СИСТЕМНЫЙ АНАЛИЗ, УПРАВЛЕНИЕ И ОБРАБОТКА ИНФОРМАЦИИ / SYSTEM ANALYSIS, MANAGEMENT AND PROCESSING OF INFORMATION

DOI: https://doi.org/10.60797/itech.2024.4.2

МЕТОДИЧЕСКИЕ АСПЕКТЫ РАЗРАБОТКИ ГИПЕРТЕКСТОВЫХ БАЗЫ ЗНАНИЙ И АЛГОРИТМОВ ИНТЕЛЛЕКТУАЛЬНОГО ПОИСКА ДЛЯ ИНФОРМАЦИОННОЙ ПОДДЕРЖКИ ПРОЦЕССОВ СТРАТЕГИЧЕСКОГО ПЛАНИРОВАНИЯ НА РЕГИОНАЛЬНОМ УРОВНЕ

Научная статья

Низамутдинов М.М.[1, *], Давлетова З.А.[2]
[1] ORCID : 0000-0001-5643-1393;
[2] ORCID : 0009-0008-4389-2113;
[1, 2] Уфимский федеральный исследовательский центр Российской академии наук, Уфа, Российская Федерация

* Корреспондирующий автор (marsel_n[at]mail.ru)

Аннотация

Исследована проблема организации интеллектуальной информационной поддержки процессов стратегического планирования с применением гипертекстовой базы знаний (ГБЗ). Предложена концептуальная схема организации информационной поддержки принятия решений, описаны основные этапы формирования структуры ГБЗ, предложена модель формирования семантических категорий документов в базе знаний на основе метода обучения искусственной нейронной сети. Рассмотрена структура элементов ГБЗ, на основе которой сформирована схема информационной поддержки процедур стратегического планирования на основе алгоритма интеллектуального информационного поиска регламентирующих документов с использованием тезауруса предметной области. Рассмотрен пример реализации процедуры поиска в ГБЗ на примере решения одной из задач в рамках процедуры стратегического планирования.

Ключевые слова: стратегическое планирование, информационная поддержка, гипертекстовая база знаний, интеллектуальный поиск.

METHODOLOGICAL ASPECTS OF HYPERTEXT KNOWLEDGE BASE DEVELOPMENT AND INTELLECTUAL SEARCH ALGORITHMS FOR INFORMATION SUPPORT OF STRATEGIC PLANNING PROCESSES AT THE REGIONAL LEVEL

Research article

Nizamutdinov M.M.[1, *], Davletova Z.A.[2]
[1] ORCID : 0000-0001-5643-1393;
[2] ORCID : 0009-0008-4389-2113;
[1, 2] Ufa Federal Research Center of the Russian Academy of Sciences, Ufa, Russian Federation

* Corresponding author (marsel_n[at]mail.ru)

Abstract

The problem of organization of intellectual information support of strategic planning processes using hypertext knowledge base (HKB) is studied. The conceptual scheme of organization of information support of decision-making is suggested, the main stages of formation of the HKB structure are described, the model of formation of semantic categories of documents in the knowledge base using the method of artificial neural network training is proposed. The structure of HKB elements is examined, on the basis of which the scheme of information support of strategic planning procedures based on the algorithm of intellectual information search of normative documents using the thesaurus of the subject area is formed. The example of implementation of the search procedure in the HKB is discussed on the example of solving one of the tasks within the framework of the strategic planning procedure.

Keywords: strategic planning, information support, hypertext knowledge base, intelligent search.

Введение

В современных условиях конкурентного рынка успешное функционирование любой сложной организационной системы требует постоянного совершенствования механизмов управления, способных обеспечить адекватную и своевременную реакцию на постоянные изменения внутренних и внешних условий. Важным направлением совершенствования процессов управления организационными системами в последние годы является повышение эффективности функций стратегического планирования и, в первую очередь, повышение качественных характеристик информации, используемой управленцами различных уровней в процессе принятия решений. Развитие отечественных информационно-управляющих систем достигло значительных результатов в области разработки технологий обработки и хранения структурированной информации, в частности, происходит активное развитие систем поддержки принятия решений на основе баз и хранилищ данных, ERP-технологий и др. [1]. Вместе с тем наблюдается некоторое отставание в развитии класса систем, ориентированных на цифровую обработку и использование слабоструктурированной, естественно-языковой информации. Значительная часть слабоструктурированных и слабо формализуемых знаний представлена в виде различных *регламентирующих документов* (законов, постановлений, указов, положений, методических рекомендаций и т.д.), представляющих собой естественно-языковое описание правил реализации различных процедур и функций принятия решений [2]. Эффективное использование накопленного десятилетиями ценнейшего опыта и знаний в области организационного и стратегического управления, отраженного в

регламентирующих документах, является сегодня необходимым условием совершенствования системы информационного обеспечения деятельности органов государственного управления. Его реализация требует выработки новых подходов, методов и технологий управления слабоструктурированными знаниями, способными одновременно обеспечить оперативность, полноту и точность информационной поддержки процессов принятия решений [3]. Поэтому разработка технологий создания систем поддержки принятия решений на основе различного типа баз знаний и реализация интеллектуальной информационной поддержки при работе базами регламентирующих документов является актуальной научной задачей.

Общая концепция реализации информационной поддержки принятия решений на основе гипертекстовой базы знаний

В рамках решения обозначенной задачи предложены основные этапы и концептуальная схема организации интеллектуальной информационной поддержки принятия решений на основе гипертекстовой базы знаний (рис.1).

Согласно разработанной схеме, первым этапом реализации системы является системный анализ и моделирование предметной области. Основной целью этого этапа является системное описание предметной области и накопление знаний о процессах управления в виде комплекса дескриптивных и когнитивных моделей. Основными источниками информации для формирования системной модели предметной области являются экспертные знания и естественно-языковые знания, определенные в тексте соответствующих регламентирующих документов.

На следующем этапе исследований разрабатываются модели и алгоритмы извлечения знаний исходя из сформированных источников информации с использованием методов интеллектуального анализа данных [4]. Для извлечения знаний из содержания регламентирующих документов предложено использовать методы лингвистического анализа текстов [5]. При этом получение знаний на основе разработанных предварительно описательных моделей предметной области предложено реализовать с применением алгоритма семантического анализа результатов системного моделирования. Предполагается также схема структурирования полученных знаний на основе методов классификации с использованием нейронных сетей.

На основе выделенных на предыдущем этапе знаний, предлагается разработать гипертекстовую базу знаний предметной области. В качестве модели представления знаний используется семантическая сеть, которая определяет множество терминов предметной области и устанавливает между ними различные типы семантических отношений. База знаний объединяет термины и отношения, сформированные на основе регламентирующих документов и моделей процессов принятия решений в форме интегрированного словаря и тезауруса предметной области.

