УДК 004.021, 004.822 DOI: 10.17587/it.27.138-146

А. О. Корней, аспирант, e-mail: korney.alena@yandex.ru,
Е. Н. Крючкова, канд. физ.-мат. наук, доц., e-mail: kruchkova_elena@mail.ru,
Алтайский государственный технический университет им. И. И. Ползунова, Барнаул

Категоризация текстов на основе сконденсированного графа

Предлагается комбинированный семантико-статистический алгоритм аспектного анализа текстов большого объема, основанный на использовании семантического графа. Метод выделения аспектов содержит фазы выделения множества значимых слов, вычисления весов вершин семантического графа методом релаксации, фильтрации аспектов на основе градиентного метода. Рассматривается алгоритм построения таких домензависимых множеств наиболее значимых слов, которые характеризуются одинаковыми статистическими характеристиками для разных доменов и учитывают структуру и лексическое разнообразие текстов.

Ключевые слова: *семантический граф, категоризация текстов, семантико-статистический алгоритм, извлечение знаний, извлечение аспектных терминов*

Введение

Резонансные мировые события последних лет привели к увеличению количества информации в сети Интернет, в том числе криминальной, недостоверной, заказных негативных отзывов. Борьба с такого рода информацией сделала особенно актуальной задачу автоматического определения тематической направленности и краткого содержания текста, его эмоциональной окраски. Ложная негативная информация может распространяться очень быстро, поэтому разработка эффективных алгоритмов выявления такого рода информации является теоретической базой для практической реализации автоматизированных систем своевременного реагирования. Для большинства обычных информационных ресурсов с оценочными характеристиками пользователей о тех или иных событиях или явлениях, сервисных центрах или частных поликлиниках, фильмах или телепередачах эффективная автоматическая обработка позволяет быстро получать актуальное краткое содержание множества документов на одну тему или об одном объекте.

В данной работе рассматривается инструмент для категоризации текстов, обладающий относительно невысокой вычислительной сложностью. В основу инструмента положен семантический граф русского языка, содержащий обобщенные знания о мире.

Задача категоризации текстов

Задача категоризации текстов может рассматриваться как частный случай классификации и обычно включает четыре этапа: предобработка и индексация документов, уменьшение размерности пространства признаков, построение и обучение классификатора, оценка качества классификации. Первые два этапа, как правило, предполагают стандартный набор действий. Предобработка текста строится на основе токенизации, лемматизации, удаления стоп-слов и т.д. [1]; индексация — построение числовой модели документа — может быть основана на методах Bag of Words [2], Word2vec [3], TF-IDF [4], учете *n*-грамм [2] и т.д.

Вычислительная сложность алгоритмов классификации напрямую зависит от размерности пространства признаков. Поэтому разумной мерой повышения эффективности является взвешивание и уменьшение числа признаков. Для этого применяют, например, латентный семантический анализ (LSA) [3, 5], поточечную взаимную информацию (PMI) [6], линейный дискриминантный анализ (LDA) [1]. Кроме того, в работе [1] рассматриваются и другие методы: стохастическое вложение соседей с *t*-распределением (*t*-SNE), метод случайных проекций и т.д.

Наиболее важным шагом является непосредственно этап классификации. Подходы,

применяемые для построения классификаторов, очень разнообразны. Наиболее известны такие решения, как наивный байесовский классификатор (NBC) [7], классификатор на основе *k*-ближайших соседей (KNN) [8], а также метод опорных векторов (SVM) [9]. Более сложные современные решения связаны с методами машинного обучения, использованием нейросетей, LSTM [10] и т.д.

В связи с тем, что производительность является одним из критических аспектов при категоризации текстов, современные системы строятся по одному из двух принципов: без понижения размерности, но с использованием "быстрого" классификатора; с понижением размерности, но с более качественным классификатором. Второй вариант более предпочтителен, поскольку область его применения включает и те задачи, где "быстрые" классификаторы работают плохо.

В рамках данной работы рассматривается метод построения пространства признаков, основанный на анализе семантических графов. Предполагается, что для каждой категории может быть определен набор семантических подграфов (кластеров), включающих в себя данные о лексико-семантических и статистических характеристиках категории. За счет построения семантических кластеров вокруг ключевых понятий можно достичь понижения размерности, а внутренние данные подграфа (веса связей и вершин) могут использоваться в качестве весов отдельных признаков.

Семантический граф в системах информационного поиска

Современные информационно-поисковые системы работают с текстами без ограничения смыслового диапазона, поэтому используют в минимальной степени знания о мире и о языке, базируясь на статистических методах анализа. И причина одна: использование семантических знаний может существенно усложнить обработку текста. Тем не менее, стремление повысить качество информационных систем приводит к появлению современных исследований в области использования знаний при обработке документов. Например, на основе онтологий в работе [11] предлагается модель семантического поиска в больших коллекциях. Таким образом, использование систем управления знаниями не теряет актуальности, однако требует эффективных решений.

