Вестн. Ом. ун-та. 2011. № 4. С. 50–55.

УДК 519.6:410

О.Г. Чанышев

АВТОРЕФЕРИРОВАНИЕ НА ОСНОВЕ АССОЦИАТИВНЫХ ПОЛЕЙ ДОМИНАНТ*

Дается формальное определение ассоциативных полей доминант естественно-языкового текста. Предлагается метод автореферирования текста на основе ассоциативных полей доминант. Приведены результаты компьютерных экспериментов.

Ключевые слова: автоматическое реферирование, ассоциативное поле, доминанты.

Введение

Автоматическое реферирование (точнее, квазиреферирование – отбор подмножества предложений исходного текста для его представления пользователю [1]) вот уже на протяжении десятилетий остается актуальным направлением исследований. Можно привести обширнейший список литературных источников, содержание которых связано с различными аспектами этого процесса. Однако оценка «правильности» предлагаемых методов, подавляющее большинство из которых основывается на интуиции разработчика, сталкивается с отсутствием онтологической (в философском смысле) теории автоматического анализа текста. Работы, основанные на лингвистических теориях (например, [2]), лишь частично допускают программную формализацию, поскольку «авторы лингвистических моделей зачастую явно или неявно апеллируют к языковой интуиции человека, носителя описываемого языка, опуская ряд «очевидных» деталей, чрезвычайно существенных при автоматической обработке текста» [3]. В качестве примера можно привести часто используемое при лингвистическом описании структуры текста понятие *межфразовое единство (или сложное синтаксическое целое)*, «которое организуется через *тема-рематическую последовательность*» ([4]). Но «любой член (или члены) предложения в соответствии с контекстом или ситуацией может выступать как тема или рема…» [5].

В работах [6; 7] была представлена «Ассоциативная модель реального текста», основанная на базовых нейролингвистических представлениях об образовании ассоциаций и ограниченности кратковременной памяти. В ее рамках модель автомата-«писателя» [8] позволила показать, что распределение весов предложений, складывающихся из суммы весов входящих слов (вес слова – частота или ассоциативная мощность) имеет пилообразный вид именно из-за ограниченности кратковременной памяти, что приводит к необходимости повторения наиболее важных слов. В свою очередь это дало обоснование отбору для автореферата предложений с локальными максимумами весов (в дальнейшем такой алгоритм и результирующий автореферат будем называть базовыми). Предложения базового автореферата тематически правильно представляют текст в целом. Коэффициент реферирования – отношение числа отобранных предложений к числу предложений текста – не зависит от жанров анализируемых текстов и оста-

* *Работа выполнена по проекту № 1.4.2.ОМН РАН за 2007 г.*

ется почти постоянным (близким к 1/3). Но такой реферат по определению состоит из несмежных предложений исходного текста, что иногда затрудняет целостное восприятие содержания.

Ассоциативная модель уточняла понятие тематического веса слова и позволяла выделять из текста наиболее важные слова (доминанты). В процессе развития модели в область задач кластеризации и классификации текстов [9] был сделан вывод о возможности использования кластеров доминант для определения смыслового подобия одинаковых доминант в различных текстах. Поскольку предложения текста являются элементарным способом образования ассоциаций между входящими словами, такие кластеры естественно назвать ассоциативными полями доминант.

В цитированной выше работе они названы (несколько поспешно) гипонимическими семантическими полями, поскольку в психолингвистике понятия семантических и ассоциативных полей тесно связаны. Последние, получаемые в ходе ассоциативного эксперимента, используются «для экспериментального исследования субъективных семантических полей слов, формируемых и функционирующих в сознании человека, а также характера семантических связей слов внутри семантического поля» [10]. Имея в виду выделенность доминанты, ее «возвышенность» над другими элементами поля, можно было бы характеризовать ассоциативное поле как «гипонимическое», однако это сразу переводит нас в область гиперонимов – суперконцептов, гипонимов, как наиболее «частотных культурно-обусловленных ассоциатов» [11]. Переводит ранее нашей готовности к алгоритмическим интерпретациям понятий лингвистической теории гипонимии.

