СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ

УДК 004:912

ОЦЕНКА СЕМАНТИЧЕСКОЙ ЗНАЧИМОСТИ НЕЧЁТКИХ КОЛЛОКАЦИЙ НА ОСНОВЕ ОБОБЩЁННОЙ ВЕКТОРНО-ПРОСТРАНСТВЕННОЙ МОДЕЛИ ТЕКСТОВОЙ КОЛЛЕКЦИИ[1]

Статья поступила в редакцию 14.01.2016, в окончательном варианте 17.02.2016.

Поляков Дмитрий Вадимович, кандидат технических наук, старший преподаватель, Тамбовский государственный технический университет, 392000, Российская Федерация, г. Тамбов, ул. Советская, 106, e-mail: dimadress@yandex.ru

Попов Андрей Иванович, кандидат педагогических наук, доцент, Тамбовский государственный технический университет, 392000, Российская Федерация, г. Тамбов, ул. Советская, 106, e-mail: olimp_popov@mail.ru

Матвеева Алёна Сергеевна, аспирант, Тамбовский государственный технический университет, 392000, Российская Федерация, г. Тамбов, ул. Советская, 106, e-mail: klenchic@mail.ru

Карасёв Павел Игоревич, аспирант, Тамбовский государственный технический университет, 392000, Российская Федерация, г. Тамбов, ул. Советская, 106, e-mail: karasevpav@rambler.ru

Балюков Дмитрий Анатольевич, аспирант, Тамбовский государственный технический университет, 392000, Российская Федерация, г. Тамбов, ул. Советская, 106, e-mail: logan.tambov@gmail.com

Рассмотрены обобщённая векторно-пространственная модель текстовой коллекции; математический аппарат сравнения семантических характеристик произвольной группы факторов, формализованных в виде нечётких множеств и заданных в виде термов. Этот аппарат позволяет определять семантическую значимость выбранной группы факторов в сравнении с термами для кластеризации текстовой коллекции или при решении на ней задач информационного поиска. Описаны постановка вычислительного эксперимента; архитектура программного обеспечения позволяющего провести такие эксперименты. Введено понятие нечёткой коллокации. Проанализированы методы построения нечётких коллокаций на основе лингвистических переменных и фаззификации расстояний между термами. Приведены результаты экспериментальных исследований для факторов, формализованных нечёткими коллокациями. Рассмотрение нечётких коллокаций в работе ограничено двумя методами их построения: на основе лингвистической переменной и с помощью фаззификации расстояния между термами в текстах. Кроме того, исследуются только коллокации, состоящие из двух термов. Сделан **вывод** о независимой природе коллокаций и об эффективности их использования для кластеризации текстовых коллекций.

Ключевые слова: анализ текстов, нечёткая коллокация, факторный анализ, *svd*-разложение, лингвистическая переменная, теория нечётких множеств, архитектура программного обеспечения, векторно-пространственная модель

EVALUTION OF SEMANTIC MEANINGFUL OF FUZZY COLLOCATION BY USING THE GENERALIZED VECTOR-SPACE MODEL OF TEXT COLLECTION

Polyakov Dmitrij V., Ph.D. (Engineering), Tambov State Technical University, 106 Sovetskaya St., Tambov, 392000, Russian Federation, e-mail: dimadress@yandex.ru

Popov Andrej I., Ph.D. (Pedagogical), Tambov State Technical University, 106 Sovetskaya St., Tambov, 392000, Russian Federation, e-mail: olimp_popov@mail.ru

[1] Работа выполнена при финансовой поддержке РФФИ (проект 15-41-03143).

Matveeva Aljona S., post-graduate student, Tambov State Technical University, 106 Sovetskaya St., Tambov, 392000, Russian Federation, e-mail: klenchic@mail.ru

Karasjov Pavel I., post-graduate student, Tambov State Technical University, 106 Sovetskaya St., Tambov, 392000, Russian Federation, e-mail: karasevpav@rambler.ru

Baljukov Dmitrij A., post-graduate student, Tambov State Technical University, 106 Sovetskaya St., Tambov, 392000, Russian Federation, e-mail: logan.tambov@gmail.com

In article are considered the generalized vector-space model of text collection; the mathematical apparatus of the comparison of semantic characteristics of an arbitrary group of factors, that are formalized in the form of fuzzy sets and terms. This mathematical apparatus allows defining the semantic significance for clustering text collection or information retrieval of the chosen groups of factors in comparison with the terms. Staging of the experiment and the architecture of software allows it is described. In article is introduced the concept of fuzzy collocation. The methods of constructing fuzzy collocations based on linguistic variables and fuzzification of distances between terms are offered. The results of experiment for factors that formalized fuzzy collocation are given. Consideration of fuzzy collocations is limited by two methods of constructing them: based on the linguistic variable and using the fuzzification of the distance between terms in texts. In addition, only the collocations, consisting of two terms are studied. Authors proved the independent nature of collocation and the effectiveness of their use for the clustering of text collections.

Keywords: texts analysis, fuzzy collocation, factor analysis, svd-decomposition, linguistic variable, fuzzy set theory, software architecture, vector-space model

Введение. На сегодняшний день становление инновационной экономики в РФ, согласно волновой теории экономического развития Н.Д. Кондратьева [12], невозможно без формирования элементов шестого технологического уклада [5]. Одна из его ключевых характеристик – развитие систем искусственного интеллекта; робототехнических комплексов; глобальной информационной сети [1, 5]. Вместе с тем биотехнологии, нанотехнологии, гибкая автоматизация производства и другие составляющие экономики, приоритетные для шестого технологического уклада, также нуждаются в информационном сопровождении (поддержке) на принципиально новом уровне.

Обеспечение эффективности такого информационного сопровождения непосредственно связано с повышением степени автоматизации при решении задач поиска и анализа информации, большая часть которой представлена в виде текстовых документов на естественных языках. Однако пока имеющиеся результаты исследований (разработок) не позволяют построить взаимно-однозначное соответствие между синтаксисом и семантикой текстовой информации. Даже частичное решение данной задачи даст толчок к развитию таких направлений, как машинный поиск, мониторинг, кластеризация и фильтрация текстовой информации; создание семантических интерфейсов, автоматизированный семантический анализ информации и многое другое. Это, в свою очередь, обеспечит возможности использования глобальной информационной сети Интернет на качественно новом уровне; откроет новые подходы к разработке систем искусственного интеллекта, способных к обучению на основе текстовой информации; позволит создавать робототехнические комплексы, способные к коммуникации с людьми на естественном языке. Безусловно, подобные технологии должны основываться на достижениях в областях микроэлектроники и информатики в рамках текущего пятого технологического уклада. Поэтому именно сегодня особенно велика значимость фундаментальных исследований в области семантического анализа информации, представленной на естественном языке [1, 13].

Одним из наиболее успешных направлений исследования семантической составляющей текстовой информации является выделение различных характеристик документов (факторов) и формализация текста как элемента некоторого семантического пространства. Такой подход позволяет вводить метрики, отражающие семантическую близость документов; отображать на пространство факторов информационные интересы пользователя. В свою очередь, это позволяет решать задачи информационного поиска в больших массивах информации; исследовать взаимосвязи синтаксиса и морфологии с семантикой в текстах на естественном языке. Комплексное исследование таких подходов и является целью настоящей статьи.

Алгоритм оценки семантической значимости факторов на основе обобщённой векторно-пространственной модели. Классическим примером указанных выше подходов является векторно-пространственная модель (ВПМ) текстовой коллекции (ТК) [13], в которой в качестве факторов используются термы.

Сформулируем строго задачу представления текстовых документов в пространстве факторов. Пусть D – множество текстовых документов или ТК. Причём $D = \{d_1, d_2, \ldots, d_N\}$, $|D| = N$, а $|.|$ – операция взятия мощности множества. Формализация ТК в рамках любой известной модели приводит к частичной потере семантической составляющей. Будем рассматривать текстовый документ как совокупность характеристических объектов (ХО).

Под таким объектом понимается любая характеристика части документа, которая потенциально может быть связана с семантической составляющей. Так известными и часто используемыми ХО для текстов являются термы. Именно представление документа в виде совокупности термов легло в основу построения расширенной булевой, векторно-пространственной и вероятностной моделей поиска [13].

Рассмотрим представление документа в виде совокупности ХО. Пусть U_p – универсум всех ХО, на наличие которых будет исследован документ в ходе формализации ТК. Тогда произвольный документ $d_i, i = \overline{1,N}$ удобно представить в виде:

$$d_i = \{p \in U_p \mid p \in d_i\}, \tag{1}$$

где $p \in d_i$ – означает присутствие ХО p в документе d_i. Обозначим для определённости $|d_i| = M_i$. Рассмотрим множество $U_F = \{\widetilde{F}_1, \widetilde{F}_2, \ldots, \widetilde{F}_K\}$, $|U_F = K|$, где U_F представляет собой пространство факторов, отражающих семантику текстовых документов. Каждый элемент $U_F = \{\widetilde{F}_i\}_{i=\overline{1,K}}$ – нечёткое подмножество U_p, задаваемое некоторой функцией принадлежности $\mu_i : U_p \to [0,\ 1]$.

Требование нечёткости множеств $\{\widetilde{F}_i\}_{i=\overline{1,K}}$ выдвинуто с целью обобщить предлагаемую модель. В частном случае любое из фактор-множеств может быть классическим. Тогда μ_i сводится к характеристической функции множества с областью допустимых значений {0, 1}.

