На правах рукописи

Бондарчук Дмитрий Вадимович

АЛГОРИТМЫ ИНТЕЛЛЕКТУАЛЬНОГО ПОИСКА НА ОСНОВЕ МЕТОДА КАТЕГОРИАЛЬНЫХ ВЕКТОРОВ

Специальность 05.13.17—
«Теоретические основы информатики»

Автореферат

диссертации на соискание ученой степени

кандидата физико-математических наук

Челябинск — 2016

Работа выполнена на кафедре естественнонаучных дисциплин ФГБОУ ВО «Уральский государственный университет путей сообщения»

Научный руководитель: **ТИМОФЕЕВА Галина Адольфовна**
доктор физико-математических наук, профессор,
заведующая кафедрой естественнонаучных дисциплин,
ФГБОУ ВО «Уральский государственный университет путей сообщения»

Официальные оппоненты: **ХОМОНЕНКО Анатолий Дмитриевич,**
доктор технических наук, профессор,
заведующий кафедрой информационных и вычислительных систем,
ФГБОУ ВО «Петербургский государственный университет путей сообщения Императора Александра I»

ВЕРЕТЕННИКОВ Александр Борисович,
кандидат физико-математических наук,
доцент кафедры вычислительной математики,
ФГАОУ ВО «Уральский федеральный университет имени Первого президента России Б.Н. Ельцина»

Ведущая организация: ФГАОУ ВО «Национальный исследовательский университет «Высшая школа экономики»

Защита состоится 15 марта 2017 г. в 12:00 часов на заседании диссертационного совета Д 212.298.18 при ФГАОУ ВО «Южно-Уральский государственный университет (национальный исследовательский университет)» по адресу: 454080, г. Челябинск, пр. Ленина, 76, ауд. 1001.

С диссертацией можно ознакомиться в библиотеке Южно-Уральского государственного университета и на сайте: http://www.susu.ru/ru/dissertation/d-21229818/bondarchuk-dmitriy-vadimovich.

Автореферат разослан

Ученый секретарь
диссертационного совета М.Л. Цымблер

Общая характеристика работы

Актуальность темы. В последнее десятилетие интеллектуальный анализ текстовых данных получил широкое распространение в связи потребностью многих отраслей экономики и науки в систематезации и автоматической категоризации больших объемов таких данных. Одним из самых перспективных подходов к решению задач автоматического поиска является подход, основанный на машинном обучении. В настоящее время исследованию интеллектуального анализа данных и развитию методов автоматической классификации и кластеризации посвящен ряд работ, подавляющее большинство из которых основано на векторной модели представления знаний, а так же на использовании семантических сетей. Источниками при проведении диссертационного исследования послужили труды отечественных и зарубежных ученых по основам интеллектуального анализа данных: труды T. Landauer, S. Deerwester, S. Streeter, А.Д. Хомоненко, И.С. Некрестьянова и А.Н. Соловьева по методу латентно-семантического анализа и методу представления знаний с помощью терм-документной матрицы, труды M. Minsky и К.В. Воронцова по вероятностным алгоритмам, труды G. Salton, С.В. Моченова, А.М. Бледнова и Ю.А. Луговских по векторной модели представления знаний и труды G. Miller, C. Fellbaum, Н.В. Лукашевич, Б.В. Доброва по семантическим БД, труды С.О. Кузнецова, Д.А. Ильвовского, А.В. Бузмакова, Д.В. Гринченкова, Б.Ю. Лемешко, С.Н. Постовалова по обработке текстовых данных на основе решеток замкнутых описаний и таксономий.

В качестве недостатка большинства существующих на сегодняшний день методов и алгоритмов можно выявить неучет взаимодействия элементов информации между собой и отношения пользователя к знанию, вследствие чего снижается релевантность поиска. Таким образом, **актуальной** является задача улучшения качества интеллектуального анализа текстовых данных за счет учета семантической и лексикографической взаимосвязи термов, и решения проблемы лексической многозначности и разработки методов, обеспечивающих непустой результат для любой обучающей выборки.

Цель и задачи исследования. *Целью* данной работы являлась разработка алгоритма интеллектуального анализа данных, гарантирующего, что пользователь на любой свой запрос получит непустую выборку, отсортированную по степени «полезности».