В рамках данной работы предлагается комбинированный семантико-статистический подход, когда в качестве знаний о мире используется семантический граф русского языка, автоматически построенный авторами на базе лингвистических словарей (толкового словаря и словаря синонимов). Вершинами графа являются канонические формы слов русского языка, связанные тремя типами нечетких отношений — синонимии, ассоциации и определения. Структура графа и анализ содержимого, извлеченного из словарей, приведены в работе [12]. Данные, извлеченные из общелингвистических словарей, не зависят от предметной области и при этом достаточно полны и потому могут быть использованы в системах информационного поиска в качестве источника общих знаний о мире, а информация о связях между отдельными понятиями может интерпретироваться различным образом в зависимости от конкретной задачи. Коллектив разработчиков успешно использовал этот граф для решения нескольких проблем, в том числе для семантического поиска в больших текстовых коллекциях[13], для классификации сложных изображений [14], в системах сентимент-анализа [15]. Семантические связи между словами в графе используются для расширения знаний системы, для извлечения неявной информации об отдельных лексических единицах. Такой подход позволяет скомпенсировать недостаточность статистической информации в системах, основанных только на статистике.

Состав и структура обучающих данных

При тестировании предлагаемого алгоритма были использованы два набора обучающих данных, соответствующие двум различным доменам — "Фильмы" и "Рестораны". По домену "Фильмы" использовался набор отзывов, автоматически извлеченных с сайта "КиноПоиск" [16], куда вошли положительные и отрицательные отзывы о фильмах различных жанров, эпох и с различным рейтингом. Для домена "Рестораны" использовался полный набор отзывов, опубликованный в рамках SemEval-2016 (Task 5, *Aspect based sentiment analysis*) [17]. На рис. 1 приведены численные характеристики, позволяющие оценить объем и лексическое разнообразие выбранных наборов данных.

Введем следующие обозначения: l — общее число слов в тексте; m — число уникальных канонических форм; $d = l/m$ — коэффициент постоянства корпуса слов. Величина d пока-

Рис. 1. Численные характеристики для оценки лексического разнообразия текстовых наборов

зывает, сколько слов в среднем приходится на одну каноническую форму, и чем выше это значение, тем менее лексически разнообразны тексты. Для домена "Фильмы" d = 12,825, для домена "Рестораны" d = 19,605, что позволяет оценить "Фильмы" как существенно более разнообразный домен. Структурные, количественные и семантические различия текстов из разных областей знаний требуют высокой универсальности алгоритмов обработки и извлечения информации.

Алгоритм построения конденсированного семантического графа на основе обучающей выборки

Для построения домен-специфичных подграфов в данной работе применяется метод конденсации исходного графа на основе данных, извлеченных из обучающей выборки. Алгоритм включает несколько этапов:

- фильтрация обучающих данных;
- релаксация на базе домен-специфичного каркаса и последующее отсечение;
- расчет центральностей и градиентов вершин;
- выбор ключевых терминов домена.

Каждый из перечисленных этапов строится на основе известных алгоритмов и концепций, а двукратное отсечение незначимых данных снижает вычислительную нагрузку.

Фильтрация обучающих данных

Вершины исходного семантического графа — это канонические формы слов, а дуги представляют связи между соответствующими словами и принадлежат одному из четырех типов: определение, ассоциация, синонимия, контекстная зависимость. Первые три типа дуг принадлежат семантическому графу изначально, а четвертый тип связи представляет собой контекстное усиление ассоциации, его будем достраивать в процессе обработки текста. Граф взвешенный, на первом этапе для построения каркаса в качестве веса вершин была выбрана частотность отдельных канонических форм, встречающихся в наборе документов в рамках домена. В качестве дополнительного источника информации при модификации графа использовалась частотность биграмм, сформированных из канонических форм слов. Выбор частотных характеристик биграмм обусловлен структурой исходного семантического графа: с точки зрения семантики наличие высокочастотных биграмм с большой вероятностью свидетельствует о том, что между словами биграммы в пределах выбранного домена существует ассоциативное отношение.

Очевидно, что для построения домен-специфичных подграфов следует выбирать слова и биграммы, несущие существенную смысловую нагрузку в пределах выбранной предметной области. Необходим механизм, позволяющий выбрать наиболее значимые слова с учетом разницы в структуре и составе текстов, характеризующих предметную область. Иными словами, необходимо ввести пороги отсечения по частотности для униграмм и биграмм, учитывая статистические характеристики доменов.