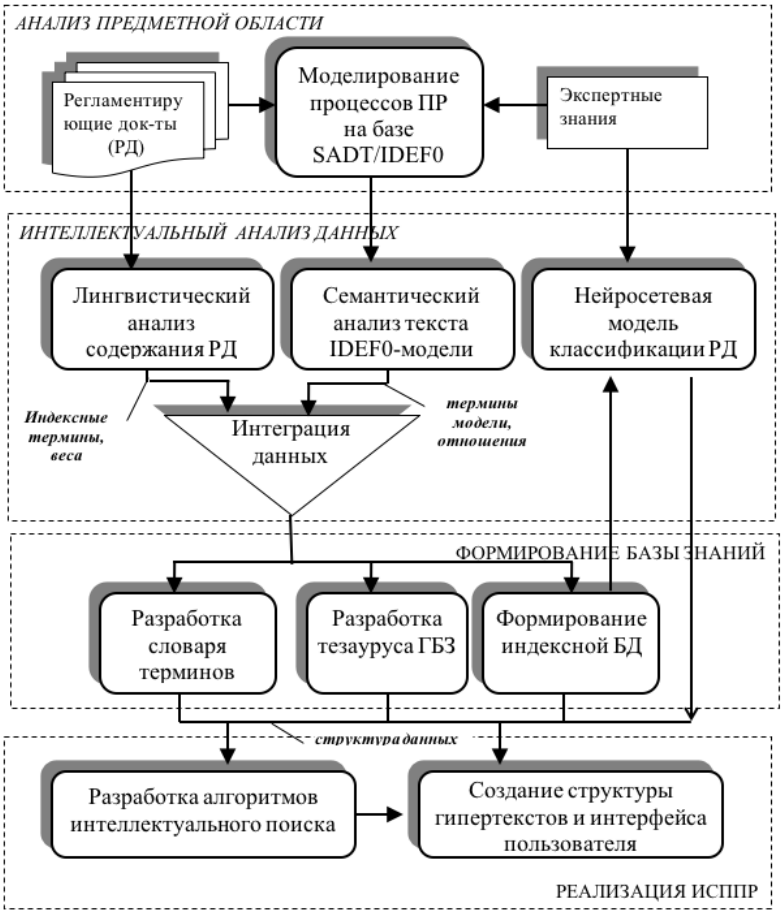

Рисунок 1 - Схема разработки информационной системы поддержки принятия решений на основе гипертекстовой базы знаний

DOI: https://doi.org/10.60797/itech.2024.4.2.1

На последнем этапе предлагаемой схемы реализации СППР исходя из структуры тезауруса предметной области формируется гипертекстовая структура регламентирующих документов, а также реализуется модель интеллектуального информационного поиска документов на основе сформированной гипертекстовой базы знаний.

Таким образом, предлагаемая концептуальная схема направлена на реализацию интеллектуальной проблемно-ориентированной информационной поддержки пользователей в процессе поиска рекомендаций по принятию решений на основе хранящихся в базе знаний регламентирующих документов.

Схема формирования семантических категорий документов на основе обучения искусственной нейронной сети

Реализация информационной поддержки процессов стратегического планирования в режиме реального времени требует обработки большого объема данных, что обуславливает необходимость использования современных методов интеллектуального анализа, позволяющих структурировать эти данные. Одним из наиболее эффективных и широко используемых методов структурирования большого объема данных являются методы классификации и кластерного анализа. Классификация – математическая процедура многомерного анализа, позволяющая на основе множества показателей характеризующих ряд объектов сгруппировать их в классы таким образом, чтобы объекты в одной группе обладали похожими свойствами, а свойства в среднем между группами максимально различались. В результате такого разбиения удается получить множество категорий объектов, соответствующих условию оптимального соотношения между потерями точности, порождаемыми представлением любого индивидуального объекта типичными средними характеристиками его класса, и получением экономии в результате замены многих индивидуальных объектов малым числом классов.

Объектом исследования с использованием методов классификации в данной работе является совокупность регламентирующих документов в сфере стратегического планирования. Априорно, в любой предметной области существует классификация документов по тем или иным критериям. Однако эффективная реализации поставленных задач информационной поддержки требует группировки документов по критерию их ориентированности на процесс принятия решений [6], т.е. порождаемые классы документов должны обеспечивать вполне определенные процедуры принятия решений. Таким образом, целью классификации множества регламентирующих документов является порождение групп документов, близких по содержанию, т.е. формирование так называемых семантических категорий

документов. Такая классификация является важным условием реализации алгоритмов интеллектуального поиска документов в реальном режиме времени.

Особенностью классификации множества регламентирующих документов являются специфические признаки классификации, в роли которых выступают термины предметной области. Возможность успешного проведения такой классификации основано на исследованиях, проведенных в области психолингвистики, в которых выдвигается и обосновывается гипотеза о наличии определенных корреляционных зависимостей понятий (терминов), описывающих определенную проблемную ситуацию [7].

Существуют различные подходы к решению задачи классификации, использующие разные алгоритмы группировки данных. Выбор конкретного метода определяется особенностями решаемой задачи, и в первую очередь способностью каждого из методов формировать желаемые результаты классификации. В данном исследовании для классификации регламентирующих документов были проанализированы два наиболее известных подхода: методы статистической классификации и классификации методом обучения искусственных нейронных сетей. Сравнительный анализ применимости каждого из рассмотренных методов для решения задачи классификации естественно-языковых документов на основе терминов предметной области будет приведен ниже.

Определим задачу классификации регламентирующих документов формально. Предположим, что можно разбить множество документов, обозначенное $\boldsymbol{D^n}$, на непересекающиеся подмножества, соответствующие классам документов (1):

$$D^n = \bigcup\ d_i,\ d_i \bigcap\ d_j = \varnothing, \quad i \neq j, \tag{1}$$

т.е. осуществить однозначные отображения: $D^n \Rightarrow W$, такое что $\forall$ ($\boldsymbol{D}_i(t) \in D^n$) $\exists$ ($w_j \in W$: Ψ ($\boldsymbol{D}(t)$) = w_j). Здесь $\boldsymbol{D}(t)$ – вектора документов, определенное в пространстве терминов предметной области, $W = \{w_1, w_2, \ldots, w_M\}$ – множество предопределенных классов документов. Классификация (распознавание) документа состоит в том, чтобы на основании измерения и анализа вектора документа $\boldsymbol{D}(t)$ отнести его к определенному классу документов $w_j \in W$.

Для рассматриваемого множества документов характерно качественное представление их признаков в виде терминов предметной области, содержащихся в описании документов на естественном языке. Следовательно, необходимо разбить исходное множество документов D^n на M классов таким образом, чтобы получить семантически близкие группы категории документов.

Множество векторов документов $\boldsymbol{D}(t)$ определяется матрицей «термин-документ», полученным в результате лингвистического анализа содержания регламентирующих документов и дескриптивной модели процесса, представленных в виде интегрированной онтологии как множество векторов документов в пространстве терминов предметной области (фрагмент приведен на рис. 2).