Для достижения поставленной цели были поставлены следующие *задачи*:

1. Разработка модели образа текстового документа и соответствующего метода отображения текста в семантическое пространство, обеспечивающих компактное представление документа в оперативной памяти.
2. Разработка алгоритма интеллектуального анализа текстов, гарантирующего непустой результат независимо от распределения обучающей выборки по категориям.
3. Разработка алгоритма перевзвешивания векторной модели представления знаний для учета семантической взаимосвязи между термами.
4. Проведение сравнительных экспериментов, оценивающих эффективность разработанных методов и подходов по сравнению с существующими.

Научная новизна работы заключается в разработке автором оригинального способа формирования семантического пространства, основанного на использовании матрицы корреспонденций термов (МКТ), которая подвергается ортогональному разложению, и метода перехода к категориальным векторам с переопределением исходных весов термов с помощью учета семантической взаимосвязи между термами.

Теоретическая ценность работы состоит в том, что в ней проведен сравнительный анализ свойств сингулярного разложения терм-документной матрицы (ТДМ) и ортогонального разложения МКТ. Доказано, что термы, содержащиеся только в коротких документах, отбрасываются при использовании сингулярного разложения ТДМ, но учитываются при использовании предлагаемого подхода. Получены условия совпадения сингулярного разложения терм-документной матрицы, соответствующей всей коллекции, с разложением матрицы, содержащей только длинные документы. **Практическая ценность** работы заключается в том, результаты работы являются основой для разработки поисковых систем, использующих интеллектуальный анализ текстовых данных. Предложенные в работе алгоритмы позволяют производить поиск и классификацию документов, формировать персональные

рекомендации пользователю, упорядоченные по степени соответствия его запросу.

Методы исследования. Методологической основой исследования являются методы линейной алгебры, статистического и системного анализа, интеллектуального анализа данных, семантического анализа.

Степень достоверности результатов. Все утверждения, связанные со свойствами ортогонального разложения матицы корреспонденций термов, сформулированы в виде теорем и снабжены строгими доказательствами. Теоретические построения подтверждены тестами, проведенными в соответствии с общепринятыми методиками.

Апробация работы. Основные результаты работы докладывались на:

1. Научно-практической конференции «Дни науки ОТИ НИЯУ МИФИ-2012» (Озерск, ОТИ НИЯУ МИФИ).
2. Научно-практической конференции «Дни науки ОТИ НИЯУ МИФИ-2013» (Озерск, ОТИ НИЯУ МИФИ).
3. Научно-практической конференции «Математические методы решения исследовательских задач» (Екатеринбург, УрГУПС).
4. Научно-практической конференции «Актуальные проблемы автоматизации и управления» (Челябинск, ЮУрГУ).
5. Международной (46-ой Всероссийской) школе-конференции "Современные проблемы математики и ее приложений"(ИММ УрО РАН, Екатеринбург, 2015).
6. IX Международной научно-практическая конференция «Отечественная наука в эпоху изменений: постулаты прошлого и теории нового времени» (Национальная ассоциация ученых, Екатеринбург, 2015)
7. 41st International Conference «Applications of Mathematics in Engineering and Economics» (Sozopol, Bulgaria, 2015).
8. International Conference and PhD Summer School "Groups and Graphs, Algorithms and Authomata"(Екатеринбург, 2015)
9. Международной (47-ой Всероссийской) школе-конференции "Современные проблемы математики и ее приложений"(ИММ УрО РАН, Екатеринбург, 2016).

Публикации. Основные результаты по теме диссертации изложены в 10 печатных работах. Работы [1–5] опубликованы в журналах, включенных ВАК в перечень изданий, в которых должны быть опубликованы основные результаты диссертаций на соискание ученой степени доктора и кандидата наук. Работы [6–7] опубликованы в изданиях, индексируемых в SCOPUS и Web of Science. В работах [3–6] научному руководителю Г.А. Тимофеевой принадлежит общее математическое руководство и консультирование, Д.В. Бондарчуку — все полученные результаты. В работе [7] А.В. Мартыненко принадлежит математическая формализация задачи, Д.В. Бондарчуку — доказательство основных теоретических утверждений.

Структура и объем работы. Диссертация состоит из введения, четырех глав, заключения и библиографии. В приложении А приведены основные обозначения, используемые в диссертации. В приложении Б приведен словарь терминов, используемых в диссертации. Объем диссертации составляет 141 страница, объем библиографии — 124 наименования.