В квантитативной лингвистике для текстов ограниченного объема применяется эмпирический закон Хипса [18], связывающий число уникальных слов в тексте с его длиной:

$$V_R(l) = Kl^{\beta}, \qquad (1)$$

где $V_R(l)$ — число разных слов в тексте длины l; $10 \leqslant K \leqslant 100$ и $0{,}4 \leqslant \beta \leqslant 0{,}6$ — свободные параметры. Для выявления оптимального соотношения порогов частотности для доменов разной структуры переформулируем соотношение (1). Очевидно, что коэффициент K существенно варьируется от текста к тексту именно по причине разнообразия лексикона этого текста. Поэтому параметр l заменим в выражении (1) на d, где d — введенный нами коэффициент постоянства корпуса слов, тогда число значимых для домена слов, а значит, и ранг r (порядковый номер в упорядоченном по убыванию частотности списке) последнего, включаемого в рассмотрение слова, удовлетворяет условию

$r \sim 1/d^{\beta}$. В соответствии с известным соотношением [19, 20] между нормированным коэффициентом частотности c и его рангом r выполняется соотношение $c = 1/r$ или $r = 1/c$. Тогда $c \sim d^{\beta}$. Поскольку лексическое разнообразие домена в выражении (1) учтено в параметре d, мы можем считать, что при обработке разных текстов в целях получения результатов с одинаковыми статистическими характеристиками достаточно выбрать коэффициент пропорциональности в сотношении $c \sim d^{\beta}$ не зависящим от текста домена данной группы текстов (отзывы, научные статьи, ленты новостей и т.п.), тогда для двух доменов справедливо равенство

$$\frac{c_1}{c_2} = \frac{r_2}{r_1} = \frac{d_1^{\beta}}{d_2^{\beta}}. \qquad (2)$$

Выражение (2) связывает относительные частотности порогов отсечения слов, выбираемых для подграфов. Для перехода к абсолютным значениям необходимо умножить коэффициенты c_1 и c_2 на максимальные частотности f_1 и f_2 соответствующих доменов (для "Фильмов" $f_1 = 3069$, для "Ресторанов" $f_2 = 1303$). Табл. 1 демонстрирует связь абсолютных значений пороговых частотностей для разных β.

Согласно табл. 1 порог частотности отдельных канонических форм для домена "Фильмы" должен не более чем вдвое превышать порог для домена "Рестораны".

Более сложную структуру, чем униграммы, имеют n-граммы. Вероятность появления слов в одной n-грамме зависит от вероятности появления отдельных слов. Чем больше n, тем меньше вероятность совместного появления данного упорядоченного набора. В рамках данной работы авторы предлагают устанавливать соотношение порогов для n-грамм исходя из их длины и вычисленного соотношения порогов для униграмм. Пусть p_1 — соотношение пороговых значений для униграмм выбранной пары доменов, тогда для n-грамм соотношение порогов следует выбирать приблизительно равным $\sqrt[n]{p_1}$. В случае доменов "Фильмы" и "Рестораны" соотношение порогов выбирали близким к $\sqrt{1,99} \approx 1,41$. При этом в рассмотрение включали биграммы, в которых оба слова подходили под порог частотности, выбранный для униграмм.

Фактические частотности униграмм и биграмм, подсчитанные по данным обучающих выборок, действительно подчиняются закону обратной зависимости ранга слова и частотности. Для проведения экспериментов в рамках данной работы были выбраны пороги, представленные в табл. 2.

Таблица 1

Зависимость соотношения порогов от β

Параметр	Домен		Отношение
	Фильмы	Рестораны	
d	12,825	19,605	
f_{max}	3069	1303	
$f_{max} d^{0,4}$	8515,69	4284,41	**1,99**
$f_{max} d^{0,5}$	10990,71	5769,36	**1,91**
$f_{max} d^{0,6}$	14185,06	7768,99	**1,83**

Таблица 2

Экспериментальные и теоретические пороговые значения частотности для построения домен-специфических подграфов

Домен	Порог для униграмм	Порог для биграмм
Фильмы	50	16
Рестораны	25	11
Расчет	$\frac{50}{1,99} = 25,12$	$\frac{16}{\sqrt{1,99}} = 11,34$

Графики зависимости между рангом r и частотностью c униграмм, а также выбранные пороги представлены на рис. 2 и 3. Схожесть графиков для двух столь различных доменов позволяет оценить теоретические расчеты как достаточно верные и адекватные. Для доменов пороги экспериментально подобраны таким

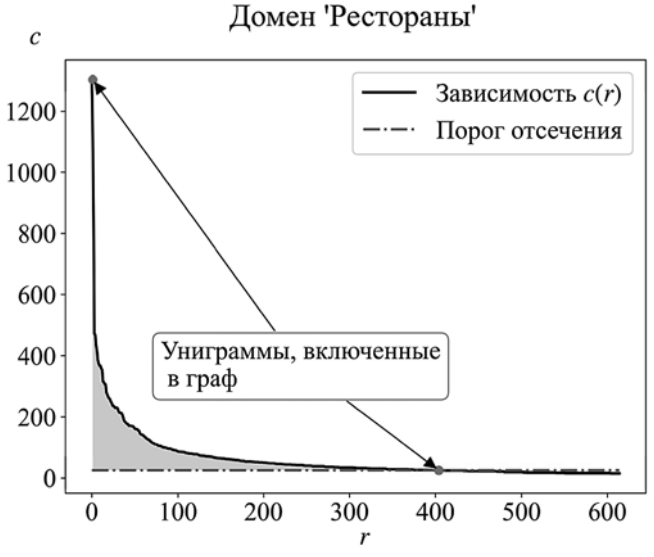

Рис. 2. Зависимость частотности униграмм от их рангов и порог отсечения для домена "Рестораны"