В настоящей статье дается точное определение понятия ассоциативного поля доминанты, описывается метод определения их состава и метод автореферирования по ассоциативным полям. Приводятся результаты компьютерных экспериментов.

Автор полагает, что положительные результаты экспериментов в различных областях автоматического анализа текста, полученные методами, основанными на использовании ассоциативных полей доминант, дадут основание рассматривать концепцию ассоциативных полей как базу для построения «вычислительной» теории текста.

1. Ассоциативные поля доминант

1.1. Ранее введенные определения

Областью существования слова (Q) называется множество включающих его предложений (или, в зависимости от контекста, множество номеров включающих предложений).

Независимыми лексемами связи (НЛС) называются слова (или постоянные сочетания слов, не обязательно контактные) текста, отвечающие следующим условиям:

– они не принадлежит множеству слов стоп-словаря,

– частоты слов (повторяемость в различных предложениях текста) больше 1,

– для произвольной пары НЛС найдется минимум два предложения, в которые они входят по отдельности.

Если Q_j есть подмножество Q_i то j-е слово называется *атрибутом* i-го.

Ассоциативной мощностью НЛС называется число других НЛС, входящих в предложения ее области существования, за исключением первого.

Частично упорядочим НЛС по убыванию ассоциативной мощности. Пронумеруем группы с одинаковыми значениями ассоциативной мощности натуральным рядом чисел от 1 до R. Слово будет иметь ранг, равный номеру группы. При анализе отдельных текстов в качестве веса слова можно использовать как значение ассоциативной мощности, так и величину, обратную рангу.

Введем обозначения:

R – максимальный ранг слова,

r – ранг слова,

ω – частота, ψ – ассоциативная мощность.

Многочисленные эксперименты показали, что зависимости $r(\psi)$ или $r(\omega)$ для НЛС в диапазоне от R до $0{,}5R$ имеют вид прямой $r = R - \omega$ или $r = R - \psi$. Далее кривая зависимости состоит из отрезков прямых, совместно образующих «псевдогиперболу» в соответствии с законом Ципфа-Мандельбродта ($r\omega = Const$). Слова со значениями $\psi > 0{,}5R$ названы *доминантами*. Их число не превышает 4 % от размера множества всех слов текста (без служебных слов).

1.2. Близости слов и ассоциативные поля доминант

Обозначим через Q_i и Q_j области существования i-го и j-го слов текста. Введем меру близости слов:

$$K_{i,j} = \frac{\rho(Q_i \cap Q_j)}{\rho(Q_j)} \qquad (1)$$

где ρ(*M*) – размер множества *M*.

Полным ассоциативным полем *(F_i^+) i*-ой доминанты называются все *j*-е слова, для которых $K_{i,j} > 0$.

Основным ассоциативным полем *(F_i) i*-ой доминанты называются все *j*-е слова, для которых $K_{i,j} > 0{,}5$.

Очевидно, что размер полного ассоциативного поля может существенно превышать размер основного.

Атрибутивные слова входят в ассоциативные поля по определению и их близость к доминанте также определяется по формуле (1).

Ассоциативные поля определяются последовательно начиная с первого элемента списка доминант, упорядоченного по убыванию ассоциативной мощности. Если *j*-ое слово является доминантой и $K_{i,j} > 0{,}5$, то для этого слова ассоциативное поле не определяется.

Фразеологическим ассоциативным полем доминанты называется объединение (без повторов) областей существования элементов ее лексического ассоциативного поля.

В связном тексте хотя бы часть полных ассоциативных полей должны иметь общие элементы. Тогда формальную меру связности текста можно определить следующим образом:

$$C = \frac{1}{N_s} \sum_{i \neq j}^{N} \rho(F_i^+ \cap F_j^+), \qquad (2)$$

где *N* – общее число полных ассоциативных полей в фиксированном тексте; *Ns* – число предложений в тексте.

