29. Popov S. I., Rogozin Ye. A., Roslov S. Yu. Analiz sovremennykh metodov i algoritmov optimizatsii na etape formirovaniya struktury i sostava kompleksa tekhnicheskikh sredstv zashchity informatsiina obekte informatazatsii [The analysis of modern optimization methods and algorithms at a stage of structure and nomenclature formation of technical means complex for information protection at object of informatization]. *Vestnik Voronezhskogo gosudarstvennogo tekhnicheskogo universiteta* [Bulletin of the Voronezh State Technical University], 2009, vol. 5, no. 6, pp. 83–85.

30. Raykova N. O. Ob integratsii sistem menedzhmenty informatsionnoy bezopasnosti i kachestva [About integration of system management of information security and quality]. *Voprosy kiberbezopasnosti* [Cybersafety Questions], 2013, no. 3, pp. 47–53.

31. Savochkin A. Ye. Algoritmizatsiya raboty sistem monitoringa i kontrolya dlya resheniya zadach identifikatsii stepeni povrezhdeniya tekhnicheski slozhnykh obektov [Algoritmization of monitoring and control systems work for identification problems solution of technically complex objects rate damage]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Management and High Technologies], 2014, no. 2, pp. 23–35.

32. Smirnov V. I. Seysmoizolyatsiya – sovremennaya antiseysmicheskaya zashchita zdaniy v Rossii [Seismoisolation – modern aseismic protection of buildings in Russia]. *Seysmostoykoe stroitelstvo. Bezopasnost sooruzheniy* [The Seismoresistant Construction. Safety of Constructions], 2013, no. 4, pp. 41–54.

33. Sobakin I. B. Sistemnyy podkhod k upravleniyu riskami informatsionnoy bezopasnosti [System approach to risk management of information security]. *Aktualnye problemy sovremennoy nauki* [Actual Problems of Modern Science], 2013, no. 3 (71), pp. 39–40.

34. Sobolev V. V., Babkin O. A. Modelirovanie i optimizatsiya usloviy primeneniya videoregistratsionnogo kontrolya kachestva pri stroitelstve zdaniy [Modeling and optimization of application conditions of video registration quality control during buildings construction]. *Internet-zhurnal Naukovedenie* [Research of Science. Internet Journal], 2014, no. 6 (25), pp. 19.

35. Starikovskiy A. V., Zhukov I. Yu., Mikhaylov D. M., Sheptunov A. A., Savchuk A. V., Krymov A. S. Povyshenie zashchishchennosti sistem avtomatizatsii upravleniya zdaniyami ot kompyuternykh atak [Increasing security from computer attacks for buildings management automation systems]. *Spetstekhnika i svyaz* [Special Equipment and Communication], 2012, no. 4, pp. 2–5.

36. Chesnokova O. Ye., Andreev V. M. Energoeffektivnye tekhnologii, ispolzuemye pri proektirovanii obshchestvennykh zdaniy [The power effective technologies used for public buildings design]. *Aktualnye problemy sovremennoy nauki, tekhniki i obrazovaniya* [Actual Problems of Modern Science, Equipment and Education], 2013, vol. 2, no. 71, pp. 223–225.

37. Sheshenya N. Kriterii inzhenerno-geologicheskogo obosnovaniya meropriyatiy po zashchite zdaniy i sooruzheniy ot opasnykh opolznevykh protsessov [Criteria of engineering and geological justification of actions for buildings and constructions protection from dangerous landslide processes]. *Inzhenernaya zashchita* [Engineering Protection], 2015, no.3 (8), pp. 44–55.

УДК 004:912

МЕТОД ФОРМАЛИЗАЦИИ НЕЧЁТКИХ КОЛЛОКАЦИЙ ТЕРМОВ В ТЕКСТАХ НА ОСНОВЕ ЛИНГВИСТИЧЕСКИХ ПЕРЕМЕННЫХ[1]

Статья поступила в редакцию 22.10.2015 г., в окончательном варианте 5.11.2015 г.

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

Митрофанов Николай Михайлович, магистрант, лаборант кафедры, Тамбовский государственный технический университет, 392000, Российская Федерация, г. Тамбов, ул. Советская, 106, e-mail: n.mitrofanow@gmail.com

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

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

Целью работы является создание и исследование математических методов формализации коллокаций в текстах. Это позволит повысить качество поиска и кластеризации текстовых коллекций путём введения в вектор признаков, представляющий в модели текст, элементов, формализующих коллокации с учётом расстояния в них между термами. Методика исследований основана на теории нечётких множеств, теории информационного поиска и теории матриц. Представленные в данной работе исследования не затрагивают вопроса использования полученных методов формализации текстовых коллекций для решения задач поиска и кластеризации. Кроме того, предложенные модель и методы ограничены рассмотрением коллокаций, состоящих из двух термов. Вместе с тем очерчен круг необходимых в дальнейшем теоретических и экспериментальных исследований с целью оценки целесообразности применения результатов данной работы для решения задач поиска и кластеризации. В работе предложен метод формализации коллокаций термов с учётом расстояния между ними на основе теории нечётких множеств. Под расстоянием между термами в коллокации понимается количество слов, появившихся между ними (термами) в тексте. Предложенный метод заключается в формализации данного расстояния посредством лингвистической переменной. По результатам исследования предложена расширенная векторно-пространственная модель коллекции документов. Она позволяет провести сравнительный анализ важности термов и коллокаций, а также обобщить алгоритмы, базирующиеся на *svd*-разложении матриц, благодаря учёту коллокаций в векторно-пространственной модели.

Ключевые слова: коллокация, текстовые коллекции, нечёткие коллокации, теория нечётких множеств, лингвистическая переменная, кластеризация текстовых коллекций, поиск в текстовых коллекциях, информационный поиск

METHOD OF FORMALIZATION OF FUZZY COLLOCATIONS IN TEXTS BASED ON LINGUISTIC VARIABLES

Polyakov Dmitriy V., Ph.D. (Engineering), senior lecturer, Tambov State Technical University, 106 Sovetskaya St., Tambov, 392000, Russian Federation, e-mail: dimadress@yandex.ru
Mitrofanov Nikolay M., undergraduate, assistant of department, Tambov State Technical University, 106 Sovetskaya St., Tambov, 392000, Russian Federation, e-mail: n.mitrofanow @gmail.com
Matveeva Alena S., post-graduate student, Tambov State Technical University, 106 Sovetskaya St., Tambov, 392000, Russian Federation, e-mail: klenchic@mail.ru

The purpose of the article is the development of mathematical methods of formalizing the collocation in the texts. This can help to improve the quality of search and clustering text collections, through the introduction of collocations in the vector space model, considering the distance between terms. In the research are used theories of fuzzy sets, information retrieval and matrices. Researches, given in this article, are not answer at such questions as how to use this collocation for informational retrieval or text clustering, moreover all given researches are limited by a consideration of collocation as a pair of terms. Method of formalization of the collocation, which considering the distance between terms using the theory of fuzzy sets, is offered. This method consists in the formalization of the distance between terms by means of linguistic variable. Moreover, in the article enhanced vector space model of the text collection is offered, which give us a tool to conduct comparative analysis of using terms and fuzzy collocations for informational retrieval.

Keywords: collocation, text collection, fuzzy collocation, theory of fuzzy sets, linguistic variable, clustering of text collection, search in text collections, information retrieval

Введение. Последние десятилетия были отмечены бурным ростом объёмов доступной человечеству информации. Развитие сетевых информационных систем, их последующая интеграция в глобальную сеть Интернет, а также быстрый рост последней привели к тому,

что если ранее проблемой было получить доступ к искомой информации, то, на сегодняшний день, сложилась ситуация, когда искомая информация, зачастую, находится в открытом доступе, но найти её крайне сложно. Это происходит из-за огромного количества «шумовой» (не отвечающей информационным потребностям конкретного пользователя) информации, которую очень сложно отделить от искомой, используя средства современных информационно-поисковых машин (ИПМ).