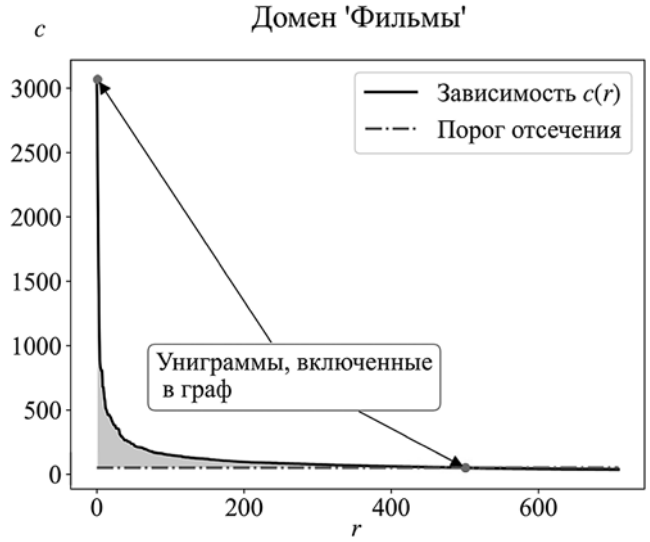

Рис. 3. Зависимость частотности униграмм от их рангов и порог отсечения для домена "Фильмы"

образом, чтобы низкочастотные и слабые по значимости биграммы и униграммы оказались за пределами рассмотрения и не включались в итоговые домен-специфические подграфы.

Релаксация на базе домен-специфичного каркаса

Пусть G — семантический граф, построенный на основе словарей. Для построения каркаса G_D домен-специфических подграфов используются слова и биграммы, отфильтрованные на основании значений из табл. 2. Пусть D — множество слов, удовлетворяющих выбранному порогу для униграмм, а B — множество биграмм, удовлетворяющих критериям выбора биграмм. Каркасы строятся на основании следующих правил:

- каждое слово $v \in D$ порождает вершину v графа G_D, которой присваивается вес $w_v = f_v$, где f_v — частота появления слова в домене;
- каждая биграмма $b \in B$ формирует двунаправленную *контекстную* ассоциативную связь в G_D между формирующими ее словами. Вес такой связи в проведенных экспериментах выбирался равным 0,8.

Полученный каркас G_D затем достраивается с использованием семантического графа G. Для этого используется процесс релаксации, в основе которого лежит классический алгоритм BFS — обход графа в ширину, модифицированный следующим образом:

- в качестве начальной вершины всегда берется очередное слово $v \in D$, которому присваивается текущий вес $\omega = \sqrt{w_v}$;
- поиск ведется параллельно по обоим графам G_D и G, G_D в ходе поиска динамически расширяется за счет включения домен-независимых вершин и связей из G. Веса w_u новых вершин u сначала принимаются равными нулю: $w_u = 0$;
- запускаем BFS от вершины v с весом ω. При посещении еще не рассмотренной очередной вершины v' ей передается релаксационный вес $\omega = r(u, v')$, где u — непосредственный предок вершины v'. Ребро (u, v'), принадлежит графу $G_D \cup G$, а $r(u, v')$ — функция релаксации;
- критерий окончания BFS: вершина не включается в очередь, если она уже была рассмотрена ранее, а также, если релаксационный вес, передаваемый ей, близок к 0 (не превышает некоторого ε).

Релаксация запускается для всех вершин из D независимо. При каждом запуске BFS соседям передаются только веса, полученные от текущей начальной вершины, а ранее накопленные веса не учитываются. Все релаксационные веса в конечном счете суммируются в итоговые значения Ω_u для всех $u = G_D$. Функция релаксации в данной работе имеет вид $r(u, v') = \alpha\omega_u f(u, v')$, где α — коэффициент затухания, не зависящий от домена; ω_u — текущий релаксационный вес вершины u и $f(u, v')$ — вес связи между u и v', определяемый как максимум из весов отношений синонимии, определения, ассоциации. В рамках данной работы эксперименты проводились для $\alpha = 0{,}5$. Веса отношений выбирались равными 0,3; 0,45 и 0,6 для определения, синонимии и ассоциации соответственно. При наличии доменной связи вес ассоциации увеличивался до 0,8. Выбор соотношения весов основан на предположении, что активное включение гипонимов и синонимов в доменный граф может понижать его специфичность, поскольку синонимия и гипонимия на общеязыковом уровне не всегда согласуется с внутридоменными отношениями.