Использование F_i либо F_i+ будет зависеть от решаемой задачи. Например, в задаче кластеризации-классификации текстов, требующей максимальной точности в определении подобия одинаковых доминат в различных текстах [9], использовались F_i поля. Если же требуется построить сеть ассоциативных полей (например, для определения различных путей в тексте) путем определения общих элементов полей, то, очевидно, следует использовать F_i+ поля.

2. Автореферирование

Предварительно введем следующие определения:

– *информативность* реферата – отношение числа доминант, присутствующих в отобранных предложениях автореферата к общему числу доминант в тексте;

– *коэффициент близости предложений* реферата – отношение суммы разностей между номерами отобранных предложений исходного текста (упорядоченных по возрастанию) к числу интервалов (число отобранных предложений – 1);

– *коэффициент реферирования* – отношение числа отобранных предложений к общему числу предложений исходного текста.

Задачу автореферирования сформулируем так: *требуется выбрать возможно меньшее число предложений исходного текста, но так, чтобы информативность реферата была равна 1 (присутствовали все доминанты) при необходимом условии: значение коэффициента близости предложений должно быть больше, чем у базового автореферата.*

Для достижения указанной цели было решено использовать фразеологические ассоциативные поля доминант, полагая, что объединение предложений области существования доминат-вершин с областями существования элементов их поля не слишком увеличит общее число предложений, но, возможно, заполнит семантические лакуны.

Процесс автореферирования состоит из следующих этапов.

1) Определение лексических и фразеологических ассоциативных полей доминант.

2) Упорядочивание вершин полей по возрастанию или убыванию их ассоциативных мощностей.

3) Последовательное объединение фразеологических ассоциативных полей.

Процесс завершается при достижении единичного значения информативности либо после исчерпания фразеологических полей.

3. Эксперимент

Основная задача эксперимента заключалась в проверке:

а) предположения о зависимости коэффициента реферирования от определенного нами критерия связности (2);

б) возможности увеличить коэффициент близости предложений реферата по сравнению с базовым, не превышая существенно аналогичный показатель последнего, при одинаковой информативности.

Экспериментальный материал – естественно-языковые тексты в линейном формате (*.txt), подразделенные на тематические группы:

лекции по тематике баз данных (С.Д. Кузнецов. *Введение в системы уп-*

равления базами данных (9 лекций); Г.М. Ладыженский. *Цикл статей по теме СУБД* (4 статьи);

монография по сетевым операционным системам (Н. А. Олифер, В. Г. Олифер. *Сетевые операционные системы,* подразделена на 10 глав);

тексты по тематике искусственного интеллекта (18 отдельных статей);

тексты по компьютерной лингвистике (15 отдельных статей);

тексты по психологии (Т.П. Пушкина. *Медицинская психология;* И. Смирнов, Е. Безносюк, А. Журавлёв. *Психотехнологии;* О.Н. Первушина. *Общая психология* (4 главы) и 12 отдельных статей – 18 текстов);

тексты по философии (Е.К. Дулуман. *Курс лекций по философии* (7 глав); Учебник *Введение в философию* (главы 2–16); А.Н. Суворова. *Введение в современную философию* (главы 2–3) – 24 текста);

романы А. Бушкова (8 романов) и Б. Акунина (4 романа).

4. Обсуждение результатов

1. Результаты автореферирования по полным ассоциативным полям (Приложение 1, Таблица 1) отражают общую тенденцию – при уменьшении значения формальной (формула (2)) связности коэффициент реферирования также уменьшается.

При отборе доминант-вершин даже в порядке возрастания ассоциативной мощности коэффициент реферирования быстро становится слишком большим (больше 0,5) при значениях информативности, меньших 1.