Содержание работы

Во **введении** обоснована актуальность темы диссертации, изложены цель и задачи исследования, научная новизна и практическая ценность полученных результатов.

В первой главе, «Основные методы интеллектуального анализа данных», рассматриваются тенденции развития интеллектуального анализа данных и дается обзор научных исследований в области современных методов. Особое внимание уделяется латентно-семантическому анализу и использованию семантических сетей.

Во второй главе, «Интеллектуальный метод подбора персональных рекомендаций, гарантирующий получение непустого результата», предлагается новый метод интеллектуального анализа данных, который на любой запрос пользователя дает пользователю непустой ответ, отсортированный по степени релевантности запросу пользователя.

Предлагаемый алгоритм обучения состоит из следующих этапов:

1. Обработка обучающей выборки текстов с использованием стеммера и определение частот вхождений термов в документы.

2. Уменьшение количества термов с использованием сингулярного разложения матрицы корреспонденций термов.
3. Вычисление категориальных векторов документов базы данных вакансий и пользовательского запроса.
4. Вычисление коэффициентов близости между пользовательским запросом и данными базы, сортировка по убыванию и выбор q первых элементов.

В качестве модели представления была выбрана векторная модель, в которой каждый текстовый документ из коллекции представляется, как вектор в векторном пространстве.

Пусть имеется некоторая обучающая выборка m текстов. Представим ее в виде $m \times n$ матрицы X, столбцами которой являются вектора термов x_j, n — количество термов. Термом будем называть слово, обработанное с помощью стеммера Портера, не содержащееся в списке стоп-слов. Вектор терма t_j представляет собой m-мерный вектор:

$$x_j = \{x_{1j}, x_{2j}, ..., x_{mj}\} \tag{1}$$

где x_{ij} — частота встречаемости терма в документе:

$$x_{ij} = tf(t_j,d_i) \tag{2}$$

где d_i – i-ый документ из обучающей выборки, $i = 1,\ldots,m$, $tf(t_j,d_i)$ частота встречаемости терма t_j в документе d_i (term frequency). Матрица X называется *терм-документной матрицей.*

Для предварительной подготовки данных к анализу используются следующие алгоритмы: удаление стоп-слов, метод стемминга Портера, метод выделения семантического ядра. *Стемминг* — это процесс нахождения основы слова для заданного исходного слова. *Семантическое ядро* — это подборка понятий, имеющих существенное значение для данной предметной области.

Определение 1. Матрица корреспонденций термов $G = \{g_{ij}\}$ — это квадратная матрица, элементами которой являются коэффициенты g_{ij}, отражающие близость i-го и j-го термов, для которых выполняются следующие условия:

1. $g_{ij} = g_{ji}$;

2. $0 \leq g_{ij} \leq 1$ для всех i и j;

3. $g_{ij} = 0$ при отсутствии взаимосвязи между термами.

В качестве меры близости можно рассматривать: скалярное произведение векторов, соответствующих нормированным векторам термов; модуль корреляции между векторами-термами; меру Дайса (*Dice measure*) или меру Жаккара (*Jaccard measure*).

Основное назначение матрицы G — отображение взаимосвязей термов внутри документов, построенное на основе знаний частоте об их совместных употреблениях. На рисунке 1 изображен случай, когда термы t_1 и t_2 совместно встречаются в документе d_2, а термы t_2 и t_3 — в документе d_1. Таким образом, термы t_1 и t_3 так же связаны между собой через терм t_2.

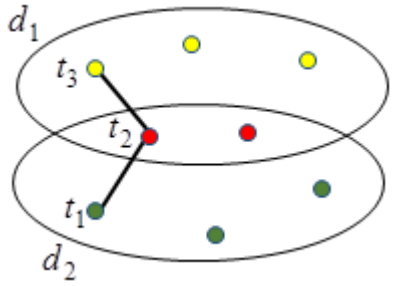

Рис. 1 — Иллюстрация взаимосвязей термов

Построим нормированную терм-документную матрицу $Y = \{y_{ij}\}$, где

$$y_{ij} = \frac{x_{ij}}{\sum\limits_j x_{ij}} = \frac{x_{ij}}{n_i} \tag{3}$$

здесь n_i — общее количество слов в документе d_i. Через y_j обозначим вектор $y_j = \{y_{1j}, y_{2j}, ..., y_{mj}\}$. Будем рассматривать матрицу, состоящую из всех возможных скалярных произведений векторов термов y_j, определяемых по формулам (1) — (3):

$$G = ((y_i, y_j))_{i,j=1}^{n} = Y^T Y \tag{4}$$

Очевидно, что эта матрица корреспонденций термов (МКТ) является симметричной и неотрицательно определенной.