Таким образом, востребованность информации зависит не только от её значимости, но и от качества работы ИПМ [3]. А качество работы ИПМ, в свою очередь, зависит от используемых моделей и алгоритмов поиска информации и ранжирования поисковой выдачи. Это ранжирование, благодаря широкому использованию в современных ИПМ подходов предложенных Л. Пэйджем и С. Брином [13], часто во многом опирается на формализацию совокупности сайтов в виде графа. Связи в этом графе задают ссылки между сайтами, а ранг сайта в поисковой выдаче определяется на основе структуры графа. Вместе с тем, по результатам исследований [3], одним из основных «барьеров» доступа пользователей к информации находящейся в сети Интернет, является низкий ранг содержащих её сайтов в ИПМ. Возникает проблема, заключающаяся в поиске информации, рассредоточенной по множеству сайтов, среди больших объёмов шумовых данных.

Подходы к решению этой проблемы можно обобщённо назвать задачами поиска и кластеризации сведений. Несмотря на большое количество работ по данной тематике [11–13, 24, 26–35], отдельные аспекты организации такого поиска остаются малоисследованными. Это касается, в частности, влияния совместного появления группы термов (коллокаций) в тексте на его семантическую составляющую. Поэтому целью данной статьи является разработка и исследование математических методов формализации коллокаций в текстах.

Общая характеристика проблематики работы. Наиболее наивным подходом к поиску информации в текстовых коллекциях по праву считается булева модель [5, 13, 24]. В её рамках запрос представляет собой логическое выражение относительно предикатов, формализующих утверждения о появлении некоторого терма в документе. Причём документ в данном случае является аргументом предиката.

Введём некоторые обозначения для формализации рассматриваемой модели. Пусть D – множество документов, на котором решается задача информационного поиска. Представим D в виде: $D = \{d_1, d_2, \ldots, d_N\}$, $|D| = N$, где d_i– некоторый документ, $1 \le j \le N$. B – бинарное множество ($B = \{0, 1\}$), а S – множество всех термов, встречающихся в элементах D. Пусть, для определённости, $S = \{s_1, s_2, \ldots, s_n\}$ и $|S| = n$. Тогда, согласно булевой модели поиска текстовой информации, каждому терму $s \in S$ будет соответствовать некоторый предикат $P_s{:}D \to B$. Причём лингвистически $P_s(d)$ означает «терм s встречается в документе d». Другими словами предикат $P_s(d)$ принимает значение «1» тогда и только тогда, когда терм S встречается в документе d.

Так как любое логическое выражение приводится к дизъюнктивной нормальной форме, то поисковый запрос q может быть записан [13] в виде:

$$q = \bigwedge_{j=1}^{n} \bigvee_{i=1}^{m} P_{s_j}^{t}(d_i), \qquad (1)$$

где $t \in B$. Причём $P_{s_i}(d_i) = P_{s_i}^{0}(d_i) = \overline{P_{s_i}^{1}(d_i)}$, а «$\bar{}$» задаёт операцию отрицания.

Главные достоинства булевой модели следующие: простота понимания и реализации; высокая скорость информационного поиска, близкая к скорости интервального поиска идентичных объектов в базе данных [5]. Основным же недостатком данной модели является упрощённая математическая формализация текстового документа, которая представляет его в виде набора термов и не учитывает следующие факторы: частоту их встречаемости; совме-

стное появление; взаимное расположение; семантические связи между ними. Это приводит к невозможности ранжирования результатов поиска по уровню их соответствия информационным запросам и крайне низкому уровню пертинентности, то есть соответствия запроса информационным потребностям пользователя [6]. При этом достигается максимальная релевантность – соответствие результатов информационного поиска запросу [6]. Действительно, в коллекцию результатов информационного поиска попадают только документы, удовлетворяющие (1).

В дальнейшем булева модель была усовершенствована путём построения расширенной булевой модели [13], предполагающей вычисление и использование весовых коэффициентов для каждого терма. В соответствии с этой моделью, каждому терму ставится в соответствие его вес – некоторое значение из интервала [0, 1].

Формализация текстового документа на основе данной модели означает его представление в виде вектора, каждый элемент которого соответствует определённому терму и представляет его вес. Вместе с тем эффективность расширенной булевой модели сильно зависит от способа вычисления весов, ведь именно они, фактически, определяют значения ненулевых элементов вектора, поставленного в соответствие текстовому документу.

Векторно-пространственная модель текстовой коллекции. Дальнейшие исследования в области поиска текстовых сведений привели к созданию векторно-пространственной модели (ВПМ) текстовой коллекции, предложенной Солтоном в 1975 г. [13, 35]. В рамках этой модели документ формализуется вектором в евклидовом пространстве, где каждому терму $s \in S$, присутствующему хотя бы в одном из документов D, ставится в соответствие его весовое значение.

Запрос, формализующий информационные интересы пользователя, представляет собой вектор той же размерности. Каждая координата вектора-запроса определяет влияние того или иного терма на пертинентность документа. Оценка релевантности произвольного документа $d \in D$ осуществляется путём вычисления скалярного произведения векторов, формализующих запрос и d. Такой подход позволяет учесть важность каждого терма для информационной потребности пользователя и особенности конкретного документа, а также получить значение релевантности как аддитивной свёртки полученных оценок по каждому терму.

Рассмотрим ВПМ более подробно. Поставим в соответствие каждому терму $s_j \in S$ в документе $d_i \in D$ неотрицательный вес w_i^j. Таким образом, документ d_i будет формализован вектором $d_i\left(w_i^1, w_i^2, ..., w_i^n\right)$.

Рассмотрим произвольный запрос q, который, как уже было ранее отмечено, представляет собой вектор весовых коэффициентов, соответствующих каждому терму: $q\left(w_q^1, w_q^2, ..., w_q^n\right)$.

Тогда релевантность документа d_i определяется по формуле:

$$rel\left(q, d_i\right) = q \cdot d_i = \sum_{j=1}^{n} w_q^j w_i^j \ . \qquad (2)$$

Важнейшим фактором, определяющим эффективность ВПМ, как и в случае с усовершенствованной булевской моделью, является метод нахождения весовых коэффициентов термов [26, 33, 34]. Классический подход к решению данной задачи предполагает использование в качестве этих коэффициентов нормированных частот термов [33, 35].

Пусть K – некоторое отображение. $K : D \times S \rightarrow Z_+$, где Z_+ – множество целых неотрицательных чисел. Причём $K\left(d, s\right)$ – количество появлений терма s в документе d. Тогда, согласно классическому подходу:

$$w_i^j = K\left(d_i, s_j\right) / \max_{t=\overline{1,N}}\left(K\left(d_t, s_j\right)\right). \qquad (3)$$

Вес терма в документе, вычисленный согласно формуле (3), принято обозначать аббревиатурой *tf* (от англ. *term frequency* – частота терма) [13, 33, 35].

Вместе с тем вычисление веса терма в конкретном документе не учитывает среднюю частоту использования данного терма в коллекции документов, на которой осуществляется поиск (D).

Во-первых, есть термы, которые свойственны естественному языку и используются практически во всех текстовых документах – поэтому частота их появления не зависит от семантической составляющей текста. Примеры таких термов: предлоги («на», «в»), некоторые глаголы («быть», «увидеть»), и широко употребляемые существительные и прилагательные («данный», «свойство»).

Во-вторых, важность терма в документе может зависеть от вида множества D. Например, если это множество представляет собой коллекцию научно-технической литературы, то глагол «формализовать» никак не повлияет на выявление документов, являющихся научными статьями по тематике «информационный поиск». С другой стороны, если множество D представляет собой коллекцию различных текстовых документов по тематике «информационный поиск», то терм «формализовать» характерен для научно-технической литературы в коллекции и вполне может быть использован как критерий при поиске и кластеризации в D.