Кроме термов в роли ХО могут выступать коллокации – коллективные локации термов с не соответствующими закону их случайного распределения частотами появления [15]. Данный подход исследован в работах Е.В. Недошивиной [15], М.В. Киселёва [11], Р.К. Бишта [28] и Л.М. Пивоваровой [17, 26]. Также в качестве U_p можно выбрать нечёткие коллокации, исследованные в работах Ю.Ю. Громова [6], Д.В. Полякова [18, 19] и О.Г. Ивановой [19].

Отметим, что ХО в общем случае не обязаны иметь численную природу. Действительно, факторы по способу задания представляют собой нечёткие подмножества наборов ХО произвольной природы. Поэтому в качестве элементов U_p можно взять, например, онтологии, которые представляют собой спецификации концептуализации предметных областей [30]. Исследование данных ХО получило развитие в работах А.Е. Ермакова [7–10], В.В. Плешко [7, 8, 10] и Г.В. Липинского [7].

Расширение понятия ХО позволяет исследовать семантические характеристики документа, обычно теряемые при формализации. К примеру, для построения многих моделей ТК используется лемманизация. Данная процедура представляет собой приведение слов, встречающихся в тексте, к единым словарным формам – термам. Например, существительные приводятся к именительному падежу, единственному числу, а при работе с прилагательными еще добавляется задача приведения их к мужскому роду. Вместе с тем в качестве ХО допустимо использовать производные моделей формообразования и определения форм слов естественных языков для выявления значимых связанных структур в тексте. Данные модели, а также алгоритмы словообразования и морфологического анализа исследованы в работах Г.Г. Белоногова [3], Дж. Голдсмита [30], А.В. Пруцкова [23, 24].

Итак, пусть документы ТК D формализованы согласно (1). Вместе с тем поставленная задача требует представления каждого документа в виде вектора в пространстве факторов. То есть по результатам построения модели необходимо получить для каждого $d_i \in D, i = \overline{1,N}$ представление в виде:

$$d_i\left(f_1^i, f_2^i, \ldots, f_K^i\right), \tag{2}$$

где $\left\|f_j^i\right\|_{i=\overline{1,N}, j=\overline{1,K}}$ – соответствие d_i фактору $\widetilde{F}_j$.

В работе «Обобщение векторно-пространственной модели для оценки семантической значимости характеристик текстовых документов» [22] авторским коллективом данной статьи показано, что при разных способах нахождения $\left\|f_j^i\right\|_{i=\overline{1,N},j=\overline{1,K}}$ обобщённая ВПМ (ОВПМ) ТК сводится к классическим, хорошо исследованным методам представления текстовой информации для решения задач поиска и кластеризации. Кроме того, в [22] предложена и обоснована логико-лингвистическая модель ТК, согласно которой нахождение $\left\|f_j^i\right\|_{i=\overline{1,N},j=\overline{1,K}}$ осуществляется по формуле:

$$f_j^i = T\left(\xi\left(\frac{1}{M_j}\sum_{k=1}^{M_j}\mu_i\left(p_k^j\right)\right),\zeta\left(N\cdot\left(1+\sum_{k=1}^{N}S\left(\mu_j\left(p_1^k\right),\mu_j\left(p_2^k\right),\ldots,\mu_j\left(p_{M_1}^k\right)\right)\right)^{-1}\right)\right), \quad (3)$$

где $\xi(\cdot)$ и $\zeta(\cdot)$ – некоторые функции, для которых верно, что ξ : [0, 1] $\rightarrow$ [0, 1], $\xi(0)=0$, $\xi(1)=1$; ζ : [1, |D|] $\rightarrow$ [0, 1], $\zeta(1)=0$, $\zeta(|D|)=1$ и $\zeta\uparrow$ на [1, |D|]; $\left\{p_k^j\right\}_{k=\overline{1,M_j}}$ – k-ый член d_i; $T(\cdot,\cdot)$ – T-норма, а $S(\ldots)$ – S-норма, обобщённая до функции с переменным числом аргументов, на основе рекуррентного соотношения:

$$S\left(x_1,x_2,\ldots,x_m\right)=\begin{cases}S\left(x_1,x_2\right),m=2,\\ S\left(x_m,S\left(x_1,x_2,\ldots,x_{m-1}\right)\right),m>2;\end{cases},\forall m\geq 2,x_i\in[0,1],i=\overline{1,m},$$

где $S(\cdot,\cdot)$ – классический вид S-нормы как функции двух переменных.

Если в качестве U_p взять τ = {t_1, t_2, …, t_n}, $|\tau|=n$, где τ – множество термов документов D; в качестве $\left\{\widetilde{F}_i=\hat{T}_i\right\}_{i=\overline{1,K}}$, где $\hat{T}_i$ – совокупность словоформ терма $t_i\in\tau$; в качестве функций $\xi(\cdot)$ и $\zeta(\cdot)$ взять $\xi(x)=x$ и $\zeta(x)=\log(x),\forall x\in R$; а в качестве операций нечёткой логики: $T(x,y)=xy$ и $S(x,y)=x+y-xy$, $\forall x,y\in R$, то ОВПМ сведётся к классической ВПМ ТК [22, 34].

Рассмотрение ВПМ с позиций формул (1–3) приводит к выводу о том, что основная потеря семантической информации происходит на этапе преобразования документа к виду (1). Действительно, зная частоты термов и их количество в документе, легко восстановить его представление в виде (1). Вместе с тем, информация о порядке термов, их словоформах, связях, определяющих семантику документов, утрачивается ещё на этапе построения (1).

Таким образом, для того чтобы повысить эффективность ВПМ ТК, необходимо добавить к множеству U_p ХО, отражающие семантику, утерянную при формализации документов. Пусть P – множество добавленных ХО. Кроме того, введём в рассмотрение факторы F, соответствующие природе добавленных ХО – то есть, такие нечёткие множества, носителем которых является P. Построим множества $U_p=\tau\cup P$, а $U_F=\Theta\cup F$, где $\Theta=\left\{\hat{T}_1,\hat{T}_2,\ldots,\hat{T}_n\right\}$.

Вместе с тем при конструировании множеств F и P необходимо убедиться, что элементы P не просто отражают семантические свойства документа, а формализуют именно то смысловое наполнение, которое не может быть отражено посредством термов. Для проверки данного свойства, удобно использовать *svd*-разложение [24, 25] матрицы $\left\|f_j^i\right\|_{K\times N}$, каждый элемент которой вычислен на основе (3). Результатом *svd*-разложения будет вектор сингулярных чисел $\Delta\left(\delta_1,\delta_2,\ldots,\delta_K\right)$, координаты которого $\left(\delta_i,i=\overline{1,K}\right)$ являются оценкой семантической значимости соответствующего фактора ($\left\|f_j^i\right\|_{i=\overline{1,N},j=\overline{1,K}}$) в сравнении с остальными [22].

Для задания алгоритма оценки значимости факторов, введём в рассмотрение некоторые обозначения. Пусть $U_{K\times N}$ – универсальное множество матриц вещественных ($\boldsymbol{Z}$) чисел с K строками и N столбцами, причём $K\geq N$. Рассмотрим отображение *SVD*: $U_{K\times N}\rightarrow\boldsymbol{Z}^K$. Пусть *SVD* формализует работу алгоритма *svd*-разложения [32, 33].

Рассмотрим алгоритм *Assessment*, позволяющий оценить значимость произвольного множества факторов F для текстовой коллекции D на основе характеристических объектов P.

Шаг 1. Осуществить ввод F, P, D.

Шаг 2. $U_p = \tau \cup P$ и $U_F = \Theta \cup F$.

Шаг 3. Построить $\left\| f_j^i \right\|_{K \times N}$, вычислив для каждого $i = \overline{1,N}, j = \overline{1,K}$:

$$f_j^i = \frac{1}{M_j} \sum_{k=1}^{M_j} \mu_i\left(p_k^j\right) \log\left(N\left(1 + \sum_{k=1}^{N} S\left(\mu_j\left(p_1^k\right), \mu_j\left(p_2^k\right), \ldots, \mu_j\left(p_{M_i}^k\right)\right)\right)^{-1}\right).$$

Шаг 4. $\Delta\left(\delta_1, \delta_2, \ldots, \delta_K\right) = SVD\left(\left\| f_j^i \right\|_{K \times N}\right)$.

Шаг 5. Создать массив пар $\left\langle \delta_i, \widetilde{F}_i \right\rangle, i = \overline{1,K}$ и отсортировать его по убыванию, сравнивая элементы по первому члену пары.

Шаг 6. Для всех k от 1 до K найти

$$\nu_k = \left(\sum_{i=\overline{1,k}}^{\widetilde{F}_i \in F} \delta_i \right) \cdot \left(\sum_{i=1}^{k} \delta_i \right)^{-1} \text{ и } \vartheta_k = \left| \left\{ \widetilde{F}_i \mid i = \overline{1,k} \right\} \cap F \right| / k,$$

где ν_k – семантическая значимость факторов F в выборке из k наиболее значимых, а ϑ_k – нормированное значение числа факторов из F [23].

