К ВОПРОСУ ПОСТРОЕНИЯ И ПРИМЕНЕНИЯ НЕЧЁТКИХ КОЛЛОКАЦИЙ ДЛЯ СЕМАНТИЧЕСКОГО АНАЛИЗА ТЕКСТОВЫХ КОЛЛЕКЦИЙ

Поляков Д. В.[1]

(*ФГБОУ ВО «Тамбовский государственный технический университет», Тамбов*)

Приводится расширенная векторно-пространственная модель текстовой коллекции в пространстве нечётких факторов. Рассматривается понятие нечёткой коллокации, как семантически значимого элемента текстовой коллекции. Показано, что при определённом значении параметров расширенной векторно-пространственной модели в качестве нечётких факторов допустимо совместное использование термов и коллокаций. На основе данной возможности предложен подход к повышению семантической значимости нечётких коллокаций посредством уточнения вида их функций принадлежности. Данный подход базируется на решении оптимизационной задачи поиска вида функции принадлежности посредством модернизированного алгоритма дифференциальной эволюции. Приведены результаты исследований в области модернизации алгоритма дифференциальной эволюции, в том числе гипотезы повышения сходимости алгоритма, постановка вычислительных экспериментов и их результаты. В качестве целевой функции Fitness эволюционного алгоритма выбрана оценка семантической значимости нечётких коллокаций на основе латентно-семантического анализа. Ключевой особенностью данного анализа является SVD-разложение матрицы расширенной векторно-пространственной модели, которое позволяет оценить семантическую значимость нечётких коллокаций в сравнении с наиболее значимыми термами. В заключение рассматриваются направления дальнейших исследований в области построения нечётких коллокаций, уточнения их функций принадлежности, а также применения нечётких коллокаций для решения прикладных задач, связанных с семантическим анализом текстовых коллекций.

Ключевые слова: текстовая коллекция, нечёткая коллокация, векторно-пространственная модель, SVD-разложение, латентно-семантический анализ, алгоритм дифференциальной эволюции.

1. Введение

Естественный язык является неотъемлемой частью человеческой культуры и проявляет себя во всех видах деятельности.

[1] *Дмитрий Вадимович Поляков, к.т.н.*(*dimadress@yandex.ru*).

Сегодня перед компьютерной лингвистикой стоит множество задач автоматизированной обработки текстовой информации, в том числе: выявление семантически значимых элементов, определение близости текстовых документов, поиск и фильтрация информации на больших текстовых коллекциях и многое другое.

Одним из известных подходов в представлении и анализе текстовой коллекции является латентно-семантический анализ [1, 6], который базируется на сингулярном разложении матрицы *tf-idf*. Вместе с тем в ходе формализации текстовой коллекции в виде этой матрицы теряется связь между термами в тексте. С целью снижения потерь семантики при формализации текстовой коллекции и учёта нечётких коллокаций, задающих связи между термами, была введена в рассмотрение расширенная векторно-пространственная модель текстовой коллекции в пространстве нечётких факторов.

2. Расширенная векторно-пространственная модель текстовой коллекции

Пусть D – множество текстовых документов, $D = \{d_1, d_2, \ldots, d_N\}$, $|D| = N$, где $|.|$ – мощность множества; а $\mathcal{T}$ – универсум термов, $\mathcal{T} = \{t_1, t_2, \ldots, t_n\}$, $|\mathcal{T}| = n$. Появление t в d_i обозначим $t \in d_i$. Поставим в соответствие каждому терму t_i, $1 \le i \le n$, множество $\hat{T}_i$ – совокупность его словоформ. Введём обозначение: $\Theta = \left\{\hat{T}_1, \hat{T}_2, ..., \hat{T}_n\right\}$ Тогда произвольный $d_i \in D$, $1 \le i \le N$, целесообразно представить в виде вектора словоформ:

(1) $d_i\left(\hat{t}_{k_1}, \hat{t}_{k_2}, ..., \hat{t}_{k_{m_i}}\right)$,

где $\hat{t}_{k_j} \in \hat{T}_{k_j} \in \Theta$, $k_j \in \{1, 2, ..., n\}$, $j = \overline{1, m_i}$.

Заметим, что представление документа в виде (1) практически не приводит к потере семантической значимости. То есть по основе (1) с точностью до авторских знаков препинания восстанавливается исходный текст.

Вместе с тем формализация документа в рамках некоторой модели приводит к утере части семантической информации.