Сингулярное разложение — это математическая операция, представляющая матрицу размера $m \times n$ в виде произведения $X = USV^T$, где U и V — ортогональные матрицы размера $m \times m$ и $n \times n$ соответственно, S — прямоугольная диагональная матрица размера $m \times n$. Под прямоуголь-

ной диагональной матрицей понимается матрица такая, что $s_{ij} = \sigma_i$ при $j = i \leq min(m, n)$ и $s_{ij} = 0$ в остальных случаях. Ортогональные матрицы U и V можно подобрать таким образом, чтобы диагональные элементы матрицы S были расположены по убыванию: $\sigma_1 \geq \sigma_2 \geq .. \geq \sigma_r > \sigma_{r+1} = ... = \sigma_n = 0$, где r — ранг матрицы X.

Основная идея латентно-семантического анализа текстов, предложенного учеными S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, состоит в том, что матрица X_k, содержащая только k первых компонент сингулярного разложения терм-документной матрицы X, отражает основную структуру различных зависимостей, присутствующих в исходной матрице.

Утверждение 1. Ортогональное разложение матрицы корреспонденций термов G, определенной по формуле (4), имеет вид: $G = VZV^T$, где V — ортогональная матрица правых сингулярных векторов в разложении нормированной терм-документной матрицы Y, матрица Z — диагональная матрица размера $n \times n$, на диагонали которой стоят $(\sigma_i)^2$, σ_i — сингулярные коэффициенты разложения матрицы Y.

Часть матрицы G, содержащая только k линейно независимых компонент, будет отражать основную структуру зависимостей между термами, присутствующих в исходной матрице. Усеченную матрицу G до размерности k обозначим G_k: $G_k = V_k Z_k V_k^T$

Теорема 1. *Пусть проводится выделение семантического ядра из коллекции k длинных документов, длиной Φ, $m - k$ коротких документов, длиной ϕ, причем длина коротких документов удовлетворяет условию $\phi < \epsilon\Phi$. Тогда сингулярные числа $\sigma_s(X)$ в разложении терм-документной матрицы X близки к сингулярным числам $\sigma_s(\Phi A)$ терм-документной матрицы ΦA, содержащей только длинные документы. Для сингулярных чисел матрицы X выполняется неравенство*

$$\sigma_s(\Phi A) \leq \sigma_s(X) \leq \sigma_s(\Phi A)\left(1 + 0.5\epsilon^2 \cdot \frac{(m-k)}{\sigma_s^2(A)}\right), \text{если } s \leq r_A. \qquad (5)$$

Здесь $\sigma_s(A)$ — s-ое сингулярное число матрицы A, $\Phi\sigma_s(A) = \sigma_s(\Phi A)$, r_A — ранг матрицы A. При $s > r_A$ сингулярные числа матрицы X удовлетворяют неравенству $0 \leq \sigma_s(X) \leq \phi\sqrt{m-k}$

Теорема 2. *Пусть K, термов $\{t_1, \ldots, t_K\}$ содержатся только в длинных документах длины Φ, остальные $n - K$ содержатся только в коротких документах длины ϕ и $\phi < \epsilon\Phi$.*

Тогда при выделении семантического ядра путем сингулярного разложения ТДМ $X = \Phi A + \phi B$, с условием отбрасывания сингулярных чисел все $n - K$ термов $\{t_{K+1}, \ldots, t_n\}$ не будут учитываться при построении семантического ядра при достаточно малых $\epsilon > 0$.

Число k устанавливается путем отбрасывания незначимых компонент диагональной матрицы Z. Незначимыми являются компоненты, значения которых ниже порогового значения σ^*. От выбора порога отсечения σ^* зависит количество термов k, а также точность результатов анализа.