Оценка семантической значимости факторов F для их носителя P посредством алгоритма *Assessment* позволяет сделать вывод о целесообразности использования тех или иных факторов для решения задачи кластеризации. Специфический вид формулы для вычисления f_j^i на шаге 3 обусловлен необходимостью вычисления элементов матрицы, соответствующих термам в рамках ВПМ. Ниже мы рассмотрим использование алгоритма *Assessment* для оценки семантической значимости факторов конкретного вида.

Понятие нечёткой коллокации и способы её формализации. Кортеж термов

$$\left\langle t_i, t_j \right\rangle,\ t_i, t_j \in \tau,\ i, j = \overline{1,n} \tag{4}$$

назовём коллокацией. Он (кортеж) задаёт термы, составляющие коллокацию и их порядок. Обычно [11, 15, 17, 26, 28] под коллокацией понимается группа термов, расположенных в тексте непосредственно рядом друг с другом. Вместе с тем, есть серьёзные основания полагать, что семантическая составляющая может проявляться и в последовательностях термов, находящихся на определённом расстоянии друг от друга. Так, например, наиболее распространённые информационно-поисковые системы, такие как Гугл или Яндекс, а также некоторые Российские юридические информационно-справочные системы, работающие с большими ТК, позволяют формулировать запросы на специальном скриптовом языке. Он обладает синтаксисом поиска с параметрами, позволяющими задавать расстояния между термами [16, 28]. Кроме того, есть работы [6, 18, 19] посвящённые исследованию коллокаций с термами, отстоящими друг от друга в тексте на некотором расстоянии.

За расстояние между парой произвольных термов t_i и t_j $i, j = \overline{1,n}$ здесь и далее примем число слов стоящих между этими термами в документе. Важно отметить, что пара термов, несущая некоторую семантическую нагрузку, совершенно необязательно всегда находится на одном и том же расстоянии друг от друга. Вместе с тем задача отнесения встретившейся в текстовом документа пары термов на определённом расстоянии друг от друга к некоторой коллокации является нетривиальной и решается в условиях неопределённости. Такое появление термов назовём ХО коллокации. Представим данный ХО в виде кортежа:

$$\left\langle t_i, t_j, k \right\rangle,\ t_i, t_j \in \tau,\ i, j = \overline{1,n}, \tag{5}$$

где k – некоторое натуральное число, задающее расстояние между термами.

Введём понятие нечёткой коллокации как нечёткого подмножества универсума ХО коллокации. Тогда нечёткую коллокацию удобно представить в виде кортежа:

$$\left\langle t_i, t_j, \mu_i \right\rangle,\ t_i, t_j \in \tau,\ i, j = \overline{1,n}, \tag{6}$$

где μ_i – функция принадлежности, ставящая в соответствие каждому целому числу k, задающему расстояние между термами в ХО коллокации, их степень принадлежности к соответствующей нечёткой коллокации.

Отметим, что элементы вида (5) представляют собой некоторый частный случай ХО текстового документа, а элементы вида (6) – соответствующие им факторы. В дальнейшем именно эти элементы будут исследоваться в рамках предложенного ранее алгоритма оценки семантической значимости факторов на основе ОВПМ. Поэтому удобно принять множество ХО в виде (5) как P, а множество факторов (6) как F. Важно отметить, что в общем случае коллокации не ограничиваются двумя термами [26]. Формализация таких коллокаций на основе теории нечётких множеств возможна с помощью многомерных функций принадлежности. Однако в данной работе мы ограничимся рассмотрением только коллокаций вида (4).

Для использования алгоритма *Assessment* с целью оценки семантической значимости нечётких коллокаций для текстов некоторой ТК D, необходимо построить множества P и F.

Множество ХО тривиально. Теоретически оно задаётся как $P = \tau \times \tau \times \mathrm{N}$, где $\times$ – декартово произведение множеств, а N – множество натуральных чисел. На практике работать с множеством с бесконечной мощностью не всегда удобно. Поэтому можно ограничиться рассмотрением лишь ХО коллокаций, которые встречаются в D. На первый взгляд, это также очень большое число, равное всем парам термов всех документов. Однако на практике достаточно рассматривать лишь ХО, входящие в носитель добавленных факторов F. Так как пары термов находящихся в одном документе на больш*и*х расстояниях не несут семантической нагрузки, то носители факторов F ограничены некоторой константой. Это означает, что P можно построить за один проход по документам коллекции.

Нетривиальным является вопрос построения F. С одной стороны, нельзя допустить, чтобы F стал больше, чем количество документов в ТК, поскольку тогда ранг матрицы $\left\| f_j^i \right\|_{K \times N}$ будет равен мощности D, что исключит ряд факторов из рассмотрения при *svd*-разложении. С другой стороны, только коллокаций в документе d_i оказывается около $C_{M_i}^2$, а число возможных нечётких коллокаций мультипликативно возрастает на величину всех возможных функций принадлежности $\mu_i : \mathrm{N} \to [0,\ 1]$. Для решения данной проблемы были предложены методы формализации нечётких коллокаций [28, 29].

В работе «Метод формализации нечётких коллокаций термов в текстах на основе лингвистических переменных» [21] рассматривается лингвистическая переменная *distance*.

$$distance = <d,\ Term,\ G,\ M>,$$

где d = «дистанция между термами в коллокации» – имя лингвистической переменной *distance; Term* = {«маленькая», «средняя», «большая»} – терм-множество значений лингвистической переменной *distance*; G – синтаксическое правило, порождающее значения *distance*, которое представляет собой метод лингвистического конструирования новых значений на основе связок и модификаторов. Множество связок – Op*{*«и», «или»}, а модификаторов Mod {«не», «очень»}. Пусть $op \in Op$, а te_1 и $te_2 \in Term$. Тогда G на основе данных элементов будет иметь вид $te_1\ op\ te_2$. Например, пусть te_1 = «большая», te_2 = «средняя», а op = «или». Тогда $te_1\ op\ te_2$ = «дистанция между термами большая или средняя».

Рассмотрим произвольный элемент $m \in Mod$. Семантическое правило, для произвольного терма $te \in Term$ имеет вид: $m\ te$. Например, при te = «большая», а m = «не», $m\ te$ означает «не большая». В рамках конструирования новых значений лингвистической переменной допускается последовательное применение различных связок и модификаторов.

Множество M представляет собой семантическое правило, которое ставит в соответствие каждому сконструированному посредством G значению нечёткой переменной некоторую функцию принадлежности $\mu: Z_+ \to [0, 1]$. Она характеризует смысловое наполнение этого значения. Эта функция отображает каждое конкретное расстояние между двумя термами, составляющими коллокацию, на отрезок [0, 1], определяя, таким образом, степень принадлежности найденной пары термов к соответствующей коллокации.

Множество $M = \{\tilde{\mu}_м, \tilde{\mu}_c, \tilde{\mu}_б, T, S, otr, power\}$, причём таким значениям лингвистической переменной *distance* как «маленькая», «средняя» и «большая» соответствуют функции $\tilde{\mu}_м$, $\tilde{\mu}_c$ и $\tilde{\mu}_б$; T,S-нормы соответствуют логическим связкам «и» и «или», операция нечёткого отрицания *otr* – модифика-

тору «не», а $power$ – «очень». В [21] предложены и обоснованы следующие выражения для отображений множества M:

$$\begin{cases} \widetilde{\mu}_{м}(x) = \max\{0, \min\{1, (R_l - x)/(R_l - L_l)\}\}, \\ \widetilde{\mu}_{c}(x) = otr(S(\widetilde{\mu}_{м}(x), \widetilde{\mu}_{c}(x))), \\ \widetilde{\mu}_{б}(x) = \max\{0, \min\{1, (x - L_r)/(R_r - L_r)\}\}, \\ power(x) = x^{\alpha}; \end{cases} \quad (7)$$

где L_l, R_l, и L_r, R_r $(L_l,\ R_l,\ L_r,\ R_r \in Z_+)$ левые и правые границы функций принадлежности по [21], т.е. точки, в которых функция достигает «0» или «1».

В [21] T,S-нормы и операцию отрицания предлагается выбирать на основе вычислительных экспериментов. Вместе с тем для корректного применения алгоритма *Assessment*, а именно чтобы (3) для термов сводились к коэффициентам *tf-idf,* необходимо положить $T(x, y) = xy$, $S(x, y) = x + y - xy$, а $otr(x) = 1 - x$ для $\forall x, y \in R$. Эти допущения позволяют существенно сократить неопределённость и использовать коллокации, формализованные посредством значений лингвистических переменных, в качестве факторов для алгоритма оценки *Assessment*. Существенным плюсом такого подхода является то, что каждая коллокация соответствует некоторому значению лингвистической переменной. Это может помочь, например, при составлении аннотаций на естественном языке к кластерам; при решении задач поиска и кластеризации текстовой информации с использованием нечётких коллокаций.

Минусами предлагаемого в [21] подхода являются его некоторая робастность (ограниченность допустимых функций принадлежности) и избыточность (большое количество коллокаций, не имеющих семантической значимости, но требующих машинного времени для обработки).

