$ (например, $\delta(x) \ge 5$) обычно содержат не слишком много точек ($n \sim 6 \div 12$) и характеризуют локальные подтемы

или эпизоды в тексте. Более разреженные кластеры с относительно небольшими значениями $\delta(x) \sim 2 \div 3$ покрывают весьма значительные фрагменты текста, но характеризуют, скорее, предметную область в целом, нежели конкретное содержание документа.

Выявленная при фиксированном значении n сильная аномалия не может исчезнуть мгновенно. Поэтому при увеличении n возникает система вложенных или пересекающихся кластеров со все меньшими значениями $\delta(x)$. Они заменяются одним (максимальным по размеру и числу точек) кластером, удовлетворяющим ограничению $\delta(x) \geq \delta_{пор}$. Среднечастотные слова, демонстрирующие позиционные аномалии, обычно характеризуются одним таким кластером. Высокочастотные слова могут иметь два — три независимых кластера.

2. Построение квазиреферата

Рассматриваются два способа формирования квазиреферата на основе позиционно кластеризованных лексических единиц. Первый связан с построением *профиля кластеризуемости* лексических единиц в тексте и отслеживанием точек изменения значений этой ступенчатой функции. Второй способ связан с приписыванием каждому предложению веса в соответствии с наличием в нем кластерообразующих лексических единиц и отбору для квазиреферата предложений с максимальным весом.

При определении профиля кластеризуемости под размером кластера будем понимать число предложений, покрываемых полностью или частично (на границах) выделенным интервалом из $d(n)$ слов. Если кластер представлен одним предложением, его целесообразно проигнорировать. Профиль кластеризуемости — это ступенчатая функция, аргументом которой является порядковый номер предложения в тексте, а значение равно числу различных кластеров, включающих в себя данное предложение. Поскольку каждый кластер связан с определенной лексической единицей, то значение профиля в каждой точке фиксирует совокупность лексических единиц, определяющих, локальное содержание данного участка текста. При этом в рассматриваемом предложении вовсе не обязаны присутствовать одновременно все слова из этой совокупности.

Простейшая стратегия построения квазиреферата состоит в фиксации моментов изменения профиля, т.е. переходов с низкой ступени на более высокую и наоборот. Это значит, что мы реагируем на начало и конец каждого кластера, отбирая для квазиреферата его первое и последнее предложение. Такая стратегия характерна для позиционных методов реферирования [4], суть которых сводится к учету начальных и конечных фрагментов в структуре текста, задаваемой автором. Так, считается, что в тексте научной статьи наиболее важными являются (наряду с заголовком) введение и заключение, в каждом из подразделов — начальный и конечный абзац, в каждом абзаце — начальное и конечное предложение. Отметим, что далеко не все научные статьи, не говоря уж о текстах других жанров, разбиты на разделы и подразделы, что является дополнительным аргументом в пользу развиваемого подхода.

Вторая стратегия построения квазиреферата состоит в назначении весов каждому предложению текста. Вес предложения определяется числом вхождений в него словоформ, демонстрирующих кластеризацию в произвольном месте текста, т.е. само предложение может и не входить в состав соответствующего кластера. Вариант, когда вес предложения фиксирует разнообразие представленных в нем кластеризованных словоформ, а не полное их количество, представляется более предпочтительным.

3. Экспериментальная проверка методики.

Апробация описанного выше подхода проводилась на полнотекстовых документах разного жанра (научные статьи, главы художественных произведений, газетные публикации). В качестве основной лексической единицы рассматривалась словоформа. Словосочетания на данном этапе не привлекались за исключением аббревиатур типа ЛФ (лексическая функция), США и т.п., которые приравнивались к словоформам). Порог отбора лексических единиц по частоте ($F_{пор}$) равнялся четырем, а порог отбора кластеров по параметру δ — трем. Из немногочисленных известных программ аналогичного назначения, ориентированных на русский язык, нам удалось воспользоваться для сравнения результатов лишь программой TextAnalist (http://www.analyst.ru), имеющей, к сожалению, ограничение на длину обрабатываемого текста.

Не имея возможности привести тексты первоисточников и разные варианты квазирефератов для них, ограничимся лишь кратким описанием результатов двух экспериментов по обработке: 1) главы 6 книги А.А. Милна «Вини-Пух и все-все-все» в переводе Б. Заходера; 2) статьи И.А. Большакова «Какие словосочетания следует хранить в словарях?» из материалов конференции «Диалог—2002».

Глава 6 из книги Милна содержит описание дня рождения ослика Иа-Иа. Приведем достаточно короткий квазиреферат этой главы, полученный описанным выше методом назначения весов предложениям и отбором предложений с пороговым весом 3 и выше.

(1) Бедный ослик ужасно расстроен, потому что у него сегодня день рождения, а все о нем забыли, и он очень понурился — ты ведь знаешь, как он умеет, — ну и вот он такой понурый, а я…

(2) Если я его как следует вымою и попрошу кого-нибудь написать на нем: «Поздравляю с днем рождения», Иа сможет держать в нем все, что хочешь.

(3) — А ты что думаешь ему подарить? — Я несу ему в подарок Полезный Горшок, в котором можно держать все, что хочешь, — сказал Пух.