Рассмотрим некоторое произвольное преобразования документа d_i, в результате которого он представляется в виде множества характеристических объектов (p). Так как d_i потерял часть семантической информации, обозначим рассматриваемое множество $\hat{d}_i$, являющееся оценкой d_i. Тогда:

(2) $\hat{d}_i = \left\{p_1^i, p_2^i, ..., p_{M_i}^i\right\}$,

где $\left|\hat{d}_i\right| = M_i$. Элементы множества $\hat{d}_i$ – это присутствующие в документе артефакты (конструкции, структуры), характеризующие семантику d_i. Представление документа в виде (2) задаёт универсальное множество характеристических объектов $U_p = \bigcup_{i=1}^{N} \hat{d}_i$. Действительно, если первоначально определить универсум U_p, то множество $\hat{d}_i, i = \overline{1, N},$ определяется как

(3) $\hat{d}_i = \left\{p \in U_p \mid p \in d_i\right\}$,

причём количество вхождений объекта p в $\hat{d}_i$ равно числу его появлений в документе d_i.

Рассмотрим множество $U_F = \left\{\tilde{F}_1, \tilde{F}_2, ..., \tilde{F}_K\right\}, \left|U_F = K\right|$. U_F представляет собой универсальное множество характеристик (пространство факторов), по которым оценивается семантический смысл текстового документа. Каждый элемент U_p – нечёткое множество $\tilde{F}_i, i = \overline{1, K}$, задаваемое функцией принадлежности μ_i: $U_p \rightarrow [0, 1]$, т.е. $\tilde{F}_i \subset U_p, \forall i = \overline{1, K}$.

Представление текстового документа представляет собой отображение некоторого $d_i \in D$ вида (1) на пространство факторов U_F. Таким образом, произвольный документ $d_i \in D$ по результатам формализации будет представлен в виде вектора:

(4) $d_i\left(f_1^i, f_2^i, ..., f_K^i\right)$,

где $f_j^i, i = \overline{1, N}, j = \overline{1, K}$ – соответствие $\hat{d}_i$ фактору $\tilde{F}_j$.

Рассмотрим семейство нечётких множеств $\tilde{W}_i \subset D, i = \overline{1,K}$. Семантически $\tilde{W}_i$ определяется как множество документов, в которых присутствует фактор $\tilde{F}_i$, и задаётся с помощью функции $\mu_\omega^i : D \to [0,\ 1]$. Заметим, что $\mu_\omega^i\left(\hat{d}_j\right)$ должна возрастать при росте кардинального числа множества $\tilde{F}_i \cap \hat{d}_j$, а максимум μ_ω^i должен достигаться, когда все признаки $\hat{d}_j$ принадлежат $\tilde{F}_i$ со степенью принадлежности 1. Минимальное значение $\mu_\omega^i\left(\hat{d}_j\right)$, равное 0, достигается, когда степень принадлежности каждого признака из $\hat{d}_j$ к фактору $\tilde{F}_i$ равно 0. Заметим, что во всех точках области определения, поведение $\mu_\omega^i\left(\hat{d}_j\right)$ совпадает с поведением $\left|\tilde{F}_i \cap \hat{d}_j\right|$.

Тогда функция принадлежности $\tilde{W}_i$ имеет вид:

$$(5)\quad \mu_\omega^i\left(\hat{d}_j\right) = \xi\left(\frac{\sum_{k=1}^{M_j} \mu_i\left(p_k^j\right)}{M_j}\right),$$

где ξ: [0, 1]→[0, 1], $\xi(0) = 0$, $\xi(1) = 1$.

Рассмотрим $\tilde{U}_F^D \subset U_F$, где $\tilde{U}_F^D$ – нечёткое множество значимых для коллекции D факторов. Пусть $\tilde{U}_F^D$ определяется функцией μ_F: $U_F \to [0, 1]$. Исследуем подробно поведение функции μ_F.

Было показано [3], что высказывание, формализующее присутствие фактора F в $\hat{d}$, будет иметь вид

$$(6)\quad \mu_F\left(\hat{d}\right) = S\left(\mu(p_1), \mu(p_2), \ldots, \mu(p_M)\right),$$

где $\mu^F(\cdot)$ – функция, отражающая присутствие $\tilde{F}$ в документе $\hat{d}$, $\mu(\cdot)$ – функция принадлежности фактора $\tilde{F}$, а $S(\ldots)$ – S-норма. Тогда $\mu_F\left(\tilde{F}\right)$ принимает вид

$$(7)\quad \mu_F\left(\widetilde{F}\right)=\zeta\left(\frac{N}{1+\sum\limits_{i=1}^{N}S\left(\mu\left(p_1^i\right),\mu\left(p_2^i\right),...,\mu\left(p_{M_i}^i\right)\right)}\right),$$

где $\mu(\cdot)$ – функция, формализующая фактор $\widetilde{F}$, а $\zeta(\cdot)$ – ζ: [1, |D|] → [0, 1], $\zeta(1) = 0$, $\zeta(|D|) = 1$ и $\zeta\uparrow$ на [1, |D|].