Для учёта данных свойств было введено понятие дискриминационной силы терма [13]. Так как при построении ВПМ доступна статистика появления термов в документах коллекции D, то хорошие показатели [13, 35] демонстрирует следующее правило вычисления веса:

$$w_i^j = \left(K(d_i, s_j) / \max_{t=\overline{1,N}} (K(d_t, s_j)) \right) \log_2 (N / n_j), \quad (4)$$

где N – общее число документов информационного массива, а n_j – количество документов, в которых встречается терм s_j. Логарифм, появившийся в формуле (4), получил название инверсная частота документа (*inverse document frequency*) или *idf*. Сама же матрица весов вида (4) стала называться *tf-idf*.

Легко видеть, что *idf* тем меньше, чем в большем числе документов встречается терм s_j. Например, если s_j появляется, во всех документах коллекции, то $w_i^j = 0$. С уменьшением числа документов включающих s_j, возрастает и $\log_2 (N / n_j)$, достигая своего максимума в точке $n_j = 1$.

Матрица *tf–idf* также используется для определения расстояния между документами и кластеризации текстовых коллекций.

Дальнейшее развитие ВПМ текстовой коллекции привело к уточнению формул для вычисления частоты терма и инверсной частоты документа. При этом само построение веса терма в документе по-прежнему представляет собой произведение *tf* и *idf*.

На сегодняшний день распространение получила формула *bm25f* [30, 32], которая отличается от (4) растяжением по координатным осям *idf* и уточнённым способом вычисления *tf*.

Латентно-семантический анализ текстовой коллекции. Использование ВПМ для представления текстовой коллекции легло в основу нескольких подходов к поиску и кластеризации текстовых сведений. Одним из таких важнейших подходов является латентно-семантический анализ, базирующийся на *svd*-разложении матрицы *tf–idf*. Рассмотрим этот подход подробнее.

Согласно представленной выше ВПМ текстовой коллекции D ставится в соответствие некоторая матрица *tf–idf*. Обозначим эту матрицу как $W_{n \times N}$. В ней каждая строка соот-

ветствует некоторому документу *d*, столбец – терму *s*, а элемент данных строки и столбца – w_i^j, рассчитывается по формуле (4).

Рассмотрим сингулярное разложение [31] матрицы $W_{n\times N}$, то есть разложение вида:

$$W_{n\times N} = U_{n\times n} \times \Sigma_{n\times N} \times V_{N\times N}^T, \quad (5)$$

где $U_{n\times n}$ и $V_{N\times N}$ – ортогональные матрицы, а $\Sigma_{n\times N}$ – диагональная матрица с неотрицательными вещественными числами (σ_{ii}) на диагонали. Иными словами верно, $(\forall i = \overline{1,n})(\forall j = \overline{1,N})(i \neq j \Rightarrow \sigma_{ij} = 0) \wedge (\sigma_{ii} \geq 0)$.

Известно [4, 31], что для любой матрицы существует разложение вида (5) и оно обладает таким свойством – если в матрице $\Sigma_{n\times N}$ оставить только *k* наибольших сингулярных значений (обозначим такую матрицу как $\Sigma_{k\times k}^k$), а в матрицах $U_{n\times n}$ и $V_{N\times N}$ только соответствующие этим значениям колонки (соответственно, матрицы $U_{n\times n}^k$ и $V_{N\times N}^k$), то матрица:

$$W_{n\times N}^k = U_{k\times n}^k \times \Sigma_{k\times k}^k \times \left(V_{k\times N}^k\right)^T \quad (6)$$

будет наилучшей по Фробениусу аппроксимацией исходной матрицы $W_{n\times N}$ с рангом, не превышающим *k* [31]. Обозначим элементы матрицы $W_{n\times N}^k$ как w_i^{jk}.

Это свойство можно переформулировать следующим образом: $W_{n\times N}^k$ будет именно той матрицей ранга *k*, которая минимизирует норму Фробениуса матрицы $\left\|W_{n\times N} - W_{n\times N}^k\right\|_F$, где данная норма ($\|\ \|_F$) определяется как:

$$\left\|W_{n\times N} - W_{n\times N}^k\right\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{N}\left(w_i^j - w_i^{jk}\right)}. \quad (7)$$

Таким образом, верно правило «Чем меньше сингулярное число, соответствующее некоторому признаку (в нашем случае терму), тем менее он важен». Это правило позволяет выявлять наиболее значимые с семантической точки зрения термы и осуществлять кластеризацию, опираясь на них.

В основе латентно-семантического анализа лежит идея о том, что термы некоторым образом семантически связаны между собой. Следовательно, существует корреляция [29] между соответствующими им коэффициентами в матрице $W_{n\times N}$. Это означает, что характеристики одних термов (w_i^j) зависят от характеристик других. Тогда, выбрав термы соответствующие *k*-ому набору наибольших сингулярных значений, мы получим (согласно (6) и (7)) матрицу максимально приближенную к исходной. Поэтому, если при выборе некоторого числа *k* наибольших значений σ_{ii}, норма Фробениуса даст малое число относительно элементов матрицы *tf–idf*, это будет означать, что имея лишь характеристики указанных *k* термов, можно с большой достоверностью восстановить веса остальных.

Исследования подтвердили данное предположение. Поэтому латентно-семантический анализ стал основой многих эффективных алгоритмов поиска и кластеризации текстовой информации [11, 13, 28].

Более того, латентно-семантический анализ преодолевает проблемы синонимии и омонимии, связанные с неоднозначностью естественных языков.

Проблема синонимии возникает, если одинаковые понятия задаются разными термами – синонимами. К примеру, пары термов «бегемот» и «гиппопотам», «солнце» и «звезда», «бежать» и «лететь» в некоторых контекстах означают одно и тоже, а в других их смысл

различен. Поэтому использованием обычного словаря синонимов данную проблему решить не представляется возможным.

Омонимия – также явление естественного языка, заключающееся в том, что один и тот же терм несёт разную семантическую нагрузку. Например, термы «лук», «коса», «замок» могут иметь разный смысл в зависимости от контекста.

Важнейшим достоинством латентно-семантического анализа является его способность к выявлению латентных (скрытых, неочевидных) семантических зависимостей между термами и отсутствие необходимости предварительного обучения или же выбора числа кластеров [13]. А основной недостаток – высокая асимптотическая сложность алгоритма [11].

Понятие нечёткой коллокации термов. Одним из путей развития алгоритмов поиска и кластеризации текстовых коллекций, стали подходы к их формализации с учётом коллокаций.

Коллокация, как лексико-фразеологически обусловленная сочетаемость термов, была известна лингвистам ещё с середины двадцатого века [27]. Вместе с тем, в рамках компьютерной лингвистики коллокации начали изучаться сравнительно недавно. Они исследованы в работах Недошивиной [14], Бишта [25], Пивоваровой [12, 16], Ягуновой [22]. Например, в работе «Учёт синтаксических связей при поиске коллокаций» Е.В. Недошивиной [14] коллокация определяется как «последовательность термов, частота совместного появления которых не соответствует ожидаемой на основе закона случайного их распределения».

Однако во всех указанных работах ([12, 14, 16, 22, 25]) под коллокациями понимается появление некоторого набора термов, находящихся непосредственно рядом друг с другом.

Вместе с тем есть основания полагать, что на семантику текстового документа влияют наборы значимых термов, появляющиеся в одном абзаце или одном предложении. Действительно, наиболее известные ИПМ, такие, например, как Яндекс или Гугл, предлагают своим пользователям формализованный язык запросов, позволяющий формулировать их задавая появление термов в искомом текстовом документе на определённом расстоянии друг от друга [15, 23]. При этом наличие знаков препинания между термами не учитывается. Похожие возможности есть и в некоторых Российских «юридических» информационно-поисковых системах.