2. При автореферировании по основным ассоциативным полям значение коэффициента реферирования не зависит от жанров анализируемых текстов (деловая или художественная проза) и остается в пределах (0,352–0,386). Это же относится и к базовым авторефератам, в которых значение этого же показателя еще более постоянно (0,324–0,333). Их объединяет только значение информативности, равное 1. Иными словами, для того, чтобы в реферате присутствовали все самые тематически важные слова (доминанты, не обязательно вершины полей), необходимо отобрать примерно треть предложений исходного текста. По мнению автора, этот феномен объясняется ограниченностью объема кратковременной памяти [8], что приводит к необходимости повторения наиболее значимых слов.

3. Как и следовало ожидать, при использовании основных ассоциативных полей коэффициент близости предложений увеличивается по сравнению с базовым рефератом. Реферат состоит из множества блоков предложений, в которых расстояние между номерами предложений не более 2 (см. Приложение 2). В этом же приложении приведены фрагменты текстов, из которых формировался блок. Опущенные предложения набраны мелким шрифтом.

5. Выводы

Автореферирование по методу объединения основных ассоциативных полей, когда доминанты-вершины последовательно выбираются из списка, упорядоченного по возрастанию ассоциативной мощности, позволяет получать информативные кусочно-связные рефераты в размере от 0,3 до 0,4 предложений исходного текста.

ЛИТЕРАТУРА

[1] *Солтон Дж.* Динамические библиотечно-информационные системы. М. : МИР, 1979. 560 с.

[2] *Мельчук И. А.* Опыт теории лингвистических моделей Смысл-Текст. М. : Наука, 1974. 314 с.

[3] *Сулейманов Д. Ш.* Аналитический обзор отечественных и зарубежных работ обработки естественного языка в аспекте прагматически-ориентированного подхода // Информационные технологии и телерадиокоммуникации : электронный журнал. 1999. № 1. URL: http://ittc.ksu.ru/?id=4.

[4] *Валгина Н. С.* Теория текста. М. : Логос, 2003. 279 с.

[5] Лингвистический энциклопедический словарь. URL: http://tapemark.narod.ru/les/022f.html

[6] *Чанышев О. Г.* Ассоциативная модель естественно-языкового текста // Вестник Омского государственного университета. 1997. Вып. 4. Омск : ОмГУ, 1997. С. 17–20.

[7] *Чанышев О. Г.* Ассоциативная модель реального текста и ее применение в процессах автоиндексирования // Труды Седьмой национальной конференции по искусственному интеллекту с международным участием КИИ'2000. М. : Физико-математическая литература, 2000. С. 430–438.

[8] *Чанышев О. Г.* Обнаружение закономерностей в реальных текстах при помощи автомата-«писатель» // Четвертый сибирский конгресс по прикладной и индустриальной математике : тез. докл. Ч. III. Новосибирск : Изд-во Ин-та математики, 2000. С. 106.

[9] *Чанышев О. Г.* Метод автоматической кластеризации текстов на основе анализа пересечений кластеров доминант // Информационные технологии. 2010. № 11(171). С. 2–7.

[10] *Глухов В. П.* Основы психолингвистики : учеб. пособие для студентов педвузов. М. : АСТ: Астрель, 2005. 351 с. URL: http://www.pedlib.ru/Books/4/0356/4_0356-299.shtml.

[11] *Котцова Е. Е.* Гипонимия в лексической системе русского языка (на материале глагола) : автореф. дис. ... д-ра филол. наук. URL: http://dibase.ru/article/02082010_kottsovaee/1.

ПРИЛОЖЕНИЯ

Принятые обозначения:
C – связность (значение определяется по формуле (2));
№ – номер текста в подборке;
Ns – число предложений исходного текста;
I – информативность;
Kn – коэффициент близости предложений;
Kr – коэффициент реферирования.

УТОЧНЕНИЯ:

b – базовый автореферат;
1 – автореферат, полученный методом последовательного объединения фразеологических ассоциативных полей.