Третий этап алгоритма основан на учете разбиения терминов на *категории* и состоит в представлении всех документов пользователей через *категориальные векторы*. Под категорией в данном случае понимается некое именованное множество текстов, объединенных по определенному признаку. Все тексты обучающей выборки разбиваются экспертом на категории $\{c_1, c_2, ..., c_l\}$, общее число категорий l невелико, обычно $l \leq 30$. Для получения вектора категории необходимо проанализировать все тексты, которые в нее входят, и построить для каждого из них векторную модель в пространстве R^k с учетом линейного преобразования V_k, полученного сингулярным разложением МКТ.

Определение 2. Векторной моделью категории (набора текстов) будем называть средний вектор между векторами ее текстов, то есть вектор $\vec{c} \in R^k$, состоящий из средних арифметических соответствующих компонент векторов текстов

$$\vec{c} = \left\{ \frac{\sum\limits_{d_i \in D_c} w_{i1}}{|D_c|}, \frac{\sum\limits_{d_i \in D_c} w_{i2}}{|D_c|}, ..., \frac{\sum\limits_{d_i \in D_c} w_{ik}}{|D_c|} \right\} \tag{6}$$

где D_c — множество документов, содержащихся в категории c, $|D_c|$ — количество документов, содержащихся в категории, w_{ij} — вес j-ого терма в i-ом документе, d_i — i-ый документ из категории c.

Найдем векторные модели $\vec{c_1}, \vec{c_2}, ..., \vec{c_l}$ для всех категорий $c_1, c_2, ..., c_l$.

Определение 3. Коэффициентом принадлежности будем называть скалярное произведение $z_{ij} = \left(\vec{d_i}, \vec{c_j}\right)$, где $\vec{c_j}$ — вектор j-ой категории, $\vec{d_i}$ вектор i-го документа.

Определение 4. Категориальный вектор Z_i документа d_i — это вектор, составленный из коэффициентов принадлежности z_{ij} текста d_i каждой из категорий: $\vec{Z_d} = \left\{(\vec{d}, \vec{c_1}), (\vec{d}, \vec{c_2}), ..., (\vec{d}, \vec{c_l})\right\} \in R^l$.

Таким образом, проведен переход из исходного пространства R^k (k — число наиболее значимых термов, полученных с помощью сингулярного разложения МКТ) в новое векторное пространство R^l , где l — число категорий.

На следующем этапе определяем определим коэффициенты близости γ_{ij} между категориальными векторами по формуле

$$\gamma_{ij} = \frac{\left(\vec{Z_i}, \vec{Z_j}\right)}{\left\|\vec{Z_i}\right\| \left\|\vec{Z_j}\right\|} \tag{7}$$

где $\left\|\vec{Z_i}\right\|$ — евклидова норма вектора $\vec{Z_i}$. Непосредственный выбор рекомендаций состоит из следующих шагов.

1. Рассчитывается категориальный вектор пользователя $\vec{Z}_{польз.}$, для которого происходит подбор рекомендаций.
2. Вычисляются коэффициенты близости между $\vec{Z}_{польз.}$ и всеми векторами $\vec{Z_j}$ из базы. Полученные значения $\gamma_{польз.j}$ сортируются по убыванию.
3. Из отсортированной выборки значений $\gamma_{польз.j}$ извлекается q первых элементов, где q – заданное количество выдаваемых каждому пользователю рекомендаций.

Алгоритм позволяет получить выборку, отсортированную по степени «полезности» конечному пользователю. Предлагаемый способ хорош в первую очередь тем, что в случае, когда данные распределены между категориями неравномерно, пользователь получит непустой результат.

Третья глава, «Векторная модель представления знаний использующая семантическую близость термов», посвящена применению семантической близости термов при обучении классификатора, а именно перевзвешиванию весов термов векторной модели представления знаний. Для вычисления се-

мантической близости термов используется авторская адаптация расширенного алгоритма Леска.

Векторная модель с учетом семантической близости термов решает проблему неоднозначности синонимов. Чтобы учесть семантическую связь между терминами, вес термина в документе будем рассчитывать несколько иначе, чем в классической векторной модели представления знаний. Настройка весов термов производится с помощью вычисления семантической близости связанных термов. Новый вес терма рассчитывается следующим образом:

$$\tilde{w}_{dt_1} = w_{dt_1} + \sum_{t=1}^{n_d} similarity(t_1, t) \tag{8}$$

где $t \neq t_1$, w_{dt_1} — вес терма в документе d до настройки, рассчитанный по формуле (2), $similarity$ — семантическая близость термов t_1 и t, рассчитываемая с помощью адаптации расширенного метода Леска. Суммирование происходит по всем термам документа d.