В качестве альтернативного подхода, лишённого данных минусов, был предложен метод формализации нечётких коллокаций на основе фаззификации расстояний между термами в текстах [19]. В основу этого метода легла идея фаззификации ХО коллокаций и их объединение в ограниченный набор факторов (нечётких коллокаций) посредством T,S-норм, формализующих в теории нечётких множеств логические связки «и» и «или». В [20] был предложен и обоснован подход к фаззификации коллокаций с произвольным числом термов. Частным случаем этого подхода (при числе термов в коллокации равном «2») стала формула:

$$\mu\ (x) = \begin{cases} \dfrac{\alpha}{\beta} x - \alpha \dfrac{k - \beta}{\beta}, x \in [k-1, k], \\ -\dfrac{\alpha}{\beta} x + \alpha \dfrac{k + \beta}{\beta}, x \in [k, k+1], \\ 0, x \in (-\infty, k-1) \cup (k+1, +\infty); \end{cases} \quad (8)$$

где k – расстояние между термами в фаззифицированном ХО коллокации, а α и β – константы, такие что $\mu\ (k) = \alpha$, $0 < \alpha \le 1$, $\mu\ (x) = 0 \Leftrightarrow |x - k| \ge \beta$.

Объединение фаззифицированных ХО коллокаций в рамках одного текстового документа осуществляется посредством S-нормы, так как она формализует в теории нечётких множеств логическую операцию «или». Полученная, таким образом, нечёткая коллокация задаёт характеристику конкретного документа. Для характеристики произвольной группы документов $D^* \subset D$ с помощью коллокации h достаточно взять T-норму, формализующую в теории нечётких множеств логическую операцию «и» от всех функций принадлежности, задающих данную коллокацию h в D^*.

Для того чтобы выявить нечёткие коллокации, отражающие семантику документов и позволяющие осуществить кластеризацию исходной коллекции D, рассмотрим некоторое подмножество текстовых документов $D^* \subset D$ и подмножество $D^{\setminus} = D \setminus D^*$. В [20] введено понятие небулева разбиения D на D^* и $D^{\setminus}$. Оно предполагает, что разбиение D на множества D^* и $D^{\setminus}$ невозможно осуществить с помощью булевой модели информационного поиска [14]. То есть нет ни одного терма $t \in \tau$, такого, что он присутствует в D^* и отсутствует в $D^{\setminus}$ или, наоборот, присутствует в $D^{\setminus}$ и отсутствует в D^*. Пусть U_D – множество небу-

левых разбиений D. Тогда показано, что для некоторой нечёткой коллокации h вида (6) функция принадлежности будет иметь вид [20]:

$$\mu_D^h = \underset{\langle D^*, D^\backslash \rangle \in U_D}{S} \left(T\left(\mu_h^D, T\left(S\left(\mu_h^{D^*}, \mu_h^{D^\backslash}\right), otr\left(T\left(\mu_h^{D^*}, \mu_h^{D^\backslash}\right)\right)\right)\right)\right), \quad (9)$$

где $\left\langle D_i^*, D_i^\backslash \right\rangle \in U_D, i = \overline{1, L}, L = \left|U_D\right|$.

Отметим, что формула (9) позволяет снизить размерность пространства факторов F, так для большинства из них $\mu_D^h \equiv 0$. Последнее, в свою очередь, означает нецелесообразность использования h в F [20].

Элементы архитектуры программного обеспечения для постановки вычислительных экспериментов. Для постановки вычислительных экспериментов было разработано программное обеспечение (ПО), отвечающее ряду требований. Основными из них являются гибкость и масштабируемость. Под гибкостью будем понимать свойство архитектуры ПО, отражающее простоту внесения изменений. Масштабируемость же означает, что архитектура ПО допускает расширение функционала без необходимости внесения изменений в какие-либо ранее разработанные модули. Для создания программных компонентов был выбран язык программирования C++. Этот выбор во многом продиктован высокой вычислительной сложностью *svd*-разложения, а также необходимостью работать с большими объёмами данных в виртуальной памяти. Язык программирования C++ является мощным языком, с уникальными возможностями метапрограммирования шаблонов и работы как на высоком (объектно-ориентированном), так и на низком уровнях [2] – последнее важно для реализации наиболее ресурсоёмких операций. Для проектирования архитектуры ПО был выбран язык объектно-ориентированного моделирования *UML* [14, 25]. Он инвариантен относительно языка программирования, выбранного для реализации ПО, но вместе с тем позволяет отобразить архитектуру ПО, а также наглядно продемонстрировать архитектурные решения.

Важнейший, с точки зрения разработчиков ПО, элемент *UML* – диаграмма классов. Именно она определяет архитектуру ПО, с учётом масштабируемости и гибкости [4]. На *UML* диаграмме можно отразить большинство использованных паттернов проектирования (шаблонных архитектурных решений), позволяющих повысить гибкость и масштабируемость ПО [4]. В данной работе мы ограничимся рассмотрением диаграммы классов. Более того, опустим ту часть ПО, которая предназначена для решения инженерных задач, таких, например, как взаимодействие с файлами ТК, лемманизации, *svd*-разложения. Рассмотрим подробно архитектурные решения для реализации в виде модулей ПО алгоритмов оценки семантической значимости факторов на основе ОВПМ.

Разработанное для проведения вычислительных экспериментов ПО представлено большим количеством классов, их шаблонов и связей различных типов между ними. Поэтому будем рассматривать диаграмму классов по частям. На рисунке 1 изображены ключевые классы и связи между ними.

Класс *Collection* задаёт интерфейс объектов, формализующих коллекции и определяет некоторые действия на основе шаблонных методов [4]. Так метод *execute_query*() является чисто виртуальным и должен быть реализован в наследниках *Collection*. Данный метод осуществляет лемманизацию и загрузку документов в виртуальную память.

На рисунке 2 представлено архитектурное решение для формализации факторов.

Интерфейс работы с фактором определён в классе *factor*. Вместе с тем при реализации классов, задающих конкретные типы факторов (терм и коллокацию) была использована идиома программирования на C++ *CRTP* [2, 30]. Она позволила в общем виде в шаблоне класса *CRTP factor methods* реализовать паттерн «Прототип» [4] (метод *copy*()) для всех его наследников и задать метод *is kind of class*(), являющийся приёмником паттерна «Визитёр» [2, 4]. Последний реализован посредствам интерфейса *selector* и шаблона класса *concret selector*. Данный вариант реализации визитёра позволяет определить принадлежность экземпляра класса к конкретному потомку *factor* при работе с общим интерфейсом, то есть проверить во время выполнения программы является ли данный фактор термом или коллокацией.

Для реализации коллокации был разработан класс *function*, задающий многочлен, формализующий функцию принадлежности коллокации. Для данного класса были перегружены операторы суммы и произведения: (*operator* *) и (*operator* +), что позволило использовать соответствующие спецификации шаблона *probabilistic_norm* для работы с функциями принадлежности.

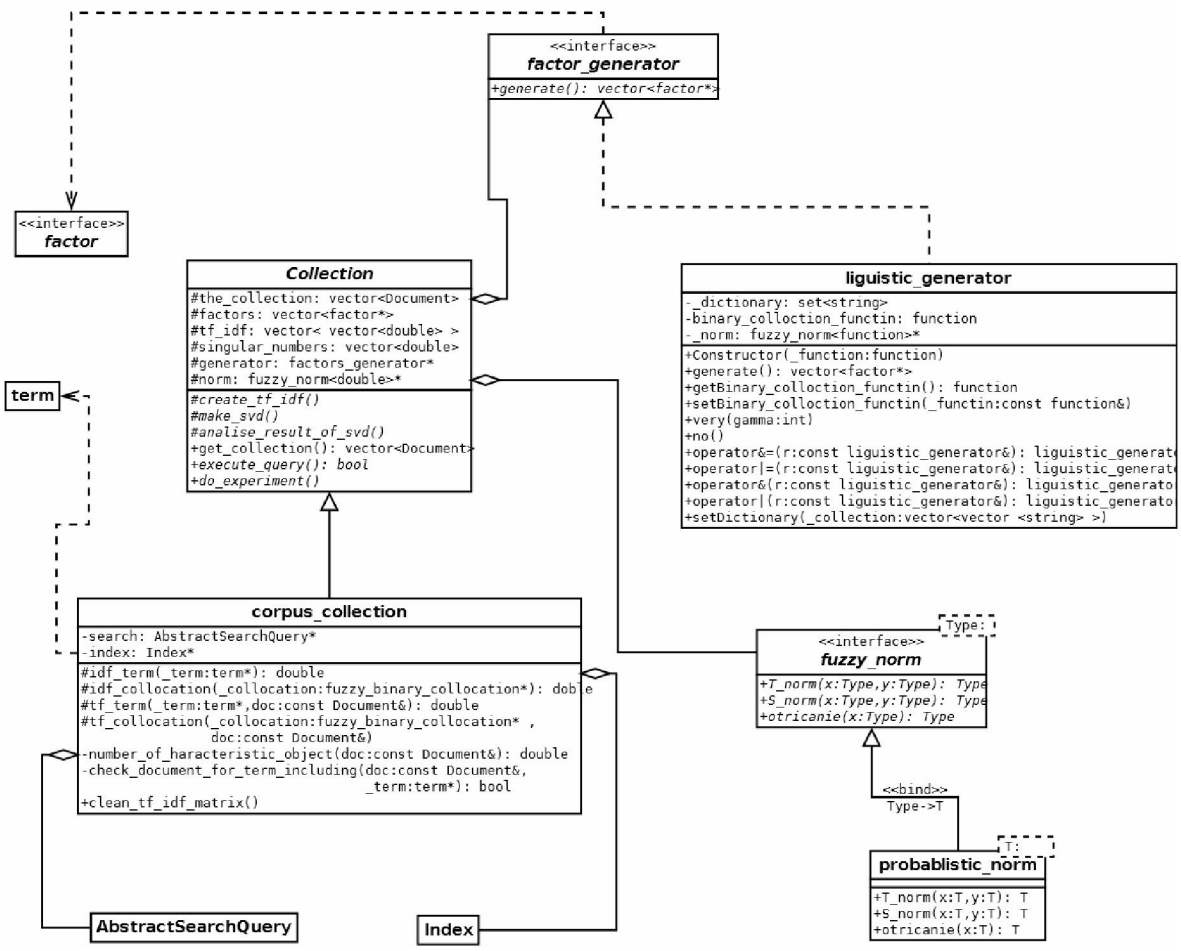

Рис. 1. Класс *Collection* и его ключевые связи в разработанном приложении

Постановка и результаты вычислительных экспериментов. Для постановки вычислительных экспериментов наряду с *liguistic generator* был спроектирован и реализован класс *fuzzification generator*, генерирующий массив коллокаций на основе модели (8)–(9). В качестве ТК D была взята подборка статей журнала «Радио» (Издательство журнала «Радио») с 1949 по 1994 г. Общее количество текстов в ТК – 453; суммарное число слов текстах – 6176763, которые в процессе лемманизации были приведены к 13012 леммам.