Под расстоянием между термами s_1 и s_2 здесь и далее будем понимать количество других термов, появившихся между s_1 и s_2 в текстовом документе.

Возникают такие вопросы: каким образом следует задать коллокацию, если термы в ней могут находиться на различном расстоянии друг от друга; является ли появление двух термов на определённом расстоянии друг от друга коллокацией или же случайным событием.

В нашей работе «Кластеризация текстовых коллекций на основе нечеткого описания коллокаций» [18] была предложена модель нечёткой коллокации как пары термов и функции, задающей расстояние между ними. В дальнейшем [8, 9] эта модель была усовершенствована путём учёта частот коллокаций и обобщения понятия коллокация до объекта, состоящего из произвольного числа термов. Множество работ авторского коллектива посвящено вопросам информационного поиска [7, 19], кластеризации текстовых коллекций [17] и построению пертинентных [6] запросов для поиска на основе нечётких коллокаций [7, 20].

В данной работе предлагается метод формализации нечётких коллокаций в текстовых коллекциях на основе лингвистической переменной, а также рассматриваются подходы к оценке значимости данных коллокаций для семантической составляющей документа и выявлению важных для коллекции коллокаций.

Формализация нечётких коллокаций на основе лингвистической переменной. Введём формализованное понятие коллокации. Здесь и далее для простоты ограничимся рассмотрением только коллокаций, состоящих из двух термов.

Определение 1. Кортеж термов:

$$\langle s_1,\ s_2 \rangle, \tag{8}$$

где $s_1,\ s_2 \in S$, будем называть коллокацией. Кортеж (8) задаёт термы, составляющие коллокацию, а также порядок их появления.

Рассмотрим в качестве примера коллокацию $\langle s_1,\ s_2 \rangle$, где s_1 задаёт терм «кредит», а s_2 – «дебет». Этой коллокацией задаются документы, в которых терм «кредит» встречается перед термом «дебет». Очевидно, что при работе с коллокацией, представленной в виде (8), расстояние между термами не рассматривается.

Для того чтобы учесть в модели коллокации расстояние между термами введём нечёткость. Рассмотрим лингвистическую переменную *distance* [10].

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

где d = «дистанция между термами в коллокации» – имя лингвистической переменной *distance;* T = {«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»} и модификаторов *Mod* {«не», «очень»}.

Пусть $op \in Op$, а t_1 и $t_2 \in T$. Тогда G на основе данных элементов будет иметь вид $t_1\ op\ t_2$. Например, пусть t_1 = «большая», t_2 = «средняя», а op = «или», тогда $t_1\ op\ t_2$ = «дистанция между термами большая или средняя».

Рассмотрим произвольный элемент $m \in Mod$. Семантическое правило, для произвольного терма $t \in T$ имеет вид: $m\ t$. Например, при t = «большая», а m = «не», $m\ t$ 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 $\mu: Z_+ \rightarrow [0,\ 1]$. 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.

Отметим, что понятия «дистанция» и «расстояние» между термами в коллокации не являются синонимичными. Термин «расстояние» был использован для обозначения конкретного числа слов, появившихся в документе между термами, составляющими коллокацию. Поэтому расстояние между термами в коллокации может принимать только целые неотрицательные значения. С другой стороны, термин «дистанция» появился в рамках определения лингвистической переменной и непосредственно связан с её значениями. Поэтому в дальнейшем под дистанцией между термами в коллокации будем понимать значение лингвистической переменной *distance*. Например, корректно сказать: «дистанция между термами в коллокации маленькая».

Определение 2. Кортеж:

$$\langle s_1,\ s_2,\ distance \rangle, \tag{9}$$

где $s_1,\ s_2 \in S$, назовём нечёткой коллокацией.

Может показаться, что нечёткая коллокация ограничена фиксацией порядка термов, однако параллельно с рассмотрением $\langle s_1,\ s_2,\ distance \rangle$, исследуется и $\langle s_2,\ s_1,\ distance \rangle$, что

даёт возможность учесть все возможные комбинации термов, а связка «и» лингвистической переменной *distance* позволит работать с объединённой коллокацией.

Рассмотрим функцию $\tilde{\mu}: R \to [0, 1]$, такую что $(\forall k \in Z_+)(\tilde{\mu}(k) = \mu(k))$.

В дальнейшем, будем рассматривать в качестве функции, формализующей семантическое правило *M*, непрерывную функцию принадлежности $\tilde{\mu}$. Это не отразится на результате вычислений, так как все преобразования, задаваемые *G* и *M*, над непрерывными функциями эквивалентны преобразованиям над значениями этих функций в каждой точке Z_+. Данный переход осуществлён исключительно для простоты обработки непрерывных функций в информационных системах, так как позволяет хранить и обрабатывать уравнение функции, вместо таблицы значений.

Пусть, согласно семантическому правилу *M*, таким значениям лингвистической переменной *distance* как «маленькая», «средняя» и «большая» соответствуют функции $\tilde{\mu}_м$, $\tilde{\mu}_с$ и $\tilde{\mu}_б$.

Функция $\tilde{\mu}_м$ задаёт коллокацию в классическом её понимании, а именно, пара термов, расположенная близко друг к другу. Примем $\tilde{\mu}_м(0) = 1$, так как, если термы находятся непосредственно рядом друг с другом, то верно, что дистанция между термами в коллокации – «маленькая». При достижении некоторой величины $R_l \in Z_+$ верно, что $\tilde{\mu}_м(R_l) = 0$. Очевидно, что дальнейшее увеличение расстояния не изменит значение рассматриваемой функции принадлежности. Сформулируем данное утверждение на языке эпсилон-дельта: для $\tilde{\mu}_м$ верно, что $(\forall x \in R \mid x > R_l)(\tilde{\mu}_м(x) = 0)$.

С другой стороны $\tilde{\mu}_м$ не обязана принимать значение равное «1» только в точке «0». Тогда пусть $L_l \in Z_+$ станет левой границей, при которой $\tilde{\mu}_м$ всё ещё равна «1». Осталось определить поведение рассматриваемой функции принадлежности в интервале $(L_l,\ R_l)$. Исходя из семантики $\tilde{\mu}_м$, можно достоверно утверждать лишь, что на данном участке она не возрастает. Классическим подходом к формализации такой функции при условии отсутствия каких-либо данных о её форме является прямая [10]. Тогда $\tilde{\mu}_м$ принимает вид:

$$\tilde{\mu}_м = \max\{0, \min\{1, (R_l - x)/(R_l - L_l)\}\}. \qquad (10)$$

График функции $\tilde{\mu}_м$, заданной выражением (10) представлен на рисунке 1.

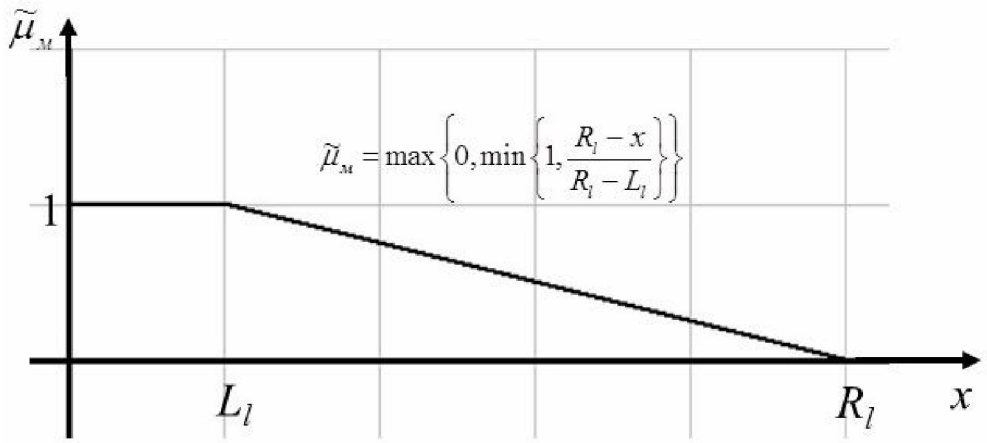

Рис. 1. График функции $\tilde{\mu}_м$

Заметим, что в реальных условиях функция принадлежности нечёткой коллокации не всегда будет задаваться линейным сплайном [1], так как могут быть построены новые значения лингвистической переменной на основе связок и модификаторов. Например, если применить модификатор «очень» и построить функцию принадлежности, формализующую значение лингвистической переменной «очень маленькая», то при классической [10] семантической формализации модификатора «очень» с помощью возведения $\tilde{\mu}_м$ в квадрат, получим функцию принадлежности, вид и график которой представлен на рисунке 2.