	T_1	T_2	T_3	T_4	T_5	T_6	T_7
D₁	0.88	0.00	0.62	0.00	0.00	0.00	0.11
D₂	0.36	0.23	0.58	0.00	0.00	0.00	0.00
D₃	0.24	0.45	0.00	0.49	0.00	0.10	0.00
D₄	0.74	0.28	0.58	0.00	0.00	0.00	0.00
D₅	0.00	0.00	0.00	0.00	0.11	0.00	0.87
D₆	0.00	0.00	0.00	0.00	0.96	0.15	0.00
D₇	0.26	0.00	0.00	0.21	0.00	0.54	0.00

Рисунок 2 - Фрагмент матрицы «термин – документ»

DOI: https://doi.org/10.60797/itech.2024.4.2.2

Здесь t_{ij} – вес i–го термина в описании j-го документа. Таким образом, множество всех регламентирующих документов, представленных в виде матрицы «термин-документ» является исходным для проведения классификации документов.

Рассмотрим сначала процедуру классификации документов статистическими методами. Методы статистической классификации до недавнего времени были одними из самых широко используемых методов, для которых разработана хорошая теоретическая база и основные методы реализованы в большинстве из современных программ статистической обработки данных. Основным преимуществом статистических методов является их непринужденность к предобработке данных, большое разнообразие математически формализованных алгоритмов, позволяющие аналитически строить классифицирующую функцию.

Основную трудность при проведении классификации методами статистического анализа является «удачный» подбор метода и меры расстояния (метрики), которая позволила бы получить более или менее однородные и сопоставимые по размерности классы документов. С этой точки зрения были проанализированы несколько различных методов и метрик, которые позволяли получить приемлемые результаты. Наиболее адекватные результаты были получены при использовании метода иерархической классификации Варда с использованием коэффициента Дейка:

$$\beta_{\text{ij}} = \frac{2n_{ij}(1,1)}{2n_{ij}^{(1,1)} + q_{ij}},$$

где $n_{ij}^{(1,1)}$ – число единичных признаков у *i*–го и *j*–го векторов документов соответственно, q_{ij} – общее число несовпадающих признаков. Коэффициент Дейка придает вдвое больший вес совпадающим терминам и, следовательно, подходит для определения сходства содержания документов.

Исходное множество векторов документов *W* было разбито на *M* предопределенных классов. На полученной дендрограмме (рис. 3) видно, что классы получились достаточно однородными и приблизительно равномощными. В нашем случае это говорит о том, что классификация удачна, поскольку нет классов либо с очень большим, либо с очень малым количеством документов. Это условие формирования классов важно при реализации алгоритма поиска документов.

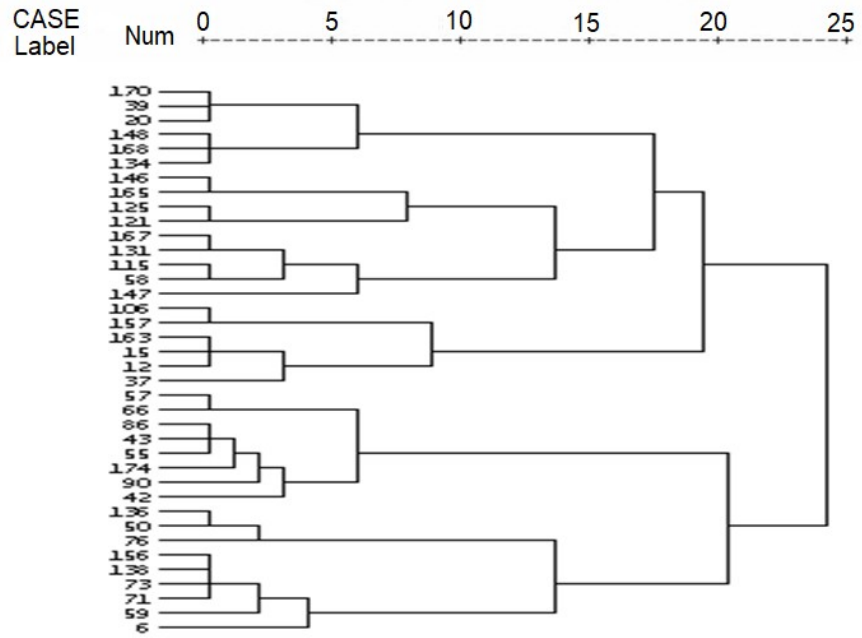

Рисунок 3 - Дендрограмма результатов классификации статистическими методами
DOI: https://doi.org/10.60797/itech.2024.4.2.3

Однако при анализе качественного состава сформированных классов обнаружилось, что документы группируются по принципу «доминантных понятий», а не по желаемому в нашем случае принципу ориентированности определенного класса документов на обеспечение соответствующего процесса принятия решений или определенного этапа рассматриваемого процесса стратегического планирования. Поэтому при сопоставительной оценке состава классов, сформированных алгоритмом классификации и сделанных специалистом-экспертом «вручную», лишь чуть более половины документов попали в «желаемые» классы. Таким образом, методы статистической классификации не дают приемлемого результата качества классификации в силу специфики решаемой задачи и используемых исходных данных.

В существенной степени разрешить недостатки статистических методов классификации удается при использовании методов обучения искусственных нейронных сетей [8]. Архитектура нейронной сети включает взаимосвязанные вычислительные элементы (нейроны), каждый из которых генерирует выходной сигнал в ответ на несколько входных. Выход элемента является входом для других. Каждый вход получает вес (в виде коэффициента в соответствующем уравнении), который корректируется в процессе обучения сети. Обучение сводится к подбору таких весов, при которых нейронная сеть безошибочно распознает эталонную выборку. Обучаемость нейронной сети является ее основным преимуществом, но в то же время предъявляет к процессу обучения достаточно жесткие требования. Для того чтобы научить нейронную сеть корректно распознавать классы, необходимо иметь достаточно большую выборку достоверных обучающих примеров, но если все же удается настроить весовые коэффициенты сети, достигнув приемлемого уровня ошибки обучения, то она способна классифицировать с большой точностью, учитывая математически трудно формализуемые особенности подаваемых на обучение примеров.

Для обучения нейронной сети формируется таблица Ψ независимых векторов документов из различных классов (тренировочное множество описаний документов): $\{w_i, \boldsymbol{t(w_i)}, d(w_i)\}^{K}_{i=1}$, где *K* – количество документов, $d(w_i) \in \{1, \ldots, M\}$ – указание эксперта об истинной принадлежности w_i к одному из *M* классов документов на множестве *D*.