После окончания релаксации проводится повторное отсечение малозначимых слов. В качестве метрики выбирается $p_u = \frac{\sqrt{w_u} + \Omega_u}{2}$. Порог для p_u устанавливается таким образом, чтобы после отсечения осталось не более 15 % от всех слов в графе. Фрагмент графа, полученного после релаксации и отсечения для домена "Рестораны", приведен на рис. 4.

В ходе экспериментов для домена "Фильмы" полный граф включал 4221 вершину, после отсечения — 521 (12,34 %). Для домена "Ресто-

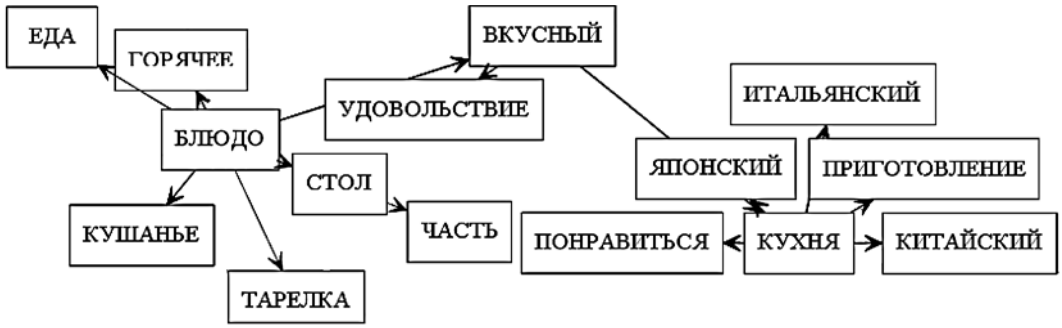

Рис. 4. Фрагмент семантического графа домена "Рестораны", образованный термином "блюдо" и его ближайшими соседями

раны" — соответственно 3142 и 253 вершины (8,05 %).

Расчет центральностей и градиентов вершин

Структура предложенного графа такова, что после релаксации на основе домен-специфической информации должны возникнуть области сгущения семантических данных за счет весов, присваиваемых вершинам. Для выявления центров сгущения авторы работы предлагают использовать две величины — центральность и градиент вершины.

Пусть $\sigma \in \{0, 1, 2...\}$ — ширина захвата контекста, т. е. при $\sigma = 0$ контекст вершины не учитывается вообще, при $\sigma = 1$ в контекст включаются вершины, непосредственно связанные с данной через исходящее ребро, — соседи первого порядка, при $\sigma = 2$ — соседи второго порядка и т. д. Использование большой ширины захвата контекста в данной задаче нецелесообразно, так как приводит к размыванию основных показателей значимости и снижению смысловой связности.

Введем величину центральности вершины. Пусть Ω_v — вес вершины, вычисленный в ходе релаксации; $G_D = (V, H)$ — домен-специфический граф с вершинами V и ребрами H. Тогда центральность вершины будем определять по формуле

$$C(v) = \begin{cases} \Omega_v, \sigma = 0; \\ \Omega_v \dfrac{\sum_{(v,u)\in H} \Omega_u}{\sqrt{\sum_{(v,u)\in H} \Omega_u^2}}, \sigma = 1; \\ \Omega_v \dfrac{\sum_{(v,u)\in H} C(u)}{\sqrt{\sum_{(v,u)\in H} C^2(u)}}, \sigma > 1. \end{cases} \quad (3)$$

Рассмотрим вектор градиента для функции $C(v)$:

$$g(v) = (C(v) - C(u_1), C(v) - C(u_2), \ldots, \ldots, C(v) - C(u_n)),$$

где $\{u_1, u_2, ..., u_n\}$ — соседи первого порядка для вершины v. Значение градиента определяется формулой

$$M_{g(v)} = \sqrt{(C(v) - C(u_1))^2 + (C(v) - C(u_2))^2 + \ldots + (C(v) - C(u_n))^2}. \quad (4)$$

В данной работе для графов, прошедших через процесс релаксации и отсечения незначимых вершин, были рассчитаны значения центральности для $\sigma = 1$ и значения градиентов. Обе вычисляемые характеристики позволяют ранжировать слова по значимости для того или иного домена, однако результаты ранжирования отличаются. В множество слов, выбранных по центральности, попадают слова, имеющие собственный высокий вес и находящиеся в центре большого "сгустка" вершин с достаточно высокими весами. Отсечение по градиенту позволяет выбрать множество вершин, окруженных большим числом соседей, но при этом обладающих собственным существенно более высоким весом по сравнению с ними. Пересечение этих множеств позволяет выбрать вершины, удовлетворяющие обеим характеристикам.

Выбор ключевых терминов домена

Для оценки согласованности ранжирования по центральности и градиенту, а также для выбора основной характеристики, позволяющей выделять ключевые термины домена, авторы работы провели ряд экспериментов.