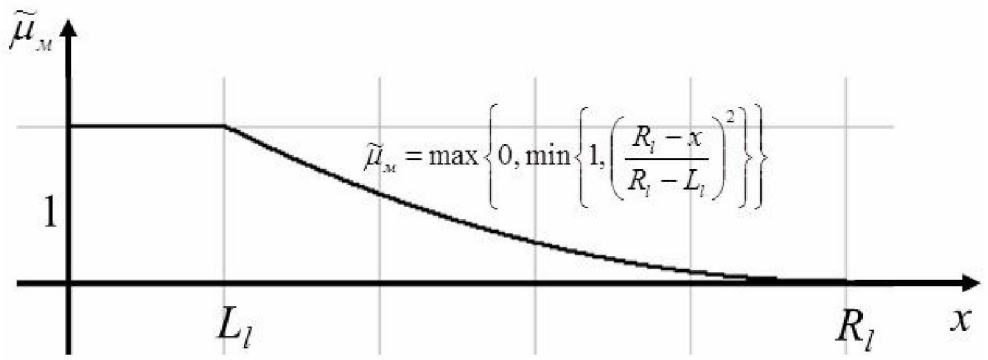

Рис. 2. График функции принадлежности, отражающей семантический смысл значения лингвистической переменной «очень маленькая»

Аналогично зададим функцию принадлежности $\tilde{\mu}_б$, формализующую значение лингвистической переменной *distance* «большая». Для этого зададим L_r, $R_r \in Z_+$, такие что для $\tilde{\mu}_б$ верно $(\forall x \in R \mid x > R_r)(\tilde{\mu}_б(x) = 1) \wedge (\forall x \in R \mid x < L_r)(\tilde{\mu}_б(x) = 0)$ и аппроксимируем значение функции в интервале $(L_r,\ R_r)$ с помощью прямой. Тогда функция $\tilde{\mu}_б$ будет иметь вид:

$$\tilde{\mu}_б = \max\{0, \min\{1, (x - L_r)/(R_r - L_r)\}\}. \quad (11)$$

График функции $\tilde{\mu}_б$, заданной выражением (11), представлен на рисунке 3.

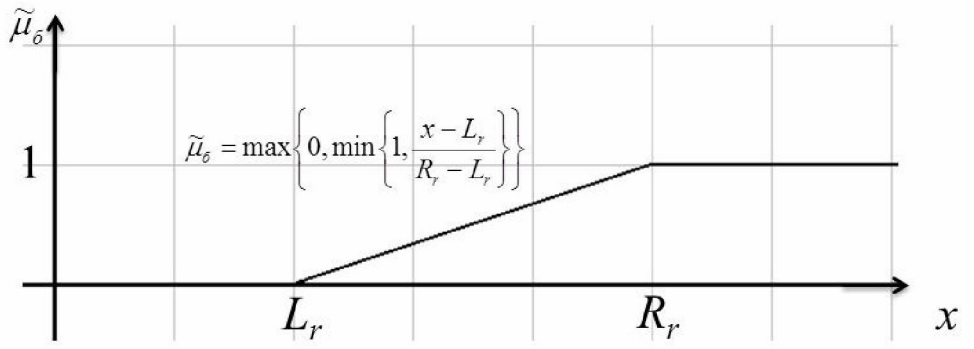

Рис. 3. График функции $\tilde{\mu}_б$

Рассмотрим функцию $\tilde{\mu}_c$, определяющую семантику значения «средняя» лингвистической переменной *distance*. Легко видеть, что конкретное значение расстояния между термами в коллокации принадлежит понятию «средняя» тем больше, чем менее мы уверены в

принадлежности данного понятия к таким значениям лингвистической переменной как «большая» или «маленькая». То есть семантика значения «средняя» однозначно определяется через значения «большая» и «маленькая» на основе модификаторов и связок множества G. А именно – логично предположить, что «средняя» = «не (маленькая или большая)».

Таким образом, чтобы определить функцию принадлежности $\tilde{\mu}_c$, необходимо, для начала, задать преобразования, осуществляемые модификаторами и связками.

Ранее было предложено использовать в качестве модификатора «очень» классическую функцию возведения в квадрат. Вместе с тем, аппроксимация интервалов $(L_l,\ R_l)$ и $(L_r,\ R_r)$ функций $\tilde{\mu}_м$ и $\tilde{\mu}_б$ с помощью прямых привела к внесению излишней робастности в модель. Поэтому, с целью сделать модель более гибкой, в качестве модификатора «очень», возьмём операцию возведения в степень γ, где $\gamma \in R, \gamma > 1$.

Модификатор «не» и связки «и» и «или» в теории нечётких множеств являются базовыми и задаются с помощью операции отрицания ($n(.)$) и T,S-норм ($T(.,.)$, $S(.,.)$) соответственно [2, 10].

Теория нечётких множеств допускает использование в качестве T,S-норм различных операций, получаемых с помощью генераторов, представляющих собой классы параметризированных функций двух переменных. На сегодняшний день известны [2] генераторы Домби, Франка, Хамахера, Швайцера-Скляра, Ягера, Майора-Торренса.

Выбор функции, формализующей операцию отрицания, целесообразно производить на основе важнейшего тождества теории нечётких множеств: $T(x,y) = n(S(n(x),n(y)))$. Очевидно, что, при условии известных T,S-норм, функция отрицания определяется единственным образом.

Для исследования коллокаций вида (9) в текстовых документах необходимо выбрать следующее: границы функций принадлежности $L_l,\ R_l,\ L_r,\ R_r \in Z_+$; генератор T,S-норм; значения параметров $\lambda, \lambda \in R, \lambda > 0$ и $\gamma, \gamma \in R, \gamma > 1$.

Выбор этих параметров целесообразно осуществлять по критерию адекватности разработанной модели, проверка которой возможна лишь при проведении вычислительных экспериментов.

Построение ВПМ текстовой коллекции с учётом нечётких коллокаций. Рассмотрим процесс построения матрицы *tf-idf* с позиций теории множеств. Поставим каждому терму $s_j \in S$ в соответствие множество $H_j = \{ s_j, s_j^{(1)}, \ldots, s_j^m \}$, где $s_j^{(k)}, k = \overline{1,m}$ – различные словоформы терма s_j. На практике, при построении таблицы *tf–idf* все словоформы s_j учитываются как один и тот же терм. Поэтому перед подсчётом частот термов в текстовых документах производится лемманизация – приведение всех термов в документе к единой словоформе [13]. Для неё используется специализированное программное обеспечение, осуществляющее морфологический анализ [21].