Для проведения классификации документов исходное множество примеров описаний документов было разбито в соотношении 2:1:1 на обучающую, тестовую и контрольную выборки. На этих примерах сеть была обучена, достигнув достаточно малой ошибки обучения, порядка 0,0098, 0,0063 и 0,0042 для обучающей, тестовой и контрольной выборки соответственно. Результаты приведены для структуры сети типа многослойного персептрона с *N* – входным слоем нейронов, равным количеству признаков-дескрипторов, с *M* – нейронами в выходном слое, соответствующем количеству классов разбиения (рис. 4).

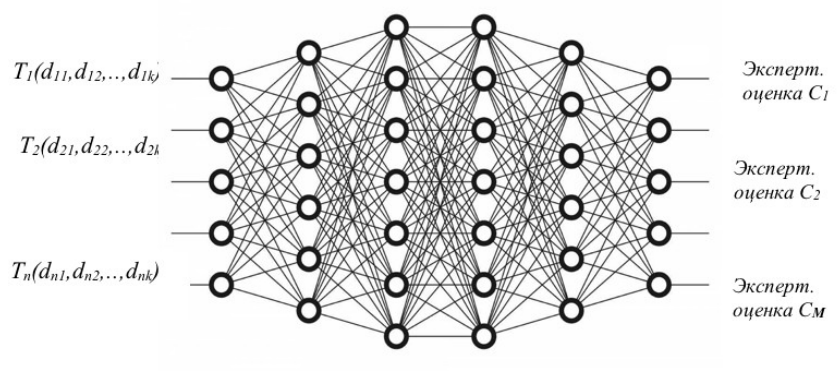

Рисунок 4 - Структура обучаемой нейронной сети
DOI: https://doi.org/10.60797/itech.2024.4.2.4

Такая структура позволяет создавать сеть средней сложности, обычно рекомендуемой в качестве базовой структуры. При наличии плохой сходимости, как правило, в промежуточный слой последовательно добавляют некоторое количество нейронов. Обучение сети проводилось по алгоритму Back Propagation (обратное распространение), для получения приемлемого уровня ошибки обучения понадобилось провести 300 итераций перенастройки весовых коэффициентов. Порог уровня значимости, т.е. минимальное расстояние, на котором принимается положительное решение о включении наблюдения в соответствующий класс, был принят на уровне 0,9.

Для решения задачи классификации обученной сети было предъявлено 50 новых примеров, лишь в одном случае сеть затруднилась отнести пример к одному из предопределенных классов, и в одном случае наблюдалось рассогласование результатов выхода нейронной сети и экспертной оценки. Достоверность порядка 0,96 является достаточно хорошим показателем при обучении нейронной сети. Однако здесь большое значение имеют качественные характеристики исходной выборки и выбор удачной структуры сети.

В качестве эталона для оценки качества классификации были приняты классы, определенные специалистом-экспертом в предметной области. Эталонные классы формировались экспертом по принципу группировки документов в соответствии основными этапами рассматриваемого процесса стратегического планирования [9].

Таким образом, применение методов обучения нейронных сетей для решения задачи классификации естественно-языковых документов на основе их описаний дает достаточно адекватные результаты, обучаясь на достоверных экспертных данных и позволяя с приемлемой точностью отслеживать логику человека-эксперта. Однако задачу приходится решать в рамках определенных ограничений, связанных в первую очередь с необходимостью иметь большой предопределенный набор качественных обучающих примеров.

Одновременно, решение задачи классификации регламентирующих позволяет одновременно структурировать сформированное множество терминов предметной области, которые являются частью описания этих документов, то есть определить категории регламентирующих документов для разработки структуры гипертекстовой базы знаний.

Разработка семантической сети гипертекстовой базы знаний

Значительная часть регламентирующей информации, выраженная в форме «официальных» документов, имеет большой объем текстового описания, поэтому невозможно рассматривать такой документ в качестве носителя знаний, определяющих правила реализации только конкретной функции или процедуры принятия решений. Каждый такой документ определяет регламент реализации целого комплекса функций и процедур, посвященных большим этапам реализации процесса подготовки решений (например, «Методические рекомендации по разработке и корректировке стратегии социально-экономического развития субъекта Российской Федерации и плана мероприятий по ее реализации, утвержденные приказом Минэкономразвития РФ от 23 марта 2017 года № 132» содержит несколько десятков отдельных рекомендаций, характеризующих различные аспекты формирования стратегий развития регионов Российской Федерации).

Вместе с тем в процессе принятия решения специалист, как правило, нуждается в конкретной рекомендации по отдельной проблемной ситуации. Предоставление в качестве такой рекомендации документа большого объема не является эффективным, поскольку специалист вынужден искать нужную ему в данный момент рекомендацию среди множества прочей информации, не отвечающей его текущим информационным потребностям. Поэтому важно определить понятие единицы или блока информации, которая описывает конкретную процедуру принятия решения и воспринимается СППР как количество информации, достаточное для удовлетворения текущих потребностей лица, принимающего решение (ЛПР).

При разработке гипертекстовых хранилищ информации, как правило, используют некоторый подход к структурированию больших документов. Самым распространенным и простым подходом к структурированию в таких случаях является формирование отдельных информационных блоков на основе уже определенного в содержании документа структуры в виде соответствующих глав, разделов и подразделов. Такие разделы изначально определяют некоторый относительно локализованный и самодостаточный блок информации. Тогда небольшой подраздел можно интерпретировать как новый документ, который уже можно воспринимать как единицу информации, способную удовлетворить текущую потребность ЛПР.

Поэтому при разработке системы ППР также было предложено формировать из документов значительного объема множество так называемых «информационных статей» документа, являющихся его частью. Таким способом

формируется новое гипертекстовое пространство документов, в котором отдельные части документов являются самостоятельными источниками информации, но в то же время связаны между собой соответствующими ссылками.

После формирования семантической сети интегрированного тезауруса (онтологии) следующим этапом является разработка на его основе структуры ГБЗ. Структуру базы знаний можно представить как совокупность следующих элементов:

$$S = \{K, T, W(T), D\},$$

где соответственно K – множество семантических категории документов, T – термины тезауруса ГБЗ, W(T) – семантические веса отношений терминов тезауруса, D – информационные статьи (документы) тезауруса ГБЗ.

На рис. 5. представлена общая структура семантической сети ГБЗ.