А элемент матрицы *tf-idf* – f_j^i принимает вид:

$$(8)\quad f_j^i=T\left(\xi\left(\frac{\sum\limits_{k=1}^{M_j}\mu_i\left(p_k^j\right)}{M_j}\right),\zeta\left(\frac{N}{1+\sum\limits_{i=1}^{N}S\left(\mu_j\left(p_1^i\right),\mu_j\left(p_2^i\right),...,\mu_j\left(p_{M_i}^i\right)\right)}\right)\right).$$

Заметим, что при $\xi(x) = x$, $\forall x \in R$, $\zeta(x) = \log(x)$, $\forall x \in R$, $T(x, y) = xy$, $\forall x, y \in R$, а $S(x, y) = x + y – xy$, (8) принимает вид

$$(9)\quad f_j^i=\frac{\sum\limits_{k=1}^{M_j}\mu_i\left(p_k^j\right)}{M_j}\log\left(\frac{N}{1+\sum\limits_{i=1}^{N}S\left(\mu_j\left(p_1^i\right),\mu_j\left(p_2^i\right),...,\mu_j\left(p_{M_i}^i\right)\right)}\right),$$

где для $\forall m \geq 2$

$$S\left(x_1,x_2,...,x_m\right)=\begin{cases}x_1+x_2-x_1x_2,\ m=2,\\ S\left(x_m,S\left(x_1,x_2,...,x_{m-1}\right)\right),\ m>2.\end{cases}$$

Причём если множество характеристических объектов – $\mathcal{T}$, а множество факторов – Θ, то показано [3], что (9) сводится к классической векторно-пространственной модели текстовой коллекции [2].

Вместе с тем представление текстовой коллекции в виде (9) позволяет строить векторно-пространственную модель не только относительно термов, но и с использованием нечётких коллокаций [3]. Более того, если в качестве характеристических объектов рассматривать группы слов текста, находящихся на некотором расстоянии друг от друга, а факторами выбрать одновременно термы и коллокации, то (9) позволит объединить термы и нечёткие коллокации в рамках одной векторно-пространственной модели. Причём латентно-семантический анализ на основе этой модели позволит оценить семантическую

значимость нечётких коллокаций в сравнении с аналогичной значимостью термов.

Вместе с тем основной проблемой является построение самих нечётких коллокаций для задания пространства факторов, а именно нахождение соответствующих семантически значимым коллокациям функций принадлежности. Эта проблема рассматривается в ряде работ [4, 5]. Однако все предложенные подходы позволяют построить лишь робастные варианты таких функций.

3. Модернизированный алгоритм дифференциальной эволюции

Для решения задачи поиска функций принадлежности было решено воспользоваться эволюционными алгоритмами. В ходе анализа последних был выбран алгоритм дифференциальной эволюции [7, 8].

Вместе с тем к настоящему моменту эволюционные процессы, происходящие в биологических системах, изучены глубже, чем на момент формирования концепции эволюционных алгоритмов. Таким образом, имеются современные исследования об эволюции живых существ, на основании которых целесообразно рассмотреть группу методов модернизации алгоритма дифференциальной эволюции. Поэтому имеет смысл сформулировать ряд гипотез, которые потенциально усовершенствуют алгоритм дифференциальной эволюции.

Был выдвинут ряд гипотез по модернизации алгоритма дифференциальной эволюции. Приведём краткое описание рассмотренных гипотез.

***Гипотеза* 1**. В основе данной гипотезы лежит идея о том, что эволюционировать могут не только особи, но и некоторые параметры алгоритма. Основная идея заключается в том, что мутации происходят с некоторой вероятностью, и эта вероятность, как и сила мутации, эволюционирует вместе с особью. Действительно, в биологических системах мутации не обязательно происходят в каждом гене. Более того, мутации довольно редки и происходят с некоторой вероятностью.

***Гипотеза* 2**. Одним из важнейших свойств наследования в биологических системах является то, что в живых организмах на каждый признак влияет множество генов, таким образом, сразу несколько генов могут влиять, усиливать или ослаблять тот или иной признак, т.е. иметь достаточно сложную связь. Кроме того, огромная часть генов является неактивной, у них нет экспрессии, т.е. они находятся, по сути, в спящем состоянии.