В качестве веса терма в формуле (8) может использоваться частота его появления в документе или другие аналогичные показатели. Для эффективной работы алгоритма использовалась меру TF-IDF:

$$w_{dt} = tf.idf(d,t) = \ln(tf(d,t) + 1) \ln \frac{|D|}{df(t)} \tag{9}$$

где $df(t)$ — документная частота терма, показывающая количество документов, в которых встречается терм, $tf(t,d)$ — число раз, с которым терм t встречается в документе d, нормализованное общим количеством термов в документе d, $|D|$ — общее количество документов.

Кроме того, предлагается способ вычисления семантической близости, основанный на предположении, что семантически близкие термы употребляются в одинаковых или схожих контекстах. В главе предлагается способ вычисления семантической близости между двумя словами или фразами, основанный на статистическом подходе. Главная идея состоит в том, что связность между словами удобнее представлять в виде *контекстного множества*, т.е. множества слов, связанных с заданным термином.

Для построения контекстного множества используется матрица корреспонденций термов G. Для построения контекстного множества i-того терма, рассматриваем i-тую строку матрицы, исключая элемент g_{ii}. Обозначим по-

лученный вектор $G_i = \{g_{ij}\}_{j=1,i\neq j}^{n}$, n — количество термов. Вычислим среднее арифметическое среди элементов вектора

$$\bar{g}_i = \frac{1}{n-1}\sum_{i\neq j} g_{ij} \tag{10}$$

Отбросим все компоненты вектора G_i, которые меньше среднего значения $\bar{g}_i$. Контекстное множество i-того терма будет состоять из термов, соответствующих оставшимся компонентам вектора G_i. Предлагаемый метод можно разделить на несколько шагов:

1. Формирование контекстных множеств слов w_i и w_2. Пусть C_1 и C_2 — контекстные множества слов w_1 и w_2, содержащие слова, с которыми w_1 и w_2 употребляются в одном контексте. Затем формируем общее контекстное множество слов: $C = C_1 \cup C_2$. Обозначим $D(C)$ мощность множества C.
2. Вычисление нормализованных близостей между словами c_i из контекстного множества C и каждым из слов w_1 и w_2:

$$\rho(c_i, w_1) = \frac{n(c_i,w_1)}{n_{max}(w_1)}; \qquad \rho(c_i, w_2) = \frac{n(c_i,w_2)}{n_{max}(w_2)}. \tag{11}$$

   где $n(c_i,w_2)$ — количество документов, где c_i и w_1 встречаются вместе, а $n_{max}(w_1)$ рассчитывается как максимум частот по всем словом из объединенного контекстного множества C: $n_{max}(w_j) = \max\{n(c_i,w_j)|c_i \in C\}$.
3. Расчет семантической близости. Для расчета семантической близости между словами w_1 и w_2. Рассчитаем коэффициенты R_i для всех слов из контекстного множества C по формуле:

$$R_i = \frac{\min\{\rho(c_i,w_1), \rho(c_i,w_2)\}}{\max\{\rho(c_i,w_1), \rho(c_i,w_2)\}} \tag{12}$$

Обозначим p — коэффициент совместной встречаемости w_1 и w_2 во всей выборке, равный 2 в случае, когда оба слова встречаются хотя бы в одном документе, и 1 в противном случае. Через s обозначим коэффициент синонимии, $s = 1$, если слова w_1 и w_2 являются синонимами, и $s = 0$, в противном случае. Семантическая близость слов w_1 и w_2 рассчитывается по формуле:

$$\hat{\rho}_{сем.}(w_1,w_2) = \frac{\sum_{i=1}^{D(C)} (\frac{pR_i}{1+R_i} + s)}{0.75 \cdot (1+s) \cdot D(C)} \quad (13)$$

где коэффициенты R_i, p находятся по формулам (12). Для семантически близких слов, полученный коэффициент $\hat{\rho}_{сем.}(w_1,w_2)$ близок к 1.

В четвертой главе, «Вычислительные эксперименты», описываются эксперименты по исследованию эффективности разработанных в диссертации моделей, методов и алгоритмов.