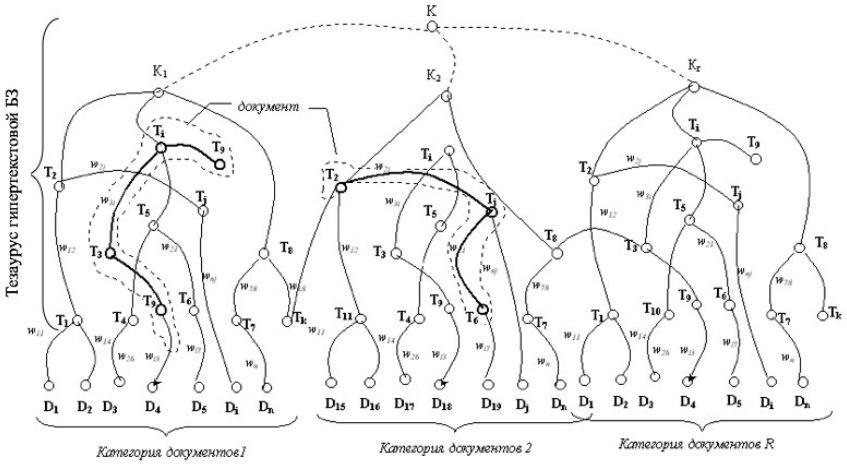

Рисунок 5 - Структура семантической сети ГБЗ

DOI: https://doi.org/10.60797/itech.2024.4.2.5

Из структуры ГБЗ видно, что множества документов образуют категории (классы), которые предварительное были сформированы на основе алгоритмов нейросетевой классификации. Путь к каждому документу на семантическом графе ГБЗ проходит через соответствующее подмножество терминов тезауруса и определяет индекс (координаты) этого документа в едином информационном пространстве терминов предметной области.

Формирование структуры индексов и реализация алгоритма интеллектуального информационного поиска на основе ГБЗ

Реализация функции полнотекстового информационного поиска на основе ГБЗ требует формирования и использования специальной базы метаданных (так называемых поисковых индексов), обеспечивающей хранение вспомогательных данных для поиска в реальном режиме времени. В индексной базе данных каждый хранящийся документ представляется некоторым набором атрибутов, необходимых и достаточных для его идентификации в процессе поиска. Благодаря заранее формируемым и хранящимся в базе данных индексам возможно существенное сокращение времени доступа и обработки информации поисковой программой, что обеспечивает интерактивность процедуры поиска пользователем необходимого документа по соответствующему запросу. Как правило, индексы полнотекстовых поисковых систем состоят из двух основных частей. Первая часть индекса хранит информацию о ключевых (индексных) терминах того или иного документа и степени важности каждого из этих терминов для раскрытия содержания документа, которая выражается соответствующим весовым коэффициентом. Вторая часть индекса хранит более общие признаки документа, в том числе его название, адрес хранения, время создания или последней модификации, размер документа и т.д. Наиболее существенное влияние на качество индекса, а значит и на качество поиска в полнотекстовых поисковых системах, оказывает эффективность организации первой части индекса, отвечающего за выдачу релевантных запросу пользователя документов [10]. На рис. 6 представлена предлагаемая общая схема формирования поискового индекса для ГБЗ на основе семантического анализа результатов моделирования.

Исходными данными для формирования индексной базы по данному алгоритму являются, во-первых, регламентирующие документы, хранящиеся в гипертекстовой базе документов информационной системы. Во-вторых, данные, полученные в результате построение функциональной IDEF-модели процесса принятия решений для рассматриваемой предметной области.

Текстовые документы подвергаются стандартной процедуре лингвистического анализа с использованием инструментов статистического анализа и реферирования текстов. Отчет о результатах моделирования не является стандартным текстовым документом и представляется в виде таблицы, определяющей различные семантические элементы диаграмм функциональной модели IDEF0. Поэтому семантический анализ результатов моделирования проводится с помощью специального программного модуля, алгоритм обработки которого отличается от стандартного лингвистического процессора.

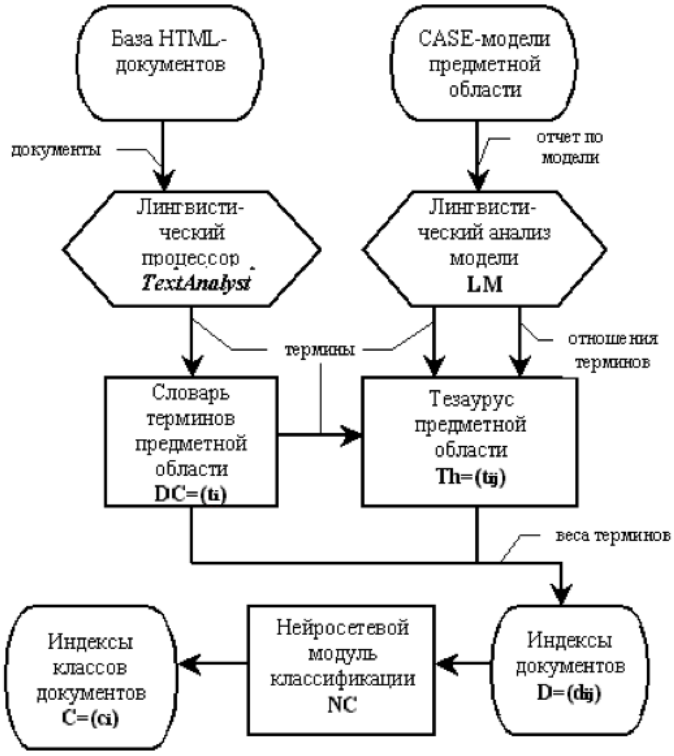

Рисунок 6 - Схема формирования структуры индексов
DOI: https://doi.org/10.60797/itech.2024.4.2.6

В результате процедуры лингвистического анализа исходных данных выделяются термины предметной области, которые заносятся в специальный электронный словарь $DC=(t_i)$. Отношения терминов, сформированные на основе модели предметной области, заносятся в тезаурус $Th=(t_{ij})$. На основе тезауруса и словаря затем формируются индексы документов $D=(d_{ij})$, в которых каждый документ определяется набором своих терминов из словаря предметной области с соответствующим весовым коэффициентом d_{ij} каждого термина в документе. Индекс классов определяет разбиение полного индекса документов на определенные категории семантически близких документов для организации 2-уровневой системы поиска. Классификация документов реализуется специальным нейросетевым модулем NC, способным к контролируемому обучению и обобщению предыдущего опыта. Разработанная по описанной схеме индексная структура является основой реализации информационного поиска регламентирующих документов в ГБЗ. Обобщенная схема реализации алгоритма 2-уровневого информационного поиска на основе тезауруса ГБЗ приведена на рис. 7.