При генерации коллокаций экземпляры класса *liguistic_generator* параметризировались значениями левых границ L_l и L_r в интервале от 1 до 9, и значениями правых границ в диапазонах $L_l < R_l < 10$ и $L_r < R_r < 10$.

В ходе экспериментов для каждого, полученного таким образом набора значений L_l, R_l, и L_r, R_r на основе (7), создавались соответствующие коллокации на основе значений лингвистических переменных. При этом не использовались связки и модификаторы, а рассматривались только значения множества *Term*. Полученные в результате вычислительных экспериментов семантически значимые коллокации, нормированные значения сингулярных чисел, которых больше единицы, представлены в таблице 1.

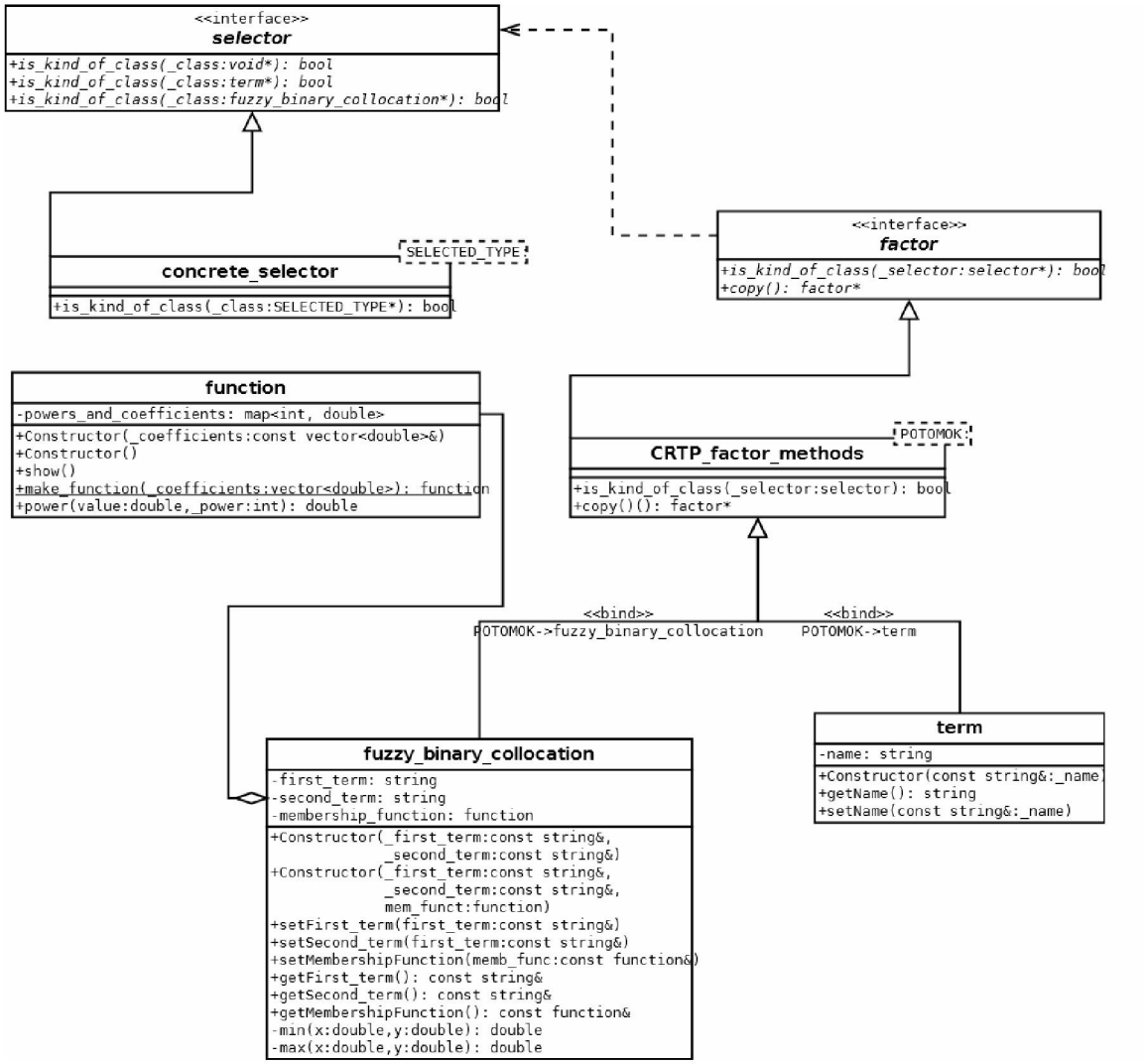

Рис. 2. Диаграмма классов, отражающая архитектурное решение для задания факторов

Таблица 1

Наиболее семантически значимые коллокации

Терм 1	Терм 2	Функция	Значение сингулярного числа $\left(\cdot 10^{-6}\right)$
программа	символ	$-0{,}167x+1{,}667$	2,602
программа	символический	$-0{,}25x+2{,}25$	2,350
работа	датчик	$-0{,}25x+2{,}25$	2,319
программа	система	$-0{,}25x+2{,}25$	2,113
программа	системный	$-0{,}143x+1{,}29$	2,062
программа	ситуация	$-0{,}333x+2{,}667$	1,968
работа	комиссия	$-0{,}25x+2{,}25$	1,959
работа	командир	$-0{,}143x+1{,}286$	1,857
система	интерфейс	$-0{,}2x+2$	1,468

транзистор	испытывать	$-0{,}167x+1{,}667$	1,455
транзистор	использовать	$-0{,}143x+1{,}429$	1,434
система	информационный	$-0{,}143x+1{,}429$	1,290
связь	документ	$-0{,}125x+1{,}25$	1,263
связь	должный	$-0{,}2x+2$	1,186
связь	донесение	$-0{,}25x+2{,}25$	1,147
связь	комплекс	$-0{,}2x+2$	1,110
связь	коммутация	$-0{,}25x+1{,}75$	1,090

Отметим, следующее. А. В таблице 1 в столбце «Функция» представлены не сами функции принадлежности, а выражения, заключённые в аргументах максиминов (7). Б. Коэффициенты при x во всех функциях таблицы 1 отрицательные. Это означает, что наибольшей семантической значимостью обладают коллокации, формализованные значением лингвистической переменной *distance* = «маленький».

По результатам вычислительных экспериментов коллокации, сгенерированные посредством экземпляров класса *fuzzification_generator*, не показали хороших результатов. Более того, для большинства из них функция принадлежности оказалась тождественно равной или близкой к «0». Подобную ситуацию можно объяснить строгостью выбранной вероятностной T-нормы, которая неоднократно используется для построения μ_D^h на основе (9). Вместе с тем выбор другой T-нормы не позволит сравнивать данные коллокации с термами, так как для последних не будет выполняться тождество коэффициентов (3) и элементов матрицы *tf-idf*. На рисунке 3 представлен общий вид функции, соответствующей наиболее семантически значимым коллокациям по результатам проведённых вычислительных экспериментов.

В таблице 2 представлен список термов с наибольшим значением сингулярных чисел. Эти числа были получены путём *svd*-разложения матрицы частот, в которой термы присутствовали наряду с коллокациями. Таким образом, данные числа позволяют сравнить семантическую значимость термов и коллокаций в исследуемой ТК и проверить связаны ли наиболее семантически значимые термы и коллокации.