Приложение 1

Таблица 1

Средние значения коэффициента реферирования и связности при автореферировании по полным ассоциативным полям

Группа файлов	*C*	*Kr*
СУБД	1,418	0,640
Сетевые операционные системы	1,332	0,699
Комп. лингвистика	1,237	0,539
Философия	1,128	0,560
Искусств. интеллект	0,914	0,492
Психология	0,893	0,471
Романы Бушкова	0,646	0,283
Романы Акунина	0,517	0,271

Приложение 2

Примеры блочной структуры рефератов, полученных по основным ассоциативным полям

2.1. Текст № 2 из подборки по философии (Е.К. Дулуман. *Курс лекций по философии.* Античная философия). Число предложений – 373, коэффициент реферирования – 0,335.

[7 9 11 13 15]... 20... 26... [30 32]... 35... 40... 45... [49 50 52 53 54 55]... [59 60 61]... [66 68 69]... 72... [81 82]... [88 89 91 92]... [95 96 97 98 100 102 103]... 109... 115... [131 132]... 135... 140... [150 151 152]... 165... [176 177]... [181 182 183 185 186 188]... 191... 195... 199... 202... 206... 209 [210 211 212]... [215 216]... [219 221 223 224 225 227 229]... 234 236... [239 241 243 244 245]... [248 250]... [253 255 257 258 259]... 262... [265 267 269 270 271 272 273 274]... [278 279 280]... [292 294]... [297 298]... [302 303 305]... 312... [320 322]... [327 328]... 333... 337 [338 339 341]... [350 351]... [355 357]... 365... [370 372]

Первый блок:

<7>Античная (от латинского слова "antiquius" – давний, старый, древний) философия – это та философская мысль, которая возникла и развивалась в государствах Средиземного моря – в рамках Греции и Римской империи – с конца 7 столетия до нашей эры до 6 столетия нашей эры.<8>Античная философия возникла и сформировалась сначала в Древней Греции, которая на заре человеческой истории появилась еще в начале второго тысячелетия до нашей эры – в эпоху бронзового века.<9>За полутора тысячи лет Греция прошла огромный исторический путь: от варварства – до цивилизации, от родоплеменных объединений – до государственного устройства всего греческого народа, и уже в 8 столетии до нашей эры стала наиболее развитым и передовым – и в политическом, и в экономическом, и в культурном отношениях – государством того времени.<10>И за все это время как в обществе, так и сознании отдельных греков единым мировоззрением было мировоззрение религиозное. <11>С начала 1-го тысячелетия до нашей эры наличное у греков религиозное мировоззрение начинает переосмысливаться сначала в художественной форме (творчество Гомера) и с позиций мифологии и здравого смысла (творчество Гесиода). <12>Правда, произведения Гомера "Илиада" и "Одиссея" и Гесиода "Труды и дни" и "Теогония"(Происхождение богов) сразу же после их появления были провозглашены священным писанием греческой религии – Олимпийского пантеону. <13>Гомер и Гесиод положили начало художественному и теоретическому переосмыслению религиозного мировоззрения. <14>Однако философская мысль Древней Греции развивалась не в русле религиозного мировоззрения, как то мы видели в Индии, и не на путях создания новой религии, как то было в Китае, а в решительном и однозначном противопоставлении себя религии.<15>Даже верующие в существование богов философы (Сократ, Платон, Ксенофан) беспощадно критикуют современную им религию, высмеивают мифы Гомера и Гесиода, неизменно утверждают, что философия с презрением смотрит на религиозные верования толпы.

2.2. Текст № 1 из подборки по СУБД (С.Д. Кузнецов. *Введение в системы управления базами данных.* Пролог.). Число предложений – 231, коэффициент реферирования – 0,346.