Пусть $C(v)$ и $M_{g(v)}$ — значения функций центральности и градиента соответственно, а V_C и V_g — множества слов, отсортированных по убыванию $C(v)$ и $M_{g(v)}$. Пусть N — число наиболее важных терминов домена, которые должны остаться после отсечения по выбранной характеристике. Тогда можно говорить об абсолютном уровне отсечения — это значение центральности или градиента для последнего слова, включаемого в топ наиболее значимых. Обозначим абсолютные уровни отсечения $A_C(v_N)$ и $A_g(v_N)$ для V_C и V_g соответственно. Кроме этого, введем понятие относительного уровня отсечения: $a_C(v_N) = A_C(v_N) / \max_{v \in V_C} C(v)$ и $a_g(v_N) = A_g(v_N) / \max_{v \in V_g} M_{g(v)}$.

В ходе экспериментов для каждого из двух выбранных доменов выбирали различные значения N, а затем вычисляли число k слов, которые попали одновременно и в топ-N по центральности, и в топ-N по градиенту. Результаты экспериментов представлены в табл. 3 и 4.

Как видно из табл. 3, 4, центральность и градиент дают достаточно разный набор максимально значимых слов. Очевидно, что с ростом N согласованность растет, поскольку N приближается к полному охвату графа. Однако выбор больших N нецелесообразен, поскольку с ростом N падает значимость отдельных слов, включаемых в рассмотрение. Возникает необходимость выбора оптимального N, а также ведущей характеристики для выбора ключевых терминов домена.

Таблица 3

Оценка согласованности ранжирования по центральности и градиенту для домена "Фильмы"

N	$A_g(v_N)$	$A_C(v_N)$	$a_g(v_N)$	$a_C(v_N)$	k/N, %	k
100	457,27	77,718	0,054	0,053	56	56
70	583,054	104	0,069	0,07	49	34
50	631,624	127,641	0,075	0,086	54	27
45	642,041	135,648	0,076	0,092	53	24
40	647,329	149,817	0,077	0,101	58	23
35	711,55	163,329	0,084	0,111	60	21
30	1284,5	177,184	0,152	0,12	60	18

Таблица 4

Оценка согласованности ранжирования по центральности и градиенту для домена "Рестораны"

N	$A_g(v_N)$	$A_C(v_N)$	$a_g(v_N)$	$a_C(v_N)$	k/N, %	k
55	204,832	61,749	0,144	0,161	64	35
50	215,645	64,583	0,152	0,168	60	30
45	224,509	70,045	0,158	0,183	58	26
40	234,126	73,381	0,164	0,191	58	23
35	238,946	81,731	0,168	0,213	57	20
30	262,065	87,102	0,184	0,227	57	17

Проведем выборочный анализ терминов, попадающих в топ-20 для домена "Рестораны". В список наиболее значимых существительных при ранжировании по градиенту вошли такие слова, как: *место, кухня, время, ресторан, интерьер, блюдо, столик, обслуживание, салат, впечатление.* При ранжировании на основе центральностей в списке существительных остаются: *кухня, место, ресторан, время, интерьер, блюдо, салат, столик, впечатление, минута, день.* Очевидно, что в списки вошли значимые для оценки ресторанов термины. Объединение или пересечение списков позволяет формировать более обширные или узкие перечни аспектов, однако информативность и значимость все равно остаются на высоком уровне. При этом при ранжировании по частотности, без релаксации и расчета характеристик захват аналогичного набора терминов происходит приблизительно на уровне топ-50 (например, важный с точки зрения ABSA термин "впечатление" по частотности имеет ранг 46 и исключается из анализа).

Для домена "Фильмы" пересечение топ-40 содержит, в частности, следующий набор существительных: *фильм, время, герой, год, действие, жизнь, конец, момент, работа.* Представленный список содержит в себе ключевые термины, используемые при описании сюжета и написании рецензий. Кроме того, исходные списки топ-40 содержат такие важные слова, как *образ* (ранг по частотности 48), *сценарий* (92), *действие* (161), которые при использовании традиционного выбора по частотности в список не попадают. Таким образом, предложенный авторами подход позволяет эффективнее отбирать значимые для домена слова по сравнению с простым подсчетом частотных характеристик по домену.

Окончательное решение по выбору основного подхода к ранжированию слов может быть принято на основании условий конкретной задачи. К примеру, для задач ABSA может быть более применимо ранжирование по градиенту — на примере домена "Рестораны" такой подход позволяет извлечь разные типы лексики — для передачи смысла и отношения.

Заключение

Предложен и реализован комбинированный семантико-статистический алгоритм аспектного анализа текстовых документов, обладающий невысокой временной сложностью и направленный на комплексное решение задачи применения знаний о языке и о мире для повышения качества автоматической обработки текстов. Показано, что предложенный алгоритм позволяет значительно улучшить качество выделения категорий. Проведенные эксперименты подтверждают применимость предложенной формализованной модели для выбора соотношения порогов отсечения при обработке разных текстов в целях получения результатов с одинаковыми статистическими характеристиками. В проведенных авторами экспериментах фаза достройки семантического словаря и релаксации занимает не более 10 % от времени работы системы. Основной и наиболее затратной частью является подсчет частотных характеристик для обучающей выборки.