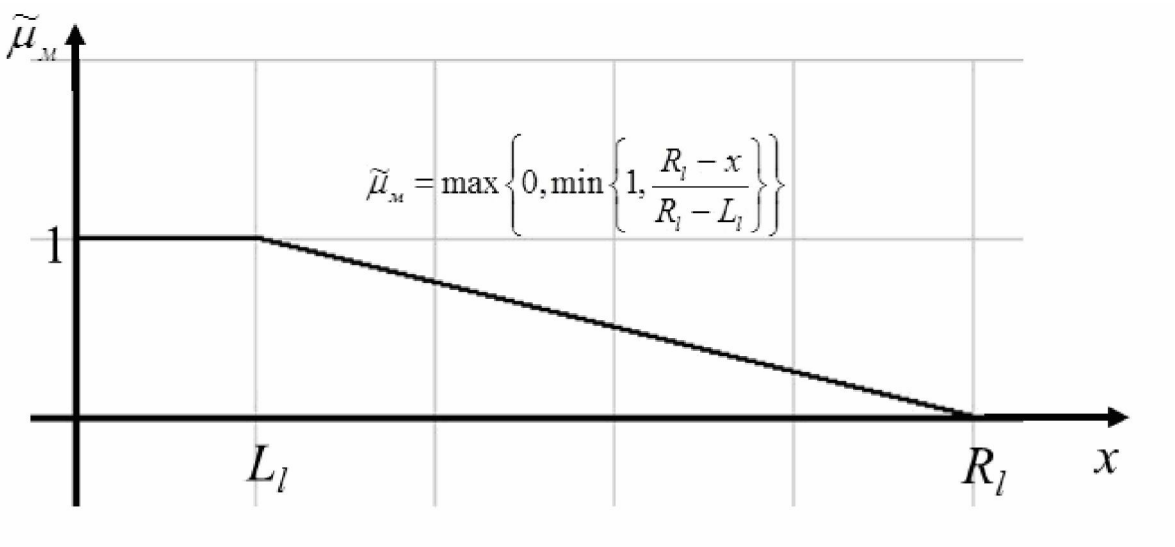

Рис. 3. График функции $\tilde{\mu}_м$, формализующей наиболее значимые коллокации в исследуемой текстовой коллекции

Таблица 2

Наиболее семантически значимые термы

Терм	Вычисленное значение сингулярного числа $\left(\cdot 10^{-4}\right)$
автоматизировать	6,238
аппаратура	3,393
база	2,327
блок	1,835
бюрократ	1,333
военный	1,252

При этом термы с наибольшими значениями сингулярных чисел не входят в состав наиболее значимых нечётких коллокаций – это свидетельствует о независимой природе последних.

В рамках вычислительных экспериментов была проведена оценка семантической значимости коллокаций (ν_k) в выборке из k максимальных элементов. В таблице 3 представлены 30 наибольших значений ν_k.

Таблица 3

Оценка семантической значимости (ν_k) коллокаций среди k первых факторов

k	$\nu_k\left(\cdot 10^{-2}\right)$	L_l	R_l	k	$\nu_k\left(\cdot 10^{-2}\right)$	L_l	R_l	k	$\nu_k\left(\cdot 10^{-2}\right)$	L_l	R_l
152	0,727	5	10	148	0,718	5	10	152	0,701	2	9
151	0,726	5	10	149	0,717	4	10	151	0,700	2	9
150	0,725	5	10	149	0,716	5	9	146	0,700	5	9
149	0,723	5	10	147	0,713	5	10	145	0,699	5	10
152	0,722	4	10	148	0,712	4	10	150	0,699	2	9
151	0,721	4	10	148	0,711	5	9	149	0,696	2	9
152	0,720	5	9	147	0,708	4	10	145	0,694	4	10
150	0,720	4	10	146	0,706	5	10	152	0,693	1	10
151	0,720	5	9	147	0,706	5	9	151	0,692	1	10
150	0,718	5	9	146	0,700	4	10	152	0,692	2	10

Анализируя таблицы 1 и 3, можно прийти к выводу, что коллокации, обладающие максимальной семантической значимостью, формализованы различными функциями принадлежности. Это, в свою очередь, свидетельствует в пользу актуальности задачи поиска семантически более значимых нечётких коллокаций путём вариации формализующих их функций принадлежности. Также в ходе вычислительных экспериментов были получены ϑ_k – нормированные значения количества нечётких коллокаций в выборке из k максимальных факторов. Данные коллокации и соответствующие им значения параметров k, L_l и R_l представлены в таблице 4.

Таблица 4

Доля коллокаций (ϑ_k) среди k наиболее значимых факторов

K	ϑ_k	L_l	R_l	k	ϑ_k	L_l	R_l
152	0,191	1	7	144	0,146	1	6
150	0,180	1	7	143	0,140	1	6
149	0,174	1	7	142	0,134	1	6
148	0,169	1	7	141	0,128	1	6
147	0,163	1	6	140	0,121	1	6
146	0,158	1	6	139	0,115	1	6
145	0,152	1	6	138	0,109	1	6
152	0,191	1	7	136	0,103	1	7

Максимальное отношение числа семантически значимых нечётких коллокаций (ϑ_k) к термам для уравнений функций принадлежности вида $-0,1(6)x+1,(6)$ и $-0,2x+1,2$ представлено на рисунке 4.

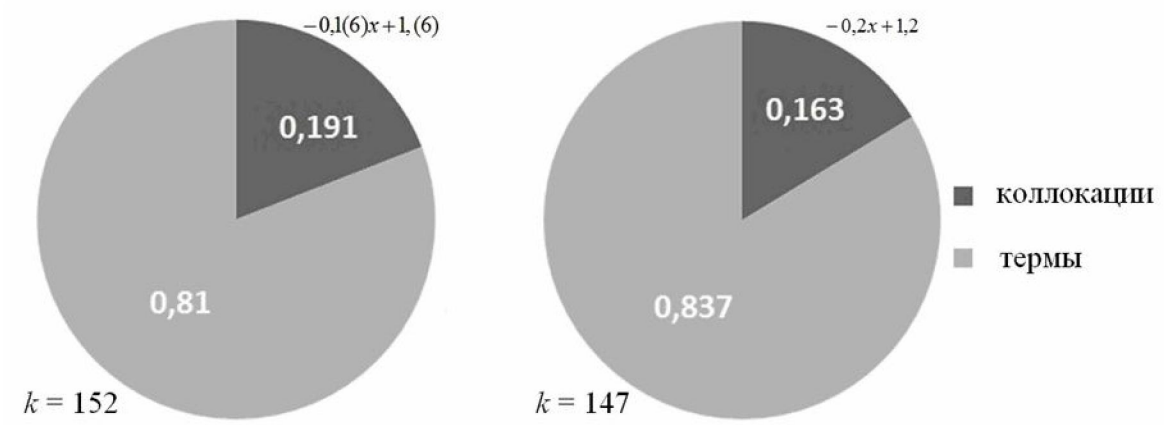

Рис. 4. Доля семантически значимых нечётких коллокаций (ϑ_k) в точках своего максимума

Согласно таблице 4 показатель, формализующий долю нечётких коллокаций в k наиболее семантически значимых факторах (ϑ_k), имеет тенденцию к росту с увеличением показателя k. На рисунке 5 показано возрастание ϑ_k при величине k, изменяющейся в диапазоне от 136 до 152.

По результатам экспериментов при k меньших 136 был сделан вывод, что k наиболее значимых факторов являются термами, то есть $\vartheta_k = 0$.

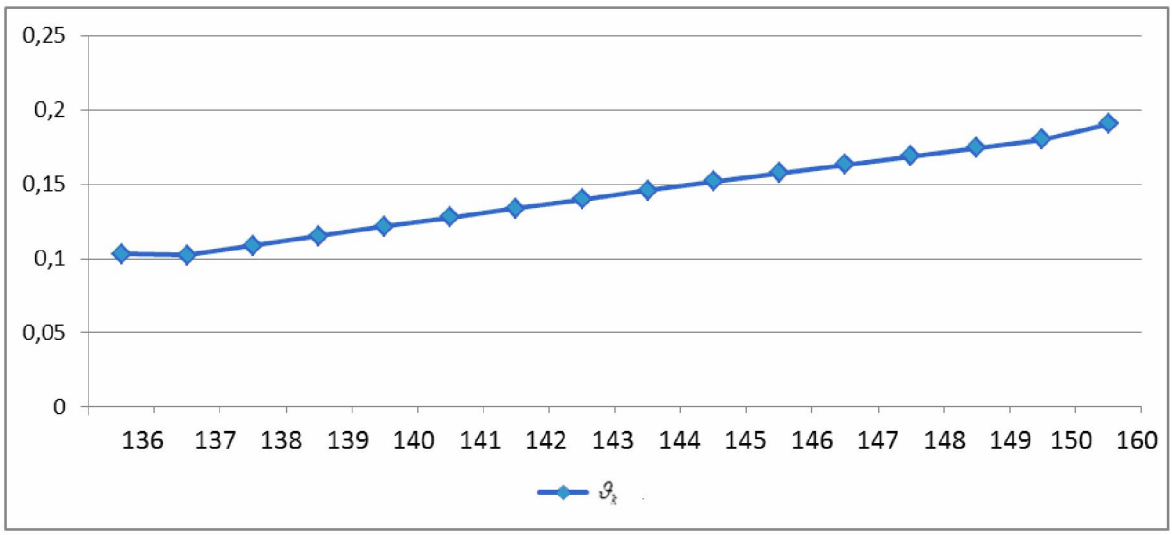

Рис. 5. Зависимость ϑ_k (вертикальная ось) от величины k (горизонтальная ось)

Заключение. Полученные по итогам вычислительных экспериментов сингулярные числа, соответствующие термам, существенно превосходят те же показатели для нечётких коллокаций. Вместе с тем при проведении вычислительных экспериментов использовался крайне робастный поход к созданию нечётких коллокаций, основанный на задании их функций принадлежности посредством лингвистической переменной без использования связок и модификаторов. Однако даже при таком огрублении модели построения коллокаций на основе лингвистической переменной показано, следующее.