Для оценки эффективности векторной модели представления знаний учитывающую семантическую близость термов использовались известные меры оценки качества классификаторов *F-measure* и *purity*. В таблице 1 представлены результаты работы классификаторов, взяты средние значения оценок за 50 тестов.

Таблица 1 — Оценка результатов работы алгоритма классифиции

	Векторная модель		Векторная модель с использованием семантической близости	
Множество	F-measure	Purity	F-measure	Purity
Объявления о работе	0.31	0.33	0.65	0.66
Новости	0.56	0.58	0.61	0.64
Литературные аннотации	0.56	0.57	0.63	0.67

В заключении в краткой форме излагаются итоги выполненного диссертационного исследования, представляются отличия диссертационной работы от ранее выполненных родственных работ других авторов, даются рекомендации по использованию полученных результатов и рассматриваются перспективы дальнейшего развития темы.

Основные результаты диссертационной работы

На защиту выносятся следующие новые научные результаты:

1. Разработаны модель образа текстового документа и соответствующий метод отображения текста в семантическое пространство, обеспечивающие компактное представление документа в оперативной

памяти на основе матрицы корреспонденций термов, которая подвергается ортогональному разложению.

2. Разработан алгоритм интеллектуального анализа текстов, гарантирующий непустой результат независимо от распределения обучающей выборки по категориям на основе использования вычисления категориальных векторов для упорядочения результирующей выборки по степени релевантности запросу пользователя.
3. Предложен метод перевзвешивания термов векторной модели с помощью вычисления их семантической взаимосвязи друг с другом на основе авторской адаптации алгоритма Леска.
4. На основе разработанных методов и подходов реализован алгоритм подбора рекомендаций. Проведены вычислительные эксперименты, подтверждающие более высокую эффективность разработанного алгоритма по сравнению с существующими.

Публикации по теме диссертации

Статьи в журналах из перечня ВАК

1. Бондарчук Д.В. Статистический способ определения семантической близости термов // Системы управления и информационные технологии. – 2015. – Т. 61, № 3. – С. 55–57.

2. Бондарчук Д.В. Алгоритм построения семантического ядра для текстового классификатора // В мире научных открытий. – 2015. – Т. 68, № 8.2. – С. 713–724.

3. Бондарчук Д.В., Тимофеева Г.А. Выделение семантического ядра на основе матрицы корреспонденций термов // Системы управления и информационные технологии. – 2015. – Т. 61, № 3.1. – С. 134–139.

4. Бондарчук Д.В., Тимофеева Г.А. Применение машинного обучения для формирования персональных рекомендаций в сфере трудоустройства // Экономика и менеджмент систем управления. – 2015. – Т. 18, № 4.2. – С. 215—221.

5. Бондарчук Д.В., Тимофеева Г.А. Математические основы метода категориальных векторов в интеллектуальном анализе данных // Вестник

Уральского государственного университета путей сообщения. – 2015. – 4(28). – С. 4–8.

Статьи в изданиях, индексируемых в Scopus и Web of Science

6. Bondarchuk D.V., Timofeeva G.A. Vector space model based on semantic relatedness // AIP Conference Proceedings, V. 1690, Proceedings of 41st International Conference «Applications of Mathematics in Engineering and Economics» (AMEE'15) . – 2015. – Pp. 1–5.

7. Bondarchuk D.V., Martynenko A.V. Spectral properties of a matrix of correspondences between terms // CEUR Workshop Proceedings, Vol. 1662, Proceedings of 47th International Youth School-Conference "Modern Problems in Mathematics and its Applications"(MPMA 2016). – 2016. – Pp. 186–190.

Статьи в изданиях, индексируемых в РИНЦ

8. Бондарчук Д. В. Использование латентно-семантического анализа в задачах классификации текстов по эмоциональной окраске // Бюллетень результатов научных исследований. – 2012. – 2(3). – С. 146–151.

9. Бондарчук Д. В. Выбор оптимального метода интеллектуального анализа данных для подбора вакансий // Информационные технологии моделирования и управления. – 2013. – 6(84). – С. 504–513.

10. Бондарчук Д. В. Интеллектуальный метод подбора персональных рекомендаций, гарантирующий получение непустого результата // Информационные технологии моделирования и управления. – 2015. – Т. 2(92).–С. 130–138.