Список литературы

1. **Kowsari K., Jafari Meimandi K., Heidarysafa M., Mendu S., Barnes L., Brown D.** Text classification algorithms: A survey // Information. 2019. Vol. 10, N. 4. P. 150.

2. **Zhang X., Zhao J., LeCun Y.** Character-level Convolutional Networks for Text Classification // Proc. of the Neural Information Processing Systems Conf. (NIPS 2015). Montreal, Canada, 2015. URL: https://arxiv.org/abs/1509.01626 (accessed July 18, 2016).

3. **Ju R.** An Efficient Method for Document Categorization Based on Word2vec and Latent Semantic Analysis // 2015 IEEE Int. Conf. on Computer and Information Technology. 2015. P. 2276—2283.

4. **Pontiki M., Galanis D., Pavlopoulos J., Papageorgiou H., Androutsopoulos I., Manandhar S.** SemEval-2014 Task 4: Aspect based sentiment analysis // The 8th Intern. Workshop on Semantic Evaluation (SemEval 2014). Dublin, Ireland. 2014. P. 27—35.

5. **Medhat W., Hassan A., Korashy H.** Sentiment analysis algorithms and applications: a survey // Ain Shams Eng. Jour. 2014. N. 5. P. 1093—1113.

6. **Xu Y., Jones G. J., Li J., Wang B., Sun C.** A Study on Mutual Information-based Feature Selection for Text Categorization // Journal of Computational Information Systems. 2007. N. 3. P. 1007—1012.

7. **Dai W., Xue G. R., Yang Q., Yu Y.** Transferring naive bayes classifiers for text classification // In AAAI. 2007. Vol. 7. P. 540—545.

8. **Guo G., Wang H., Bell D., Bi Y., Greer K.** Using kNN model for automatic text categorization // Soft Computing. 2006. Vol. 10, N. 5. P. 423—430.

9. **Joachims T.** Text categorization with Support Vector Machines: Learning with many relevant features // NÉdellec C., Rouveirol C. (eds) Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence). Vol 1398. Berlin, Heidelberg: Springer. URL: https://doi.org/10.1007/BFb0026683

10. **Luan Y., Lin S.** Research on Text Classification Based on CNN and LSTM // 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China. 2019. P. 352-355. doi: 10.1109/ICAICA.2019.8873454.

11. **Курейчик В. В., Бова В. В., Лещанов Д. В.** Модель семантического поиска в системах управления знаниями на основе генетических процедур // Информационные технологии. 2017. Т. 23, № 12. С. 876—883.

12. **Крайванова В. А., Кротова А. О., Крючкова Е. Н.** Математическая модель естественного языка в задачах нечеткого ассоциативного поиска // XIV Международная конференция "Речь и компьютер" (SPECOM'2011). С. 402—406.

13. **Савченко В.** Алгоритм семантического поиска в больших текстовых коллекциях" // Supplementary Proceedings of the 3rd International Conference on Analysis of Images, Social Networks and Texts (AIST'2014). P. 161—166.

14. **Казаков М. Г.** Классификация сложных изображений на основе семантического графа понятий // Прикладная информатика. 2014. Т. 54, № 6. С. 79—89.

15. **Корней А. О., Крючкова Е. Н.** Анализ тональности коротких текстов на основе семантического графа // Робототехника и искусственный интеллект. Матер. X Всеросс. науч.-техн. конф. с международным участием. 2018. С. 168—174.

16. **КиноПоиск.** Все фильмы планеты. URL: https://www.kinopoisk.ru/.

17. **Pontiki M.** et al. SemEval-2016 Task 5: Aspect Based Sentiment Analysis // Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). 2016. P. 19-30. San Diego, California: The Association for Computational Linguistics. URL: https://doi.org/10.18653/v1/S16-1002.

18. **Heaps, Harold Stanley**. Information Retrieval: Computational and Theoretical Aspects. Inc.6277 Sea Harbor Drive Orlando, FL, United States, Academic Press, 1978.

19. **Zipf G. K.** Human behavior and the principle of least effort. Cambridge (Mass.): Addison—Wesley, 1949, pp. 573.

20. **Li W.** Random texts exhibit Zipf's-law-like word frequency distribution // IEEE Transactions on information theory. 1992. Vol. 38, N. 6. P. 1842—1845.

A. O. Korney, Postgraduate Student, korney.alena@yandex.ru,
E. N. Kryuchkova, Cand. Ph.-Math. Sc., Associate Professor, kruchkova_elena@mail.ru,
Polzunov Altai State Technical University, Barnaul, 656038, Russian Federation

Text Categorization Based on Condensed Graph

The resonant world events of 2020 led to an increase in the amount of information on the Internet, including criminal, fake news, and fake negative reviews. False negative information can spread very quickly, and methods are needed to suppress this process. The development of effective algorithms for automatic text analysis is especially relevant today. The most important

subtasks include thematic categorization, sentiment analysis, including ABSA (aspect-based sentiment analysis). The paper proposes a combined semantic-statistical algorithm for the aspect analysis of large texts, based on the use of a semantic graph. The aspect extraction method contains the phases of selecting a set of significant words, calculating the weights of the vertices of the semantic graph by the relaxation method, filtering aspects based on the gradient method. The method proposed allows to extract domain-dependent aspect terms from training data. Different aspect term sets extracted from different domains have the same statistical features, and in the same time lexical diversity and structure are taken into account.