1. Нечёткие коллокации имеют независимую от термов природу, то есть их семантическая значимость не является побочным продуктом соответствующей значимости, входящих в неё термов.

2. Целесообразность учёта нечётких коллокаций при кластеризации текстовых коллекций возрастает при увеличении числа искомых кластеров.

3. Наибольшей семантической значимостью обладают нечёткие коллокации, формализованные невозрастающими функциями принадлежности.

Проведённые вычислительные эксперименты продемонстрировали необходимость проведения дальнейших исследований нечётких коллокаций. В рамках новых исследований планируется увеличение числа ТК, а также усовершенствование методов построения функций принадлежности нечётких коллокаций с целью повышения их семантической значимости.

Список литературы

1. Авербух В. М. Шестой технологический уклад и перспективы России (краткий обзор) / В. М. Авербух // Вестник Ставропольского государственного университета. – 2010. – № 71. – С. 159–166.

2. Александреску А. Современное проектирование на C++: Обобщённое программирование и прикладные шаблоны проектирования / А. Александреску. – Москва : Вильямс, 2002. – 335 с.

3. Белоногов Г. Г. Языковые средства автоматизированных информационных систем / Г. Г. Белоногов, Б. А. Кузнецов. – Москва : Наука, 1983. – 288 с.

4. Гамма Э. Приёмы объектно-ориентированного проектирования. Паттерны проектирования / Э. Гамма, Р. Хелм, Р. Джонсон, Д. Влиссидес. – Санкт-Петербург : Питер, 2015. – 367 с.

5. Глазьев С. Ю. Эволюция технико-экономических систем: возможности и границы централизованного регулирования / С. Ю. Глазьев, Д. С. Львов, Г. Г. Фетисов – Москва : Наука, 1992. – 207 с.

6. Громов Ю. Ю. Формализация текстовой коллекции на основе нечетких частот коллокаций / Ю. Ю. Громов и другие // Приборы и системы. Управление, контроль, диагностика. – 2013. – № 2. – С. 15–17.

7. Ермаков А. Е. TopSOM: визуализация информационных массивов с применением самоорганизующихся тематических карт / А. Е. Ермаков, В. В. Плешко, Г. В. Липинский // Информационные технологии. – 2001. – № 8. – С. 1–7.

8. Ермаков А. Е. Ассоциативная семантическая сеть: статистическая модель восприятия и порождения текста / А. Е. Ермаков, В. В. Плешко // ООО «Гарант-Парк-Интернет». – Режим доступа: http://www.dialog-21.ru/Archive/2001/volume2/2_20.htm (дата обращения 28.01.2016), свободный. – Заглавие с экрана. – Яз. рус.

9. Ермаков А. Е. Полнотекстовый поиск: Проблемы и их решение / А. Е. Ермаков // Мир ПК. – 2001. – № 5. – Режим доступа: http://www.osp.ru/pcworld/2001/05/161575 (дата обращения 28.01.2016), свободный. – Заглавие с экрана. – Яз. рус.

10. Ермаков А. Е. Тематическая навигация в полнотекстовых базах данных / А. Е. Ермаков, В. В. Плешко // Мир ПК. – 2001. – № 8. – Режим доступа: http://www.osp.ru/pcworld/2001/08/162037 (дата обращения 28.01.2016), свободный. – Заглавие с экрана. – Яз. рус.

11. Киселев М. В. Метод кластеризации текстов, учитывающий совместную встречаемость ключевых терминов, и его применение к анализу тематической структуры новостного потока, а также его динамики / М. В. Киселев, М. М. Шмулевич, В. С. Пивоваров. – Москва : Компания Megaputer Intelligence, 2005. – 24 с.

12. Кондратьев Н. Д. Большие циклы конъюнктуры и теория предвиденья: Избранные труды / Н. Д. Кондратьев. – Москва : Экономика, 2002. – 264 с.

13. Ландэ Д. В. ИНТЕРНЕТИКА: Навигация в сложных сетях: модели и алгоритмы / Д. В. Ландэ, А. А. Санарский, И. В. Безсуднов. – Москва : ЛИБРОКОМ, 2009. – 264 с.

14. Леоненков А. Самоучитель UML / А. Леоненков. – Москва : Книга по требованию, 2006. – 417 с.

15. Недошивина Е. В. Учёт синтаксических связей при поиске коллокаций / Е. В. Недошивина // Natural Language Processing. – 2008. – № 4. – С. 1–3.

16. Операторы в поисковых запросах. – Режим доступа: https://support.google.com/websearch/answer/2466433?hl=ru&rd=1 (дата обращения 28.01.2016), свободный. – Заглавие с экрана. – Яз. рус.

17. Пивоварова Л. М. Извлечение и классификация терминологических коллокаций на материале лингвистических научных текстов / Л. М. Пивоварова, Е. В. Ягунова // Терминология и знание : материалы Симпозиума. – Москва, 2010. – С. 121–129.

18. Поляков Д. В. К вопросу построения математической модели кластеризации текстовых сведений / Д. В. Поляков и другие // Математические методы и информационно-технические средства : труды VIII Всероссийской научно-практической конференции. – Краснодар : Краснодарский университет Министерства внутренних дел России, 2012. – С. 164.

19. Поляков Д. В. Кластеризация текстовых коллекций на основе нечеткого описания коллокаций / Д. В. Поляков, О. Г. Иванова, А. Ю. Громова, В. Е. Дидрих // Информация и безопасность. – 2011. – № 3. – С. 459–462.

20. Поляков Д. В. Метод формализации нечётких коллокаций на основе фаззификации расстояний между термами в текстах / Д. В. Поляков, А. И. Елисеев, С. А. Дузькрятченко // Приборы и системы. Управление, контроль, диагностика. – 2015. – №12. – С. 50–61.

21. Поляков Д. В. Метод формализации нечётких коллокаций термов в текстах на основе лингвистических переменных / Д. В. Поляков, Н. М. Митрофанов, А. С. Матвеева //Прикаспийский журнал: Управление и высокие технологии. – 2015. – № 4 (32). – С. 167–183.

22. Поляков Д. В. Обобщение векторно-пространственной модели для оценки семантической значимости характеристик текстовых документов / Д. В. Поляков, Н. М. Митрофанов, Е. Н. Лепёшкин. // Приборы и системы. Управление, контроль, диагностика. – 2016. – № 1. – С. 35–44.

23. Пруцков А. В. Генерация и определение форм слов естественных языков на основе их последовательных преобразований / А. В. Пруцков // Вестник Рязанского государственного радиотехнического университета. – 2009. – № 27. – С. 51–58.

24. Пруцков А. В. Морфологический анализ и синтез текстов посредством преобразований форм слов / А. В. Пруцков // Вестник Рязанского государственного радиотехнического университета. – 2004. – № 15. – С. 70—75.

25. Фаулер М. Основы UML. Краткое руководство по стандартному языку объектного моделирования / М. Фаулер. – Санкт-Петербург : Символ, 2005. – 185 с.

26. Ягунова Е. В. От коллокаций к конструкциям / Е. В. Ягунова, Л. М. Пивоварова // Русский язык: конструкционные и лексико-семантические подходы. – Санкт-Петербург : Труды Института лингвистических исследований Российской академии наук, 2011. – С. 24–29.

27. Язык запросов Яндекса. – Режим доступа: https://yandex.ru/support/search/query-language/qlanguage.xml (дата обращения 28.01.2016) , свободный. – Заглавие с экрана. – Яз. рус.

28. Bisht R. K. Fuzzy Set Theoretic Approach To Collocation Extraction. / R. K. Bisht, H. S Dhami // International Journal of Computer Applications. – 2010. – Vol. 5, № 3. – P. 43–49.

29. Coplien J. O. Curiously Recurring Template Patterns / J. O. Coplien // C++ Report. – 1995. – P. 24–27.

30. Goldsmith J. Unsupervised Learning of the Morphology of a Natural Language / J. Goldsmith // Chicago: Association for Computational Linguistics. – 2001. – Vol. 27, № 2. – P. 173–194.

31. Gruber T. R. A translation approach to portable ontologies / T. R. Gruber // Knowledge Acquisition. – Stanford : Stanford University, 1993. – Vol. 5 – P. 199–220.

32. Jia Y. B. Singular Value Decomposition / Y. B. Jia // ComNotes. – 2015. – № 477. – P. 1–9.

33. Press H. W. Numerical Recipes in C. The Art of Scientific Computing Second Edition / H. W. Press, S. A. Teukolsky, W. T. Vetterling, B. P. Flannery. – Cambridge : Cambridge University Press, 1998. – 994 p.

34. Salton G. A Vector Space Model for Automatic Indexing / G. Salton, A. Wong, C. Yang // Communications of the ACM. – 1975. – № 11. – P. 613–620.

References

1. Averbukh V. M. Shestoy tekhnologicheskiy uklad i perspektivy Rossii (kratkiy obzor) [The sixth techno-economic paradigm and prospects of Russia (short overview)]. *Vestnik Stavropolskogo gosudarstvennogo universiteta* [Bulletin of the Stavropol State University], 2010, no. 71, pp. 159–166.

2. Aleksandresku A. *Sovremennoe proektirovanie na C++: Obobshchennoe programmirovanie i prikladnye shablony proektirovaniya* [Modern C++ Design: Generic Programming and Design Patterns Applied], Moscow, Vilyams Publ., 2002. 335 p.