***Гипотеза* 3**. В эволюционной биологии довольно давно существовала гипотеза о синергетическом эффекте отрицательных (негативных) мутаций. В ряде новых исследований данный эффект был установлен экспериментальным путем. Суть синергетического эффекта негативных мутаций состоит в том, что последние усиливают отрицательный эффект друг друга в случае появления у одной и той же особи. Таким образом, гибнут или оставляют мало потомства особи, в которых скопились сразу несколько аллелей (вариантов генов) с негативными мутациями. Такая ситуация позволяет кардинально уменьшить число аллей с негативными мутациями в результате элиминации относительно небольшого количества особей.

***Гипотеза* 3**. Одним из известных и широко признанных драйверов эволюционного процесса в живой природе является так называемая «эволюционная гонка» – взаимодействие «хищник – жертва», в рамках которой обеим сторонам приходится непрерывно наращивать эволюционное преимущество. Количественно такое взаимодействие описывается уравнением Лотки – Вольтерры. Поэтому, согласно данной гипотезы, уравнение Лотки – Вольтерры можно использовать для контроля численности популяции, так как сохранение численности неизменной является весьма искусственным приёмом, не соответствующим реальности в биологических системах.

Для анализа данных гипотез были спроектированы и поставлены вычислительные эксперименты. Было показано, что гипотеза 2 не позволяет сократить время сходимости классического алгоритма дифференциальной эволюции. Среднее число итераций алгоритма, построенного на гипотезе 2, стремительно растет при повышении сложности *Fitness*.

Данная гипотеза затрачивает на нахождение решения в пятимерном пространстве в среднем 281 шаг. В это же время другие функции затрачивают на поиск решения в среднем от 19 до 28 шагов. В то время как оставшиеся гипотезы демонстрировали положительное влияние на алгоритм дифференциальной эволюции (рис. 1).

При проведении экспериментов осуществлялась автоматическая генерация полиномиальных функций с заранее известными точками экстремума. В силу стохастического характера работы алгоритма на каждой функции осуществлялось 1000 запусков алгоритма. В ходе вычислительных экспериментов было сгенерировано более 10000 функций для каждой исследуемой размерности пространства (от 5 до 14). Данные границы исследуемых размерностей связаны с тем, что между термами предполагается не больше пятнадцати нечётких коллокаций [2]. Таким образом, было проведено более 100 млн запусков базового алгоритма и модернизированных по каждой из гипотез.

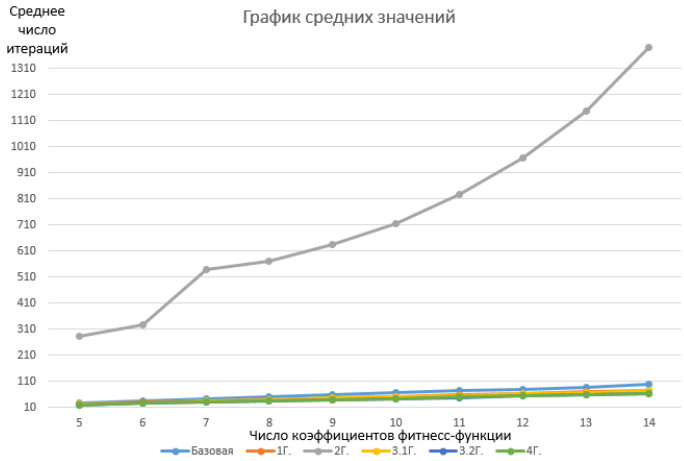

Рис. 1. Оценка сходимости алгоритма при его модернизации на основе представленных гиптез

На рис. 2 представлены те же 5 графиков, но в другом масштабе.

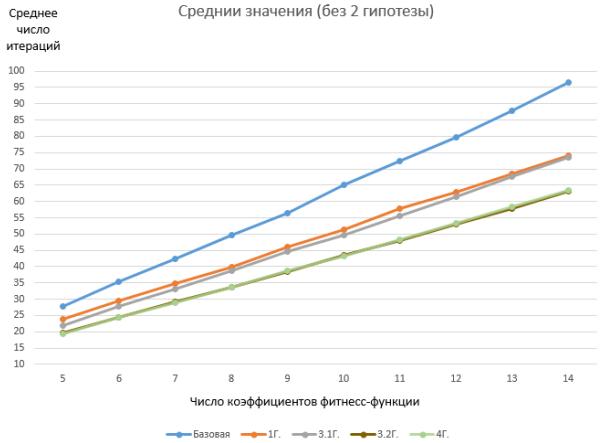

Рис. 2. Сравнение сходимости алгоритма при его модернизации на основе гипотез с положительным эффектом

Из рис. 2 легко видеть, что 1, 3 и 4 гипотезы повышают сходимость исследуемого алгоритма. Вместе с тем наибольший положительный эффект на алгоритм оказывает 4 гипотеза.