(4) — Тут когда-то держали мед, — сказала Сова.

(5) — В нем можно что хочешь держать, серьезно сказал Пух.

(6) — Славный горшочек, — сказала Сова, оглядев горшок со всех сторон.

(7) Воздушные шары не входят в горшки.

(8) — Это другие шары не входят, а мой входит, — с гордостью сказал Иа-Иа.

(9) — Мне очень приятно, — радостно сказал Пятачок, — что я догадался подарить тебе Полезный Горшок, куда можно класть какие хочешь вещи!

(10) Ему было не до того: он то клал свой шар в горшок, то вынимал его обратно, и видно было, что он совершенно счастлив!

Если не считать некоторого дублирования во фразах (2) и (3), качество реферата можно признать вполне удовлетворительным. Реферат, сделанный программой TextAnalyst, существенно длиннее и содержит серьезный изъян: общеупотребительное слово «сказать» (с высокой частотой встречаемости) принято за «тематически значимое»; в итоге, в реферат попало много неинформативных предложений, содержащих это слово.

Второй эксперимент — построение квазиреферата статьи И.А. Большакова нашим методом и с помощью программы TextAnalyst —- дал сопоставимые результаты. Реферат, построенный по профилю кластеризуемости, более детален и разнообразен по части учета ключевых слов. Реферат TextAnalyst'а удачнее по концовке, но несколько монотонен в плане непрерывного (почти в каждом предложении) использования термина «словосочетание». Оно указано автором как ключевое, но по сути является общетематическим для «Диалога» и не несет конкретной информации о содержании статьи. Реферат, полученный путем назначения весов, имеет содержательную концовку, но из него исчезали два весьма информативных подзаголовка, вошедшие в «профильный реферат», а именно: «Терминологические словосочетания языково специфичны» и «Статистический подход сомнителен». Причина здесь в том, что короткие предложения объективно имеют меньше шансов набрать большой вес.

4. Факторы, влияющие на качество квазиреферата

1). Существенным фактором является правильная разбивка текста на слова и предложения в автоматическом режиме (аббревиатуры, многоточия, прямая речь и т.п.)
2). Ошибки в работе процедуры нормализации при «опознании» новых слов могут явиться причиной раздробления кластера или его потери.
3). Использование в качестве базовых единиц лишь одних словоформ, т.е. игнорирование словосочетаний, может привести к потере значимой информации. Напрмер, в статье Большакова словосочетание «лексическая функция» кластеризовано, а словоформа «лексическая» — нет, поскольку встречается и в других комбинациях (лексическая омонимия, лексическая единица и т.п.).
4). Учет семантических повторов (типа: ЛФ = «лексическая функция») и анафорических ссылок может повлиять на решение о наличии или отсутствии кластера либо изменить его границы. Раскрытие анафорических ссылок в тексте самого квазиреферата желательно с целью повышения связности текста.
5). Обработка формульных выражений и списка литературы может привести к появлению ложных кластеров из-за явления омонимии (P — как обозначения вероятности события в тексте и как обозначения номера страницы в списке литературы и т.п.).

Учет перечисленных выше факторов вполне реален и может способствовать повышению качества квазиреферата.

Заключение

Предложен метод квазиреферирования текста, основанный на формальном выделении сверхфразовых единств с помощью сканирующих статистик и извлечении из них наиболее информативных предложений для квазиреферата. Метод ориентирован на полнотекстовые документы (научные статьи, газетные публикации, информация, размещенная на Интернет-сайтах) и может работать с неструктурированными данными. Временны'е затраты незначительны, поскольку имитационное моделирование необходимо лишь на этапе обучения.

Литература

1. Пащенко Н.А., Кнорина Л.В., Молчанова Т.В. и др. Проблемы автоматизации индексирования и реферирования // Итоги науки и техники. Информатика, 1983 г. Т. 7, С. 7—164.
2. Бондаренко Г.В. Распределение повторов в связном тексте как основа для обнаружения суперсинтаксических единиц // НТИ, 1975. сер.2, № 12, С. 20—31.

3. Бондаренко Г.В., Яровенко О.И. Использование структурных закономерностей текста при автоматической обработке информации // НТИ, 1984. сер.2, № 3, С. 23—29.
4. Гиндин С.И. Позиционные методы автоматического фрагментирования текста, их теоретико-текстовые и психолингвистические предпосылки // Семиотика и информатика. М.: ВИНИТИ, 1978. Вып. 10, С. 32 — 73.
5. Naus J.I. The distribution of the size of the maximum cluster of points on a line // J. Amer. Statist. Assoc, 1965. Vol. 61, № 310, P. 532—538.
6. Wallenstein S.R., Naus J.I. Probabilities for a *k*-th nearest neighbor problem on the line // The Annals of Probability, 1973. Vol. 1, № 1, P. 188—190.
7. Гусев В.Д., Немытикова Л.А., Саломатина Н.В. Выявление аномалий в распределении слов или связных цепочек символов по длине текста // Интеллектуальный анализ данных. — Новосибирск, 2002. — вып. 171: Вычислительные системы. С.51—74.