[2 4]... 7... 12... 15... [19 21 23]... 28... 31... 36... [40 42]... [46 48 50]... [53 55]... 58... [62 64 66]... 70... 75... [78 80 82]... 85... 88... 91... 94... 97... 102... 105... [108 110 112 114]... [117 119 121 123 125 127]... [130 132]... 135... 138... [142 144]... 148... 151... [154 156]... 160... [163 165 167]... 170... 173... 176... 179... 184... [187 189 191]... [194 196 198 200]... 204... [207 209]... 214... 219... [223 225]... 229

Второй блок:

<19>В классической форме будут определены реляционная алгебра и реляционное исчисление, два эквивалентных механизма, на которых базируются современные языки манипулирования базами данных. <20>Третья часть цикла посвящается проектированию реляционных БД.<21> Будут описаны два подхода: на основе пошаговой нормализации исходного универсального отношения и с использованием семантических моделей данных, из которых наиболее распространена практически модель "сущность-связь". <22>В четвертой части рассматриваются методы внутренней организации многопользовательских реляционных СУБД.<23>Будут проанализированы методы организации внешней памяти реляционных БД и специальных служебных структур внешней памяти – индексов, служащих для оптимизации выполнения запросов к БД.

Вестн. Ом. ун-та. 2011. № 4. С. 55–60.

УДК 544.344

В.В. Прудников, П.В. Прудников, И.С. Попов, С.В. Алексеев

ИССЛЕДОВАНИЕ ЭФФЕКТОВ СТАРЕНИЯ И НАРУШЕНИЯ ФЛУКТУАЦИОННО-ДИССИПАТИВНОЙ ТЕОРЕМЫ В ДВУМЕРНОЙ XY-МОДЕЛИ ПРИ МОДЕЛИРОВАНИИ ИЗ НАЧАЛЬНОГО СОСТОЯНИЯ С МАЛЫМ ЗНАЧЕНИЕМ НАМАГНИЧЕННОСТИ*

Исследуется явление старения в низкотемпературном неравновесном критическом поведении двумерной XY-модели методами Монте-Карло из начального неупорядоченного состояния. Исследуется неравновесное поведение двухвременных автокорреляционной функции и функции отклика.

Ключевые слова: двумерная XY-модель, явление старения, флуктуационно-диссипативная теорема, коротковременная динамика;

В последние годы исследование систем, характеризующихся медленной динамикой, вызывает значительный интерес как с теоретической, так и экспериментальной точек зрения. Это обусловлено предсказываемыми и наблюдаемыми в них свойствами старения при медленной эволюции систем из неравновесного начального состояния и нарушениями флуктуационно-диссипативной теоремы [1]. Хорошо известными примерами подобных систем с медленной динамикой и эффектами старения являются такие комплексные неупорядоченные системы, как стекла: дипольные, металлические и спиновые стекла. Однако данные особенности неравновесного поведения, как показали различные аналитические и численные исследования, могут наблюдаться и в структурно однородных системах в критической точке или вблизи нее при фазовых переходах второго рода, так как критическая динамика таких систем характеризуется аномально большими временами релаксации. К системам с медленной динамикой относится и двумерная XY-модель при температурах ниже и равной температуре T_{KT} фазового перехода Березинского–Костерлица–Таулеса [2]. Под процессом старения материалов понимают явление роста времени релаксации системы к состоянию равновесия с увеличением «возраста» материала, т. е. времени, прошедшего после приготовления образца [3]. Явление старения проявляется математически прежде всего в двухвременных характеристиках системы, таких как корреляционные функции и функции отклика. При неравновесных процессах эти функции зависят от двух переменных временной природы: t и t_w, при $t > t_w$, и не только от их разницы, но и от каждой в отдельности. Причем эта зависимость сохраняется и при достаточно больших временах наблюдения t. Временная переменная t_w характеризует возраст образца, т. е. время, прошедшее после его приготовления, и называется временем ожидания. При явлении старения процесс релаксации

* Работа поддержана грантами Минобрнауки 2.1.1/13956 и 2010-1.1-121-011-047, грантом РФФИ 10-02-00507 и грантом Президента РФ МК-3815.2010.2.