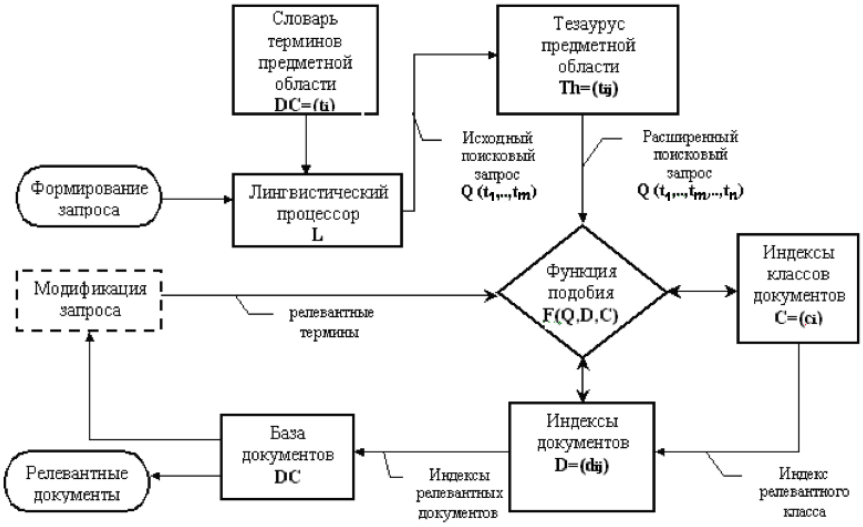

Рисунок 7 - Обобщенная схема информационного поиска с релевантной обратной связью
DOI: https://doi.org/10.60797/itech.2024.4.2.7

Пользователь системы формирует исходный естественно-языковый запрос на поиск требуемого регламентирующего документа. Запрос принимается лингвистическим процессором, который отсекает из предложения все служебные слова, остальные слова приводятся к нормальной форме и сопоставляются с терминами из словаря допустимых терминов **DC=*(ti)*** предметной области. Таким образом, первоначальный естественно-языковый запрос преобразуется в некоторый бинарный вектор запроса $\mathbf{Q}(t_1,..,t_m)$, формально определяющий образ исходного поискового запроса. Если пользователь затрудняется подобрать нужные термины при формулировании запроса, ему могут быть выданы нерелевантные (несоответствующие) документы. Во избежание этого исходный поисковый запрос дополняется новыми терминами из тезауруса **Th=*(t_{ij})*** предметной области, связанными с исходными терминами каким-либо из определенных типов отношений, что позволяет более точно определить контекст поиска. В результате формируется расширенный вектор поискового запроса $\mathbf{Q}(t_1,..,t_m,..,t_n)$, который определяет полное множество допустимых понятий, характеризующих семантику исходного поискового запроса пользователя.

В то же время в индексной базе данных хранятся взвешенные векторные поисковые образы каждого из документов $\boldsymbol{d_i} = (t_1,..,t_m,..,t_n)$ и образы классов документов $\mathbf{C_k} = (t_1,..,t_m,..,t_n)$. Поэтому основной функцией системы далее является определение образа документа $\boldsymbol{d_i}$, геометрически наиболее близкого к сформированному образу поискового запроса $\boldsymbol{q_j}$ в *N*-мерном пространстве множества терминов тезауруса ГБЗ. Эта процедура в теории информационного поиска называется определением меры сходства документов и запроса и в общем случае определяется реализуемой моделью поиска и используемой функцией подобия ***F(q,d)***. Наиболее простым примером функции подобия в векторной модели поиска является определение косинуса угла между соответствующими векторами запроса и документа:

$$F_{vw} = \frac{\sum v_i w_i}{\sqrt{\sum (v_i)^2 \sum (w_i)^2}}. \qquad (2)$$

Таким образом, чем ближе документы в пространстве, тем меньше косинус угла между ними. В предложенной модели 2-уровневого поиска такая мера подобия определяется дважды: сначала на вход нейросетевого модуля распознавания подается вектор расширенного поискового запроса, на основе которого определяется наиболее семантически близкий к поисковому запросу класс документов $\mathbf{C_k}$**.** На следующем этапе (уровне) вектор запроса сопоставляется с вектором каждого документа $\boldsymbol{d_{ik}}$ из «класса-победителя». Такая схема позволяет на первом же этапе локализовать релевантные документы внутри класса и проводить дальнейшую селекцию только внутри этого класса. Это существенно уменьшает время поиска документов в ГБЗ, при этом эффективность поиска сохраняется.

В результате завершения процедуры сопоставления вектора запроса с векторами соответствующих документов формируется некоторый список индексов документов, ранжированный по степени их релевантности поисковому запросу. Выбирается пороговое значение степени релевантности, информация об индексах документов, превышающих этот порог, передается в хранилище документов, где документы, соответствующие этим индексам, отбираются и выдаются в качестве ответа на запрос пользователя. По такой схеме реализуется прямая связь пользователя с поисковой системой.

Однако среди множества выданных документов нужного документа может не оказаться или его ранг может быть необоснованно занижен системой. Это в первую очередь связано с неточностью формулирования первоначального поискового запроса, когда пользователь в силу объективных причин «спрашивает не совсем то, что хочет получить». Поэтому одновременно с документами пользователю выдается множество всех терминов, использованных для идентификации всех релевантных документов. В результате пользователь имеет возможность переформулировать или уточнить первоначальный поисковый запрос, исключив «случайные» термины или включив более «удачные» термины из тезауруса системы. Переформулированный запрос обрабатывается системой и результаты поиска снова возвращаются пользователю. Как правило, эффективность поиска при этом оказывается существенно лучше предыдущего. Таким образом осуществляется процедура поэтапного улучшения качества поиска до получения приемлемых результатов по критериям полноты и точности.

В математическом смысле процесс интеллектуального поиска можно представить как нахождение вектора документа D_i, являющегося ближайшим к сформированному пользователем вектору запроса Q_i.

Все пространство поиска документов образует в этом случае поверхность N-мерной гиперсферы, разделенное на области C_k соответствующие классам документов (рис. 8).

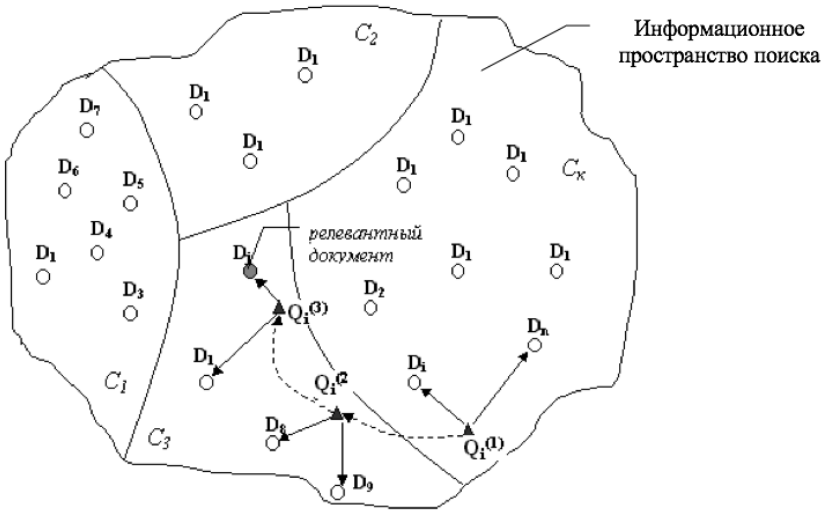

Рисунок 8 - Геометрическое представление процесса информационного поиска
DOI: https://doi.org/10.60797/itech.2024.4.2.8

В таком пространстве реализация поиска происходит следующим образом: пользователь формулирует свои информационные потребности с помощью некоторого множества выдаваемых ему системой терминов тезауруса предметной области. Поисковый алгоритм на первом этапе формирует на основе этого множества терминов первоначальный бинарный вектор запроса $Q_i^{(1)} = (t_1,\ldots,t_n)$ и подает его на вход обученной нейронной сети, которая распознает категорию документов (область C_i на рисунке), ближайшую для вектора запроса $Q_i^{(1)}$. На втором этапе поиска определяется мера близости каждого из документов класса к вектору запроса $Q_i^{(1)}$. Первые *k*-ближайших документа выдаются в качестве ответа на запрос, а также выдается множество всех терминов, которые определяют в тезаурусе эти *k* документов.