Пусть $\mu_j : S \to B$ – характеристическая функция множества H_j, то есть:

$$\mu_j(s) = \begin{cases} 0, s \notin H_j, \\ 1, s \in H_j; \end{cases} . \tag{12}$$

Тогда $K(d,s)$ – количество появлений терма $s \in S$ в документе $d \in D$ вычисляется по формуле:

$$K(d_i, s_j) = \sum_{s \in d_i} \mu_j(s). \tag{13}$$

В силу (13) формула (3) принимает вид:

$$w_i^j = \sum_{s\in d_i} \mu_j(s) / \max_{t=\overline{1,N}} \left(\sum_{s\in d_t} \mu_j(s) \right). \quad (14)$$

Найдём теперь n_j – количество документов, в которых встречается терм s_j. Возьмём и зафиксируем произвольный терм s_j. Пусть P_j – некоторое отображение, $P_j : D \to B$ и верно, что, $P_j(d) = 1$, если $s_j \in D$ и $P_j(d) = 0$, если $s_j \notin D$. Тогда, легко видеть, что:

$$P_j(d) = \bigvee_{s\in d} (\mu_j(s)), \quad (15)$$

где $\vee$ – операция дизъюнкции. Действительно, если хотя бы один терм d совпадает с s_j, то результирующая дизъюнкция даёт «1». В противном случае – если в документе d терм s_j не появляется, то $P_j(d) = 0$.

Важно отметить, что выражение (15) корректно и с семантической точки зрения. Так, если представить документ d в виде кортежа термов: $d(s_{i_1}, s_{i_1}, ..., s_{i_k})$, где k – количество термов в d, то P_j будет предикатом вида: «терм $s_{i_1} \in H_j$ или $s_{i_2} \in H_j$ или … или $s_{i_k} \in H_j$». Истинность или ложность принадлежности определяется характеристической функцией μ_j, а связка «или» формализуется при помощи дизъюнкции.

Согласно определению P_j и n_j легко видеть, что $n_j = \sum_{d\in D} P_j(d)$ или в силу (15) получаем:

$$n_j = \sum_{d\in D} \bigvee_{s\in d} (\mu_j(s)). \quad (16)$$

Формулы (13), (14) и (16) позволяют выразить (4) через характеристические функции семейства множеств H_j.

Проведём аналогичные рассуждения для нечётких коллокаций, заданных в виде (9). Поставим каждой нечёткой коллокации $\langle s_i,\ s_j,\ distance \rangle$ в соответствие нечёткое множество $\widetilde{H}_{ij}^{distance}$, характеризующееся функцией принадлежности $\mu_{ij}^{distance} : S^2 \times Z_+ \to [0,1]$. $\mu_{ij}^{distance}$ конструируется с помощью семантического правила M, лингвистической переменной *distance*, на основе её значения. Множество $\widetilde{H}_{ij}^{distance}$ является нечётким аналогом H_j для нечеткой коллокации, а соответствующая ему функция принадлежности – обобщением характеристической функции μ_j.

Тогда (13) для учёта количества появления коллокаций в тексте принимает вид:

$$K(d_i, \langle s_r, s_t, distance \rangle) = \sum_{s_r, s_t \in d_i} \mu_{sr}^{distance}(s_r, s_t, k_i), \quad (17)$$

где $w_{ij}^{distance}$ весовой коэффициент *tf–idf*, соответствующий коллокации $\langle s_i,\ s_j,\ distance \rangle$, а k_i – расстояние появившееся в документе d_i между выбранными термами s_r, s_t.

А формула (16), в силу того, что операция дизъюнкции при переходе к нечёткости обобщается S-нормой, принимает вид:

$$n_{ij}^{distance} = \sum_{d\in D} S_{s_r, s_t \in d} (\mu_{ij}^{distance}(s_r, s_t, k)), \quad (18)$$

где «S» – S-норма, формализующая в нечёткой логике связку «или».

Равенства (17) и (18) показывают, что для нечёткой коллокации, заданной в виде (9), можно вычислить весовые коэффициенты матрицы *tf–idf* согласно (4).

Теория множеств является частным случаем теории нечётких множеств и имеет место в случае, если функция принадлежности μ принимает только два значения: «0» или «1». Легко видеть, что (17) и (18), задающие на основе (4) правило вычисления весового коэффициента коллокации в документе, при дискретных значениях функции принадлежности сводятся к (13) и (15) соответственно. То есть представленная расширенная ВПМ с учётом нечётких коллокаций является более общим случаем классической ВПМ, что косвенно свидетельствует об адекватности предложенного расширения.

Заключительные замечания. Расширенная матрица *tf–idf*, состоящая из весов, вычисляемых по формуле (4) с использованием (13) и (15) для термов, а также (17) и (18) для коллокаций, позволяет оценить значимость нечётких коллокаций в сравнении с термами. Для этого достаточно провести *svd*-разложение данной матрицы, и получить, согласно (5), диагональную матрицу Σ. Рассмотрим для каждого терма и коллокации соответствующий им диагональный элемент матрицы Σ. В силу (5)–(7) эти элементы отражают относительную важность признака вне зависимости от того терм это или коллокация.

Важнейшими достоинствами предложенной модели формализации нечётких коллокаций являются их учёт в матрице *tf–idf* и возможность провести сравнительный анализ значимости частот коллокаций и термов для текстовых документов. Кроме того, немаловажным достоинством является и то, что все коллокации содержат в себе значения лингвистической переменной. Поэтому они удобны для интерпретации на естественном языке и использования при составлении поисковых запросов.

К недостаткам предложенной модели можно отнести её некоторую робастность, которая не позволяет учесть коллокации с функциями принадлежности не соответствующими ни одному из возможных значений лингвистической переменной.

В дальнейших исследованиях предполагается выполнить постановки задач, планирование и проведение вычислительных экспериментов, позволяющих осуществить сравнительный анализ значимости коллокаций и термов для различных выборок текстовых документов.

Планируется также выбрать такие параметры предложенной модели: $L_l,\ R_l,\ L_r,\ R_r \in Z_+$; используемый генератор T,S-норм; ; $\lambda, \lambda \in R, \lambda > 0$ и $\gamma, \gamma \in R, \gamma > 1$. Выбор параметров необходимо проводить по критерию максимизации значимости нечётких коллокаций в сравнении с термами.

Список литературы

1. Алберг Дж. Теория сплайнов и её приложения / Дж. Алберг, Э. Нильсон, Дж. Уолш. – Москва : Мир, 1972. – 320 с.

2. Батыршин И. З. Основные операции нечёткой логики их обобщения / И. З. Батыршин. – Казань : Отечество, 2001. – 100 с.

3. Брумштейн Ю. М. Системный анализ вопросов, связанных с востребованностью информации на web-сайтах / Ю. М. Брумштейн, Е. Ю. Васьковский // Прикаспийский журнал: управление и высокие технологии. – 2015. – № 1 (29). – С. 59–74.

4. Гантмахер Ф. Р. Теория матриц / Ф. Р. Гантмахер.– Москва : Наука, 1996. – 576 с.

5. Гасанов Э. Э. Теория хранения и поиска информации / Э. Э. Гасанов, В. Б. Кудрявцев. – Москва : ФИЗМАТЛИТ, 2002. – 288 с.

6. ГОСТ 7.73-96. Система стандартов по информации, библиотечному и издательскому делу. Поиск и распространение информации. Термины и определения. – Взамен ГОСТ 7.27-80; введен 1998–01–01. – Минск : Издательство стандартов, 1997. – 20 с.

7. Громов Ю. Ю. Нечеткий подход к определению пертинентности результатов поиска и выбору оптимального запроса / Ю. Ю. Громов и другие // Вестник Воронежского института ФСИН России. – 2011. – № 2. – С. 49–55.

8. Громов Ю. Ю. Построение многомерных функций принадлежности / Ю. Ю. Громов и другие // Приборы и системы. Управление, контроль, диагностика. – 2012. – № 11. – С. 21–26.

9. Громов Ю. Ю. Формализация текстовой коллекции на основе нечетких частот коллокаций / Ю. Ю. Громов, Д. В. Поляков, Т. О. Авдеева // Приборы и системы. Управление, контроль, диагностика. – 2013. – № 2. – С. 15–17.

10. Заде Л. Понятие лингвистической переменной и её применение к принятию приближённых решений / Л. Заде. – Москва : МИР, 1973. – 167 с.