4. Заключение

Результаты проведённых исследований говорят о целесообразности использования алгоритма дифференциальной эволюции, модернизированного посредством четвёртой гипотезы (контроль размера популяции на основе уравнения Лотки – Вольтерры) для поиска вида функций принадлежности нечётких коллокаций.

В рамках дальнейших исследований планируется проектирование и постановка вычислительных экспериментов по поиску нечётких коллокаций с высокой семантической значимостью. Идея экспериментов состоит в том, чтобы посредством латентно-семантического анализа найти наиболее значимые термы и, воспользовавшись расширенной векторно-пространственной моделью (9), оценивать семантическую значимость коллокаций. Предполагается, что использование данной оценки в качестве функции *Fitness* в предложенном эволюционном алгоритме поз-

волит найти нечёткие коллокации с наибольшей семантической значимостью для исследуемой текстовой коллекции.

Литература

1. БЕЛОНОГОВ Г.Г., КУЗНЕЦОВ Б.А. *Языковые средства автоматизированных информационных систем*. – М.: «Наука», 1983. – 288 с.
2. ЛАНДЭ Д.В., САНАРСКИЙ А.А., БЕЗСУДНОВ И.В. *Интернетика: Навигация в сложных сетях: модели и алгоритмы*. – М.: ЛИБРОКОМ, 2009. – 264 с.
3. ПОЛЯКОВ Д.В., МИТРОФАНОВ Н.М., ЛЕПЁШКИН Е.Н. *Обобщение векторно-пространственной модели для оценки семантической значимости характеристик текстовых документов* // Приборы и системы. Управление, контроль, диагностика. – 2016. – №1 – С. 35–44.
4. ПОЛЯКОВ Д.В., ЕЛИСЕЕВ А.И., ДУЗЬКРЯТЧЕНКО С.А. *Метод формализации нечётких коллокаций на основе фаззификации расстояний между термами в текстах* // Приборы и системы. Управление, контроль, диагностика. – 2015. – №12 – С. 50–61.
5. ПОЛЯКОВ Д.В., МИТРОФАНОВ Н.М., МАТВЕЕВА А.С. *Метод формализации нечётких коллокаций термов в текстах на основе лингвистических переменных* // Прикаспийский журнал: Управление и высокие технологии. – 2015. – №4(32) – С. 167–183.
6. GOLDSMITH J. *Unsupervised Learning of the Morphology of a Natural Language* // Chicago: Association for Computational Linguistics. – 2001. – Vol. 27, No. 2. – P. 173–194.
7. PRICE K., STORN R., Lampinen J. *Differential Evolution: A Practical Approach to Global Optimization*. – Springer, 2005. – 538 p.
8. STORN R., PRICE K. *Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces* // Journal of Global Optimization. – 1997. – Vol. 11. – P. 341–359.

TO THE PROBLEM OF CONSTRUCTION AND APPLICATION OF FUZZY COLLOCATIONS FOR SEMANTIC ANALYSIS OF TEXT COLLECTIONS

Dmitry Polyakov, Tambov State Technical University, Tambov, Cand.Sc. (dimadress@yandex.ru)

Abstract: The paper presents an expanded vector-space model of a text collection in the space of fuzzy factors. The concept of fuzzy collocation as a semantically significant element of a text collection is considered. It is shown that for a certain value of parameters of the extended vector-space model, the combined use of terms and collocations is possible as fuzzy factors. Base on it an approach to increase the semantic significance of fuzzy collocations by specifying their membership functions is proposed. This approach based on solving the optimization problem of finding the membership function by means of a modernized differential evolution algorithm. The results of research of modernization of the differential evolution algorithm are presented, including the hypothesis of increasing the convergence of the algorithm, the setting of computational experiments and their results. Evaluation of the semantic significance of fuzzy collocations based on latent semantic analysis have been chosen as a Fitness function of evolutionary algorithm. The key latent semantic analysis procedure - SVD decomposition of the matrix of the extended vector-space model allows one to assess the semantic significance of fuzzy collocations in comparison with the most significant terms. In conclusion the directions for further research in the constructing fuzzy collocations, finding their membership functions, and using fuzzy collocations for solving semantic analysis applied problems of text collections are considered.

Keywords: text collection, fuzzy collocation, vector-space model, SVD decomposition, latent semantic analysis, differential evolution algorithm.

УДК 004.91
ББК 16.0