Если пользователь не удовлетворен результатами выдачи, на основе предоставленного ему множества терминов он формулирует новый запрос. Вектор нового запроса $Q_i^{(2)}$ имеет координаты, отличные от $Q_i^{(1)}$ и поэтому происходит его смещение в пространстве поиска в сторону области релевантных документов. Далее процедура поэтапного смещения вектора запроса повторяется до тех пор, пока очередной запрос не окажется ближайшим к необходимому (релевантному) пользователю документу. Такая процедура также может быть использована для обучения (перенастройки весов) тезауруса ГБЗ. Тогда, при формировании аналогичного запроса другим пользователем, система предоставит ему документ, который оценил, как релевантный предыдущий пользователь. Наличие тезауруса ГБЗ позволяет пользователю непосредственно участвовать в процессе регулирования и выборе правильного направления поиска необходимой ему информации.

Рассмотрим пример реализации поискового запроса в процессе решения одной для одной из задач стратегического планирования – оценку и отбор инвестиционных проектов для включения в стратегический план развития региона [11]. Пусть пользователь осуществляет функцию оценки эффективности реализуемых в рамках стратегии проектов и хочет узнать *«как рассчитать влияние инфляции на реализуемость проекта»*. При ручном поиске было установлено, что интересующая его информация содержится в документе *«Рекомендации по учету инфляции при оценке эффективности инвестиционного проекта»*. Проследим как будет развиваться процесс информационной поддержки пользователя. Допустим, он формирует запрос вида: *«инфляция, финансовая реализуемость»*. При поиске в «бестезаурусном режиме», нужный ему документ был определен под 8-м номером из 10-ти выданных системой ранжированного по близости запросу списка «релевантных» документов. При этом первые 5 документов были посвящены общим вопросам оценки финансовой реализуемости, 6 и 7 документы определяли общие понятие о видах инфляции и их влиянии на эффективность проекта в целом.

На первой итерации поиска с использованием тезауруса нужный документ был выдан под 6-м номером, поскольку система расширила первоначальный вектор запроса понятием *«прогноз»*, имеющим функциональную зависимость от термина *«инфляция»*. Кроме того, ему было предложено уточнить запрос с помощью терминов *«цена, темп, процент, индекс, дефлирование, прирост»*, которые связаны в тезаурусе соответственно с терминами «*инфляция, финансовая реализуемость, прогноз*».

На второй итерации поиска он расширил свой поиск следующим образом: *«инфляция, темп, финансы, реализуемость, цена, дефлирование, прогноз»*. В результате, нужный ему документ был выдан под 1-м номером в ранжированном списке. Для сравнения, нужный документ был проиндексирован в системе следующими терминами (приводятся без весов): *«прирост, инфляция, темп, финансовая реализуемость, цена, прогноз»*.

Сформированный пользователем на основе ГБЗ на второй итерации поисковый запрос оказался наиболее близким к вектору нужного документа. Однако, если сравнить этот вектор с его первоначальным запросом, это оказалось возможным только в результате тех уточнений, которые он ввел в свой исходный запрос, опираясь на подсказки системы ППР.

Таким образом, предложенная информационная система ППР на основе интегрированной в ее структуру ГБЗ предметной области позволяет работать специалисту и системе работать в едином пространстве терминов, тем самым позволяя, с одной стороны, ЛПР «правильно выразить» свои информационные потребности, а с другой, информационной системе, адекватно «понять» его потребности и обеспечить полноту и точность их удовлетворения за счет используемого механизма релевантной обратной связи.

Заключение

Таким образом, в рамках проведенного исследования для реализации интеллектуальной информационной поддержки процессами стратегического планирования разработана концептуальная схема организации интеллектуальной информационной поддержки принятия решений на основе ГБЗ. Определены основные этапы проектирования системы, включающей этапы системного моделирования процессов стратегического планирования, семантического анализа результатов моделирования и лингвистического анализа регламентирующих документов, формирования интегрированного тезауруса предметной области, а также этапов формирования структуры семантических категорий ГБЗ на основе нейросетевой классификации и разработки алгоритма интеллектуального поиска документов на основе тезауруса ГБЗ.

Рассмотрены основные этапы формирования структуры ГБЗ как основного модуля проектируемой системы ППР. Обоснован выбор и предложена модель формирования семантических категорий регламентирующих документов в ГБЗ на основе метода обучения искусственной нейронной сети. Классификацию предложено проводить в сформированном информационном пространстве терминов тезауруса предметной области, объединяя в единую категорию документы близкие по своему содержанию и функциональному назначению. Рассмотрена структура элементов формируемой ГБЗ, в которой множество документов образует семантические категории, путь к каждому документу на семантическом графе ГБЗ проходит через соответствующее подмножество терминов тезауруса и определяет индекс (координаты) этого документа в едином информационном пространстве терминов предметной области.

В соответствии с разработанной структурой ГБЗ предложена схема реализации ППР при управлении процессами стратегического планирования на основе алгоритма интеллектуального информационного поиска регламентирующих документов с использованием тезауруса предметной области. Предложенный алгоритм реализует расширенный семантический поиск документов в соответствии со сформированным информационным пространством терминов предметной области. Кроме того, реализация алгоритма поиска предполагает механизм релевантной обратной связи системы с пользователем в процессе формирования им своих информационных потребностей и оценке полноты и точности выданной системой информации. Использование тезауруса ГБЗ позволяет пользователю непосредственно участвовать в процессе регулирования и выбора правильного направления поиска необходимой ему информации. Рассмотрен фрагмент реализации процедуры поиска регламентирующего документа в ГБЗ на примере решения одной из задач в рамках процедуры стратегического планирования.

Финансирование

Исследование выполнено за счёт гранта Российского научного фонда, проект № 23-28-00871.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Funding

The research was carried out at the expense of a grant from the Russian Science Foundation, project № 23-28-00871.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы / References

1. Васильева Е.В. Адаптивное хранилище данных как технологический базис экосистемы банка / Е.В. Васильева, К.С. Солянов, Т.Д. Коневцева // Финансы: теория и практика. — 2020. — № 24(3). — С. 132–146. — DOI: 10.26794/2587.5671.2020.24.3.132.146.