11. Кириченко К. М. Обзор методов кластеризации текстовой информации / К. М. Кириченко, М. Б. Герасимов. – Режим доступа: http://www.dialog-21.ru/Archive/ 2001/volume2/2_26.htm, свободный. – Заглавие с экрана. – Яз. рус.

12. Киселев М. В. Метод кластеризации текстов, учитывающий совместную встречаемость ключевых терминов, и его применение к анализу тематической структуры новостного потока, а также ее динамики / М. В. Киселев, В. С. Пивоваров, М. М. Шмулевич. – Компания Megaputer Intelligence, 2005. – 24 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. Недошивина Е. В. Учёт синтаксических связей при поиске коллокаций / Е. В. Недошивина // Natural Language Processing. – 2008. – С. 1–3.

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

16. 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.

17. 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 МВД России, 2012. – С. 164.

18. 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.

19. Поляков Д. В. Определение пертинентности результатов запроса с использованием нечеткой логики / Д. В. Поляков и другие // Приборы и системы. Управление, контроль, диагностика. – 2012. – № 3. – С. 29–33.

20. Поляков Д. В. Построение пертинентного запроса к информационно-поисковой машине на основе математического аппарата нечеткой логики / Д. В. Поляков и другие // Математические методы и информационно-технические средства : труды VIII Всероссийской научно-практической конференции. – Краснодар : Краснодарский университет МВД России, 2012. – С. 167.

21. Пруцков А.В. Методы морфологической обработке текстов / А. В. Пруцков, А. К. Розанов // Прикаспийский журнал: управление и высокие технологии. – 2014. – № 3 (27). – С. 119–133.

22. Ягунова Е. В. От коллокаций к конструкциям / Е. В. Ягунова, Л. М. Пивоварова // Русский язык: конструкционные и лексико-семантические подходы / отв. ред. С. С. Сай. – Санкт-Петербург : Труды Института лингвистических исследований Российской академии наук, 2011. – 43 с.

23. Язык запросов Яндекса. – Режим доступа: https://yandex.ru/support/search/query-language/qlanguage.xml, свободный. – Заглавие с экрана. – Яз. рус.

24. Baeza-Yates R. Современный информационный поиск / R. Baeza-Yates, B. Ribeiro-Neto. – Нью-Йорк : ACM Press Series ; AddisonWesley, 1999. – 513 с.

25. Bisht R. K. Подход к выделению коллокаций на основе нечётких множеств / R. K. Bisht, H. S. Dhami // International Journal of Computer Applications. – 2010. – Vol. 5, № 3. – С. 43–49.

26. Egghe L. Соотношение между коэффициентом корреляции Пирсона и косинусной мерой Солтона / L. Egghe, L. Leydesdorff // Journal of the American Society for Information Science & Technology (forthcoming). – 2009. – Vol. 60, № 2. – С. 232–239.

27. Firth J. R. Лингвистическая теория / J. R. Firth // Studies in Linguistic Analysis. – Oxford : Philological Society, 1968. – С. 1–32.

28. Hofmann T. Вероятностное латентное семантическое индексирование / T. Hofmann // Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in en:Information Retrieval. – 1999. – С. 50–57.

29. Pelleg D. Расширение алгоритма *k*-средних за счёт эффективной оценки числа кластеров / D. Pelleg, A. Moore. – Pittsburgh : Carnegie Mellon University, 2000. – С. 1–8.

30. Perez-Aguera J. R. Использование bm25f для семантического поиска / J. R. Pérez-Agüera, J. Arroyo, J. Greenberg, J. P. Iglesias, V. Fresno // Proceedings of the 3rd International Semantic Search Workshop. – 2010. – С. 1–10.

31. Perez-Iglesias J. Использование bm25f для патентного поиска / J. Perez-Iglesias, A. Rodrigo, V. Fresno. – Режим доступа: http://ceur-ws.org/Vol-1176/CLEF2010wn-CLEF-IP-PerezEt2010.pdf, свободный. – Заглавие с экрана. – Яз. рус.

32. Press H. W Численные методы в С. Искусство научных вычислений / H. W Press и другие. – 2 изд. – Cambridge : Cambridge University Press, 1997. – 994 с.

33. Salton G. Автоматический информационный поиск / G. Salton. – Ithaca : Cornell University, 1980. – С. 41–54.

34. Salton G. Векторно-пространственная модель для автоматического индексирования / G. Salton, A. Wong, C. Yang // Communications of the ACM. – 1975. – С. 613–620.

35. Salton G. Выборочный обход текста / G. Salton, A. Singhal. – Ithaca : Department of Computer Science, Cornell University, 1995. – С. 131–144.

References

1. Alberg J., Nilson J., Uolsh J. *Teoriya splaynov i ee prilozheniya* [Theory of spline and its application], Moscow, MIR Publ., 1972. 320 p.

2. Batyrshin I. Z. *Osnovnye operatsii nechetkoy logiki i ikh obobshcheniya* [Base operations of fuzzy logic and their generalization], Kazan, Otechestvo Publ., 2001. 100 p.

3. Brumshteyn Yu. M., Vaskovskiy Ye. Yu. *Sistemnyy analiz voprosov, svyazannykh s vostrebovannostyu informatsii na web-saytakh* [The system analysis of questions, connected with information demand of the web-sites]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Management and High Technologies], 2015, no. 1 (29), pp. 59–74.

4. Gantmaher F. R. *Teoriya matrits* [Theory of Matrices], Moscow, Nauka Publ., 1996. 576 p.

5. Gasanov E. E., Kudryavtsev V. B. *Teoriya khraneniya i poiska informatsii* [Theory of informational storage and retrieval], Moscow, FIZMATLIT Publ., 2002. 288 p.

6. GOST 7.73-96. System of standards on information, librarianship and publishing. Search and dissemination of information. Terms and Definitions. Instead of GOST 7.27-80, introduced 1998–01–01. Minsk, Izdatelstvo standartov Publ., 1997. 20 p.

7. Gromov Yu. Yu., et. al. *Nechetkiy podkhod k opredeleniyu pertinentnosti rezultatov poiska i vyboru optimalnogo zaprosa* [Fuzzy approach to calculation of the pertinence of search results and the choice of optimal query] . *Vestnik Voronezhskogo instituta FSIN Rossii* [Bulletin of the Voronezh Institute of Russian Federal Penitentiary Service], 2011, no. 2, pp. 49–55.

8. Gromov Yu. Yu., et. al. *Postroenie mnogomernykh funktsiy prinadlezhnosti* [Creation of the multidimensional fuzzy functions]. *Pribory i sistemy. Upravlenie, kontrol, diagnostika* [Instruments and Systems. Management, Monitoring, Diagnostics], 2012, no. 11, pp. 21–26.

9. 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.

10. Zade L. *Ponyatie lingvisticheskoy peremennoy i ee primenenie k prinyatiyu priblizhennykh resheniy* [The concept of linguistic variable and approach to the adoption of approximate solutions], Moscow, MIR Publ., 1973. 167 p.

11. Kirichenko K. M. *Obzor metodov klasterizatsii tekstovoy informatsii* [Overview of clustering methods of textual information]. Available at: http://www.dialog-21.ru/Archive/2001/volume2/2_26.htm.

12. Kiselev M. V., Pivovarov V. S., Shmulevich M. M. *Metod klasterizatsii tekstov, uchityvayushchiy sovmestnuyu vstrechaemost klyuchevykh terminov, i ego primenenie k analizu tematicheskoy struk-*

tury 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.

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, LI-BROKOM Publ., 2009. 264 p.

14. Nedoshivina Ye. V. Uchet sintaksicheskikh svyazey pri poiske kollokatsiy [Accounting syntactic links when searching collocations]. *Natural Language Processing*. 2008, pp. 1–3.