Keywords: *semantic graph, text categorization, semantic-statistical algorithm, knowledge extraction, aspect term extraction*

DOI: 10.17587/it.27.138-146

References

1. **Kowsari K., Jafari Meimandi K., Heidarysafa M., Mendu S., Barnes L., Brown D.** Text classification algorithms: A survey, *Information*, 2019, vol. 10, no. 4, pp. 150.

2. **Zhang X., Zhao J., LeCun Y.** Character-level Convolutional Networks for Text Classification, *Proc. of the Neural Information Processing Systems Conf. (NIPS 2015),* Montreal, Canada, 2015, available at: https://arxiv.org/abs/1509.01626 (accessed July 18, 2016).

3. **Ju R.** An Efficient Method for Document Categorization Based on Word2vec and Latent Semantic Analysis, *2015 IEEE Int. Conf. on Computer and Information Technology*, 2015, pp. 2276—2283.

4. **Pontiki M., Galanis D., Pavlopoulos J., Papageorgiou H., Androutsopoulos I., Manandhar S.** SemEval-2014 Task 4: Aspect based sentiment analysis, *The 8th Intern. Workshop on Semantic Evaluation (SemEval 2014*), Dublin, Ireland, 2014, pp. 27—35.

5. **Medhat W., Hassan A., Korashy H.** Sentiment analysis algorithms and applications: a survey, *Ain Shams Eng. Jour.*, 2014, no. 5, pp. 1093—1113.

6. **Xu Y., Jones G. J., Li J., Wang B., Sun C.** A Study on Mutual Information-based Feature Selection for Text Categorization, *Journal of Computational Information Systems*, 2007, no. 3, pp. 1007—1012.

7. **Dai W., Xue G. R., Yang Q., Yu Y.,** Transferring naive bayes classifiers for text classification, *In AAAI*, 2007, vol. 7, pp. 540—545.

8. **Guo G., Wang H., Bell D., Bi Y., Greer K.** Using kNN model for automatic text categorization. Soft Computing, 2006, vol. 10, no. 5, pp. 423—430**.**

9. **Joachims T.** Text categorization with Support Vector Machines: Learning with many relevant features, NÉdellec C., Rouveirol C. (eds) *Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence),* 1998, vol 1398, Springer, Berlin, Heidelberg, available at: https://doi.org/10.1007/BFb0026683.

10. **Luan Y., Lin S.** Research on Text Classification Based on CNN and LSTM, *2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA),* Dalian, China, 2019, pp. 352—355, doi: 10.1109/ICAICA.2019.8873454.

11. **Kureichik V. V., Bova V. V., Leshchanov D. V.** Semantic search model for knowledge management systems based on genetic procedures, *Information Technology*, 2017, vol. 23, no. 12, pp. 876—883.

12. **Krotova A, Krayvanova V., Kryuchova E**. Mathematical model of natural language for fuzzy associative search tasks, *SPECOM 2011 Proceedings*, Kazan, 2011, pp. 402—406 (in Russian).

13. **Savchenko V.** Semantic Search Algorithms in Large Text Collections, *in Supplementary Proceedings of AIST* 2014, pp. 161—166 (in Russian).

14. **Kazakov M., Kruchkova E.** Classication of complex images based on semantic graph, *Journal of Applied informatics*, vol. 6, no. 54, pp. 79—89 (in Russian).

15. **Korney A., Kryuchkova E0,** Short text sentiment analysis based on semantic graph, *ROBOTICS AND ARTIFICIAL INTELLIGENCE X Proceedings*, Zheleznogorsk, 2018, pp. 168—174 (in Russian).

16. **КиноПоиск.** Все фильмы планеты, available at: https://www.kinopoisk.ru/.

17. **Pontiki M.** et al. SemEval-2016 Task 5: Aspect Based Sentiment Analysis // Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 19-30. San Diego, California: The Association for Computational Linguistics, available at: https://doi.org/10.18653/v1/S16-1002.

18. **Heaps Harold Stanley.** Information Retrieval: Computational and Theoretical Aspects. Academic Press, Inc.6277 Sea Harbor Drive Orlando, FL, United States, 1978.

19. **Zipf G. K.** Human behavior and the principle of least effort. Cambridge, Mass., Addison—Wesley, 1949, pp. 573.

20. **Li W**. Random texts exhibit Zipf's-law-like word frequency distribution, *IEEE Transactions on information theory,* 1992, vol. 38, no. 6, pp. 1842—1845.