2. Трофимов В.В. Организация и технология документационного обеспечения управления / В.В. Трофимов. — Москва : КНОРУС. — 2019.

3. Костин А.В. База знаний как инструментарий решения задач в экономической, социальной и производственной сферах / А.В. Костин, под ред. М.А. Ягольницера. — Новосибирск: ИЭОПП СО РАН. — 2023. — 160 с.

4. Храмов А.Г. Методы и алгоритмы интеллектуального анализа данных / А.Г. Храмов. — Самара: СамГУ. — 2019. — 176 с.

5. Яшина Н.Г. Онтологический инжиниринг в информационной науке (зарубежный опыт) / Н.Г. Яшина // Вестник Казанского государственного университета культуры и искусств. — 2015. — №1. — С. 94–97.

6. Сидоренко Э.Л. Эффективность цифрового государственного управления: теоретические и прикладные аспекты / Э.Л. Сидоренко, И.Н. Барциц, З.И. Хисамова // Вопросы государственного и муниципального управления. — 2019. — № 2. — С. 93–114.

7. Полетаева Н.Г. Классификация систем машинного обучения / Н.Г. Полетаева // Вестник Балтийского федерального университета им. И. Канта. Серия: Физико-математические и технические науки. — 2020. — № 1. — С. 5–22.

8. Ковалев М.В. Семантические модели и средства разработки искусственных нейронных сетей и их интеграции с базами знаний / М.В. Ковалев // Информатика. — 2023. — Т. 20. — № 3. — С. 90–105.

9. Низамутдинов М.М. Концептуальная модель прогнозирования влияния качества жизни населения на миграционные и демографические процессы / М.М. Низамутдинов, З.А. Давлетова // Экономика и управление: научно-практический журнал. — 2024. — № 1. — С. 150–155.

10. Суконщиков А.А. Разработка интеллектуальной системы поиска, обработки, генерации контекстной информации / А.А. Суконщиков, А.Н. Егоров, А.Н. Швецов // Вестник Череповецкого государственного университета. — 2023. — № 5 (116). — С. 75–86.

11. Морозова М.Е. Среднесрочное прогнозирование российской экономики с использованием когнитивной модели / М.Е. Морозова, В.В. Шмат // Проблемы прогнозирования. — 2017. — №. 3. — С. 19–25.

Список литературы на английском языке / References in English

1. Vasilyeva E.V. Adaptivnoe hranilishhe dannyh kak tehnologicheskij bazis jekosistemy banka [Adaptive data warehouse as a technological basis for a bank ecosystem] / E.V. Vasilyeva, K.S. Solyanov, T.D. Konevtseva // Finansy: teorija i praktika [Finance: Theory and Practice]. — 2020. — № 24(3). — P. 132–146. — DOI: 10.26794/2587.5671.2020.24.3.132.146. [in Russian]

2. Trofimov V.V. Organizacija i tehnologija dokumentacionnogo obespechenija upravlenija [Organization and technology of documentation support of management] / V.V. Trofimov. — Moscow: KNORUS. — 2019. [in Russian]

3. Kostin A.V. Baza znanij kak instrumentarij reshenija zadach v jekonomicheskoj, social'noj i proizvodstvennoj sferah [Knowledge base as a tool for solving problems in the economic, social and production spheres] / A.V. Kostin, edited by M.A. Yagolnitser. — Novosibirsk: Institute of Economics, Industrial Problems and Problems of the Siberian Branch of the Russian Academy of Sciences. — 2023. — 160 p. [in Russian]

4. Khramov A.G. Metody i algoritmy intellektual'nogo analiza dannyh [Methods and algorithms for data mining] / A.G. Khramov. — Samara: Samara State University. — 2019. — 176 p. [in Russian]

5. Yashina N.G. Ontologicheskij inzhiniring v informacionnoj nauke (zarubezhnyj opyt) [Ontological engineering in information science (foreign experience)] / N.G. Yashina // Vestnik Kazanskogo gosudarstvennogo universiteta kul'tury i iskusstv [Bulletin of the Kazan State University of Culture and Arts]. — 2015. — № 1. — P. 94–97. [in Russian]

6. Sidorenko E.L. Jeffektivnost' cifrovogo gosudarstvennogo upravlenija: teoreticheskie i prikladnye aspekty [Efficiency of digital public administration: theoretical and applied aspects] / E.L. Sidorenko, I.N. Bartsits, Z.I. Khisamova // Voprosy gosudarstvennogo i municipal'nogo upravlenija [Issues of public and municipal administration]. — 2019. — № 2. — P. 93–114. [in Russian]

7. Poletaeva N.G. Klassifikacija sistem mashinnogo obuchenija [Classification of machine learning systems] / N.G. Poletaeva // Vestnik Baltijskogo federal'nogo universiteta im. I. Kanta. Serija: Fiziko-matematicheskie i tehnicheskie nauki [Bulletin of the Immanuel Kant Baltic Federal University. Series: Physical, Mathematical and Technical Sciences]. — 2020. — № 1. — P. 5–22. [in Russian]

8. Kovalev M.V. Semanticheskie modeli i sredstva razrabotki iskusstvennyh nejronnyh setej i ih integracii s bazami znanij [Semantic models and tools for developing artificial neural networks and their integration with knowledge bases] / M.V. Kovalev // Informatika [Informatics]. — 2023. — Vol. 20. — № 3. — P. 90–105. [in Russian]

9. Nizamutdinov M.M. Konceptual'naja model' prognozirovanija vlijanija kachestva zhizni naselenija na migracionnye i demograficheskie processy [Conceptual model for forecasting the impact of the quality of life of the population on migration and demographic processes] / M.M. Nizamutdinov, Z.A. Davletova // Jekonomika i upravlenie: nauchno-prakticheskij zhurnal [Economics and Management: Scientific and Practical Journal]. — 2024. — № 1. — P. 150–155. [in Russian]

10. Sukonshchikov A.A. Razrabotka intellektual'noj sistemy poiska, obrabotki, generacii kontekstnoj informacii [Development of an intelligent system for searching, processing, and generating contextual information] / A. A. Sukonshchikov, A. N. Egorov, A.N. Shvetsov // Vestnik Cherepoveckogo gosudarstvennogo universiteta [Bulletin of Cherepovets State University]. — 2023. — № 5 (116). — P. 75–86. [in Russian]

11. Morozova M.E. Srednesrochnoe prognozirovanie rossijskoj jekonomiki s ispol'zovaniem kognitivnoj modeli [Medium-term forecasting of the Russian economy using a cognitive model] / M.E. Morozova, V.V. Shmat // Problemy prognozirovanija [Problems of forecasting]. — 2017. — № 3. — P. 19–25. [in Russian]