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

16. Pivovarova L. M., Yagunova Ye. 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.

17. Polyakov D. V., et al. *K voprosu postroeniya matematicheskoy modeli klasterizatsii tekstovykh svedeniy* [The problem of constructing a mathematical model for clustering text information]. *Matematicheskie metody i informatsionno-tekhnicheskie sredstva : trudy VIII Vserossiyskoy nauchno-prakticheskoy konferentsii* [Mathematical methods and information technology equipment: Proceedings of VIII scientific-practical conference], Krasnodar, Krasnodar University of the Ministry of Internal Affairs of Russia Publ. House, 2012, pp. 164.

18. Polyakov D. V., Ivanova O. G., Gromova A. Yu., Didrikh V. Ye. *Klasterizatsiyka tekstovyh kollektsiy na osnove nechetkogo opisaniya kollokatsiy* [Clustering of text collections based on fuzzy collocations]. *Informatsiya i bezopasnost* [Information and security], 2011, no. 3, pp. 459–462.

19. Polyakov D. V., et al. *Opredelenie pertinentnosti rezultatov zaprosa s ispolzovaniem nechetkoy logiki* [Determination pertinence of query results using fuzzy logic]. *Pribory i sistemy. Upravlenie, kontrol, diagnostika* [Instruments and Systems. Management, Monitoring, Diagnostics], 2012, no. 3, pp. 29–33.

20. Polyakov D. V., et al. *Postroenie pertinentnogo zaprosa k informactsionno-poiskovoy mashine na osnove matematicheskogo apparata nechetkoy logiki* [Creating a pertinent request to a search engine based on the mathematical apparatus of fuzzy logic]. *Matematicheskie metody i informactsionno-tehnicheskie sredstva : trudy VIII Vserossiyskoy nauchno-prakticheskoy konferentsii* [Mathematical Methods and Information Technology Equipment. Proceedings of VIII All-Russian Scientific and Practical Conference], Krasnodar, Krasnodar University of the Ministry of Internal Affairs of Russia Publ. House, 2012, pp. 167.

21. Prutskov A. V., Rozanov A. K. *Metody morfologicheskoy obrabotki tekstov* [Ways of natural language morphological processing]. *Prikaspiyskiy zhurnal: upravlenie i vysokie tekhnologii* [Caspian Journal: Management and High Technologies], 2014, no. 3 (27), pp. 119–133.

22. Yagunova Ye. V., Pivovarova L. M. *Ot kollokatsiy k konstruktsijyam* [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.

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

24. Baeza-Yates R., Ribeiro-Neto B. *Sovremennyy informatsionnyy poisk* [Modern Information Retrieval], New York, ACM Press Series, AddisonWesley Publ., 1999. 513 p.

25. Bisht R. K., Dhami H. S. Podhod k vydeleniyu kollokatsiy na osnove nechetkikh mnozhestv [Fuzzy Set Theoretic Approach To Collocation Extraction]. *International Journal of Computer Applications*, 2010, vol. 5, no. 3, pp. 43–49.

26. Egghe L., Leydesdorff L. Sootnoshenie mezhdu koeffitsientom korrelyatsii Pirsona i kosinusnoy meroy Soltona [The relation between Pearson's correlation coefficient r and Salton's cosine measure]. *Journal of the American Society for Information Science & Technology (forthcoming)*, 2009, vol. 60, no. 2, pp. 232–239.

27. Firth J. R. *Lingvisticheskaya teoriya* [A synopsis of linguistic theory]. *Studies in Linguistic Analysis*, Oxford, Philological Society Publ., 1968, pp. 1–32.

28. Hofmann T. Veroyatnostnoe latentnoe semanticheskoe indeksirovanie [Probabilistic latent semantic indexing]. *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in en:Information Retrieval*, 1999, pp. 50–57.

29. Pelleg D., Moore A. *Rasshirenie algoritma k-srednikh zaschet effektivnoy otsenki chisla klasterov* [X-means: Extending K-means with Efficient Estimation of the Number of Clusters], Pittsburgh, School of Computer Science, Carnegie Mellon University Publ. House, 2000, pp. 1–8.

30. Perez-Aguera J. R., Arroyo J., Greenberg J., Iglesias J. P., Fresno V. Ispolzovanie BM25F dlya semanticheskogo poiska [Using BM25F for semantic search]. *Proceedings of the 3rd International Semantic Search Workshop*, 2010, pp. 1–8.

31. Perez-Iglesias J., Rodrigo A., Fresno V. *Ispolzovanie bm25f dlya patentnogo poiska* [Using bm25f and kld for patent retrieval]. Available at: http://ceur-ws.org/Vol-1176/CLEF2010wn-CLEF-IP-Perez Et2010.pdf.

32. Press W. H Teukolsky S. A., Vetterling W. T., Flannery B. P. *Chislennye metody v Si. Iskusstvo nauchnykh vychisleniy* [Numerical Recipes in C. The Art of Scientific Computing], Cambridge, Cambridge University Press Publ. House, 1997. 994 p.

33. Salton G. *Avtomaticheskiy informatsionnyy poisk* [Automatic Information Retrieval], Ithaca, Cornell University Publ. House, 1980, pp. 41–54.

34. Salton G., Wong A., Yang C. *Vektorno-prostranstvennaya model dlya avtomaticheskogo indeksirovaniya* [Vector Space Model for Automatic Indexing]. *Communications of the ACM*, 1975, pp. 613–620.

35. Salton G., Singhal A. *Vyborochnyy obkhod teksta* [Selective Text Traversal], Ithaca, Department of Computer Science, Cornell University Publ. House, 1995, pp. 131–144.

УДК 618.19 – 073.65:51 – 7

ПРИМЕНЕНИЕ ДВУХМЕРНОГО ФРАКТАЛЬНОГО АНАЛИЗА ДЛЯ ДИФЕРЕНЦИАЦИИ НОРМЫ И ПАТОЛОГИИ КОНТАКТНЫХ ТЕРМОГРАММ МОЛОЧНЫХ ЖЕЛЕЗ

Статья получена в редакцию 29.09.2015 г., в окончательном варианте 06.11.2015 г.

Горшков Олег Георгиевич, преподаватель, Донецкий национальный медицинский университет, 83003, ДНР, г. Донецк, пр. Ильича, 16, e-mail: olgor22@yahoo.com

Старченко Ирина Борисовна, доктор технических наук, профессор, Южный федеральный университет, 347922, Российская Федерация, г. Таганрог, ул. Шевченко, 2, e-mail: star@fep.tti.sfedu.ru

Соботницкий Иван Сергеевич, аспирант, Южный федеральный университет, 347922, Российская Федерация, г. Таганрог, ул. Шевченко, 2, e-mail: pyogenic@mail.ru

Показано, что термограммы поверхности молочных желез имеют двухмерную фрактальную структуру. Это позволяет применить методы двухмерного фрактального анализа для оценки фрактальных свойств распределения температуры при норме и патологии (в т.ч. при онкологических заболеваниях). Для дифференциации нормы и патологии предлагается использовать метод DMA (detrending moving average) расчета показателя Херста для многомерных фракталов. Авторами были обработаны данные по термограммам 478 женщин в возрасте от 15 до 80 лет. В результате проведенных расчетов было выявлено статистически значимое различие между показателями Херста для распределений разницы температур двух симметричных точек левой и правой молочных желез термограмм для контрольной группы, группы больных раком молочной железы; группы больных фиброзно-кистозной мастопатией. Значение этих показателей Херста для контрольной группы H = 0,14 (0,08; 0,19 – нижняя и верхняя границы 95 % доверительного интервала) меньше значений по сравнению с группой больных раком молочной железы H = 0,19 (0,11; 0,26) и группой больных фиброзно-кистозной мастопатией H = 0,17 (0,12; 0,22). Контактные термографические методы могут быть рекомендованы для массового (скринингового) контроля состояния молочных желез как способ выявления пациентов, нуждающихся в дополнительных обследованиях.

Ключевые слова: молочные железы, выявление заболеваний, термография, фрактальный анализ изображений, фрактальная структура, показатель Херста, метод DMA, статистический анализ