3. Belonogov G. G., Kuznetsov B. A. *Yazykovye sredstva avtomatizirovannykh informatsionnykh system* [Language means of automated information systems], Moscow, Nauka Publ., 1983. 288 p.

4. Gamma E., Helm R., Dzhonson R., Vlissides J. *Priemy obektno-orientirovannogo proektirovaniya. Patterny proektirovaniya* [Design Patterns: Elements of Reusable Object-Oriented Software], Saint Petersburg, Piter Publ., 2015. 367 p.

5. Glazev S. Yu., Lvov D. S., Fetisov G. G. *Evolyutsiya tekhniko-ekonomicheskikh sistem: vozmozhnosti i granitsy tsentralizovannogo regulirovaniya* [The evolution of technical and economic systems: the possibilities and limits of centralized regulation], Moscow, Nauka Publ., 1992. 207 p.

6. Gromov Yu. Yu., Polyakov D. V., Avdeeva T. O. Formalizatsiya tekstovoy kollektsii na osnove nechetkikh chastot kollokatsiy [The formalization of the text based on fuzzy collection frequency collocations]. *Pribory i sistemy. Upravlenie, kontrol, diagnostika* [Instruments and Systems. Management, Monitoring, Diagnostics], 2013, no. 2, pp. 15–17.

7. Ermakov A. E., Pleshko V. V., Lipinskiy G. V. TopSOM: vizualizatsiya informatsionnykh massivov s primeneniem samoorganizuyushchikhsya tematicheskikh kart [TopSOM: visualization of text collections using self-organizing thematic maps]. *Informatsionnye tekhnologii* [Information Technology], 2001, no. 8, pp. 1–7.

8. Ermakov A. E., Pleshko V. V. Assotsiativnya model porozhdeniya teksta v zadache klassifikatsii [The associative model of generating text in classification problem]. *Informatsionnye tekhnologii* [Information Technology], 2000, no. 12. Available at: http://www.dialog-21.ru/Archive/2001/volume2/2_20.htm (accessed 28.01.2016).

9. Ermakov A. E. Problemy polnotekstovogo poiska i ikh reshenie [Problems of full text search and there solution]. *Mir PK* [PC World], 2001, no. 5. Available at: http://www.osp.ru/pcworld/2001/05/161575 (accessed 28.01.2016).

10. Ermakov A. E., Pleshko V. V. Tematicheskaya navigatsiya v polnotekstovykh bazakh dannykh [The problem of thematic navigation in the full-text databases]. *Mir PK* [PC World], 2001, no. 8. Available at: http://www.osp.ru/ /2001/08/162037 (accessed 28.01.2016).

11. Kiselev M. V., Pivovarov V. S., Shmulevich M. M. *Metod klasterizatsii tekstov, uchityvayushchiy sovmestnuyu vstrechaemost klyuchevykh terminov, i ego primenenie k analizu tematicheskoy struktury novostnogo potoka, a takzhe ego dinamiki* [The method of text clustering, that take into account the co-occurrence of key terms and use for analysis of the thematic structure of the news flow and its dynamics], Moscow, Megaputer Intelligence Publ., 2005. 24 p.

12. Kondratev N. D. *Bolshie tsikly konyunktury i teoriya predvidenya: Izbrannye trudy* [Big cycles of the conjuncture and the theory of forecusting: Selected Works], Moscow, Ekonomika Publ., 2002. 264 p.

13. Lande D. V., Sanarskiy A. A., Bezsudnov I. V. *Internetika: Navigatsiya v slozhnykh setyakh: modeli i algoritmy* [Internetika: Navigation in complex networks: models and algorithms], Moscow, LIBROKOM Publ., 2009. 264 p.

14. Leonenkov A. *Samouchitel UML* [Teach UML], Moscow, Kniga po trebovaniyu Publ., 2006. 417 p.

15. Nedoshivina E. V. Uchet sintaksicheskikh svyazey pri poiske kollokatsiy [Accounting syntactic links when searching collocations]. *Natural Language Processing*, 2008, no. 4, pp. 1–3.

16. *Operatory v poiskovyh zaprosah* [Operators in search queries]. Available at: https://support.google.com/websearch/answer/2466433?hl=ru&rd=1 (accessed 28.01.2016).

17. Pivovarova L. M., Yagunova E. V. Izvlechenie i klassifikatsiya terminologicheskikh kollokatsiy na materiale lingvisticheskikh nauchnykh tekstov [Extraction and classification of collocation from the matirial of linguistic, scientific texts]. *Terminologiya i znanie : materialy Simpoziuma* [Terminology and Knowledge. Proceedings of the Symposium], Moscow, 2010, pp. 121–129.

18. Polyakov D. V., Samoylov V. V., Al'-Balushi M. P., Hak D. L. K voprosu postroeniya matematicheskoy modeli klasterizatscii tekstovkyh svedeniy [The problem of constructing a mathematical model for clustering text information]. *Matematicheskie metody i informatsionno-tekhnicheskie sredstva : trudy VIII Vserossiyskoy nauchno-prakticheskiy konferentsii* [Mathematical Methods and Information Technology Equipment. Proceedings of VIII Scientificn and Practical Conference], Krasnodar, Krasnodar University of the Ministry of Internal Affairs of Russia Publ. House, 2012, pp. 164.

19. Polyakov D. V., Ivanova O. G., Gromova A. Yu., Didrikh V. E. Klasterizatsya tekstovykh kollektsiy na osnove nechetkogo opisaniya kollokatsiy [Clustering of text collections based on fuzzy collocations]. I*nformatsiya i bezopasnost* [Information and safety], 2011, no. 3, pp. 459–462.

20. Polyakov D. V., Eliseev A. I., Duzkryatchenko S. A. Metod formalizatsii nechetkikh kollokatsiy termov v tekstakh na osnove lingvisticheskikh peremennykh [Method of formalization of fuzzy collocations based on distance analysis location of terms in text]. *Pribory i sistemy. Upravlenie, kontrol, diagnostika* [Instruments and Systems. Management, Monitoring, Diagnostics], 2015, no. 12, pp. 50–61.

21. Polyakov D. V., Mitrofanov N. M., Matveeva A. S. Metod formalizatsii nechetkikh kollokatsiy termov v tekstakh na osnove lingvisticheskikh peremennykh [Method of formalization of fuzzy collocations in texts based on linguistic variables]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Management and High Technologies], 2015, no. 4 (32), pp. 167–183.

22. Polyakov D. V., Mitrofanov N. M., Lepeshkin E. N. Obobshchenie vektorno-prostranstvennoy modeli dlya otsenki semanticheskoy znachimosti kharakteristik tekstovykh dokumentov [Generalized LSA method and its use for assessing the significance of fuzzy collocation in text collections]. *Pribory i sistemy. Upravlenie, kontrol, diagnostika* [Instruments and Systems. Management, Monitoring, Diagnostics], 2016, no. 1, pp. 35–44.

23. Prutskov A. V. Generatsiya i opredelenie form slov estestvennykh yazykov na osnove ikh posledovatelnykh preobrazovaniy [Generation and identification of forms of words of natural languages based on their successive transformations]. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta* [Bulletin of the Ryazan State Radio Engineering University], 2009, no. 27, pp. 51–58.

24. Prutskov A. V. *Morfologicheskiy analiz i sintez tekstov posredstvom preobrazovaniy form slov* [Morphological analysis and synthesis of texts using transforming the forms of words]. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta* [Bulletin of the Ryazan State Radio Engineering University], 2004, no. 15, pp. 70–75.

25. Fauler M. *Osnovy UML. Kratkoe rukovodstvo po standartnomu yazyku obektnogo modelirovaniya* [UML Basics. Quick guide to the language of objective modeling], Saint Petersburg, Simvol Publ., 2005. 185 p.

26. Yagunova E. V., Pivovarova L. M. Ot kollokatsiy k konstruktsiyam [From collocations to constructions]. *Russkiy yazyk: konstruktsionnye i leksiko-semanticheskie podkhody* [Russian Language: Structural and Lexical and Semantic Approaches], Saint Petersburg, Proceedings of the Institute of Linguistic Studies Publ. House, 2011. 43 p.

27. *Yazyk zaprosov Yandeksa* [The query language of Yandex]. Available at: https://yandex.ru/support/search/query-language/qlanguage.xml (accessed 28.01.2016).

28. Bisht R. K., Dhami H. S. Fuzzy Set Theoretic Approach To Collocation Extraction. *International Journal of Computer Applications*, 2010, vol. 5, no. 3, pp. 43–49.

29. Coplien J. O. Curiously Recurring Template Patterns. *C++ Report*, 1995, pp. 24–27.

30. Goldsmith J. Unsupervised Learning of the Morphology of a Natural Language. *Association for Computational Linguistics*, 2001, vol. 27, no. 2, pp. 173–194.

31. Gruber T. R. A translation approach to portable ontologies. *Knowledge Acquisition*, Stanford, Stanford University Publ. House, 1993, vol. 5, pp. 199–220.

32. Jia Y. B. Singular Value Decomposition. C*omNotes*, 2015, no. 477. pp. 1–9.

33. Press H. W., Teukolsky S. A., Vetterling W. T., Flannery B. P. *Numerical Recipes in C. The Art of Scientific Computing Second Edition*, Cambridge, Cambridge University Press Publ. House, 1998. 994 p.

34. Salton G., Wong A., Yang C. A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 1975, no. 11, pp. 613–620.