УДК 004.9

НЕЯВНЫЕ ЗНАНИЯ В ИНФОРМАЦИОННОМ ПОИСКЕ

Курдюков Никита Сергеевич[1],
e-mail: nskurdyukov@gmail.com,
[1]Российский технологический университет (РТУ МИРЭА), г. Москва, Россия

В статье исследуются неявные знания в информационном поиске. Существует тенденция к увеличению объемов информации, в том числе в информационных сетях. Эта тенденция мотивирует исследования в области поиска информации. Рост объемов информации влечет рост скрытых знаний и информационной неопределенности. Дана таксономия причин неадекватности поиска в информационной сети. Эти причины делятся на объективные и когнитивные. Показана разница между морфологическим и семантическим поиском. Предложена теоретико-множественная модель описания поиска информации и его результатов. Описаны три основных типа поиска информации. Показаны разница между поиском информации и поиском знаний; качественная разница между онтологическим и традиционным информационным поиском; причины появления неявного знания в практической деятельности и научных исследованиях. Дано формальное описание онтологического поиска, включающего компонент неявного знания. Показаны критерии наличия неявного знания, которое можно получить посредством онтологического поиска. Результаты данного исследования способствуют развитию методов поиска информации, онтологического поиска и выявления неявных знаний.

Ключевые слова: информационный поиск, информационный онтологический поиск, теоретико-множественные модели, неявное знание, морфологический поиск, семантический поиск

IMPLICIT KNOWLEDGE IN INFORMATION RETRIEVAL

Kurdyukov N.S.[1],
e-mail: nskurdyukov@gmail.com,
[1]Russian Technological University (RTU MIREA), Moscow, Russia

The article explores implicit knowledge in information retrieval. There is a tendency to increase the volume of information, including in information networks. This trend motivates research in the field of information retrieval. An increase in the volume of information leads to an increase in hidden knowledge and information uncertainty. The taxonomy of the reasons for the inadequacy of the search in the information network is given. These reasons are divided into objective and cognitive. The difference between morphological and semantic search is shown. A set-theoretic model for describing the search for information and its results is proposed. Three main types of information retrieval are described. The difference between the search for information and the search for knowledge is shown. The qualitative difference between ontological and traditional information search is shown. The reasons for the appearance of implicit knowledge in practical activities and scientific research are shown. A formal description of the ontological search, including a component of implicit knowledge, is given. The criteria for the presence of implicit knowledge, which can be obtained through an ontological search, are shown. The results of this study contribute to the development of information retrieval methods, ontological search and the identification of implicit knowledge.

Keywords: information search, information otological search, set-theoretic models, implicit knowledge, morphological search, semantic search

DOI 10.21777/2500-2112-2024-1-80-87

Введение

Объем данных в мире и в Интернете, в частности, растет в геометрической прогрессии [1]. Рост информации опережает рост числа методов извлечения нужной информации. Это отражает проблему обработки больших данных [2; 3], делает актуальным исследования в этой области и разработку методов извлечения информации. Информационный поиск является основным методом извлечения информации [4] в сетях и в информационных системах хранения информации. Можно ввести понятие «информационный сетевой поиск» для выделения технологий информационного поиска в сетевых системах.

Информационный сетевой поиск не дает адекватных результатов по ряду причин.

Первая причина – полисемия (многозначность, многовариантность) поисковых образцов. Она состоит в том, что поисковые паттерны могут иметь разные значения и быть синонимами. Например, «лук». Это растение и орудие для стрельбы. В результате полисемии результаты поиска могут содержать неопределенность и быть нерелевантными.

Вторая причина есть информационная неопределенность разных видов. Информационная неопределенность существует в информационном множестве, в котором проводят поиск. Информационная неопределенность характеризует состояние объекта, который выполняет поиск. Во всех научных исследованиях информационная неопределенность выражается в том, что искомая информация всегда известна только приближенно. Научные исследования характеризует информационная неопределенность [5], которую необходимо уменьшить или устранить. Информационная неопределенность в условиях поиска приводит к неточности или к нерелевантности результатов поиска.

Третья причина неточности поиска является технической. Она связана со временем поиска. Чем длиннее паттерн запроса, тем длительнее время поиска. Чем сложнее паттерн запроса, тем длительнее время поиска. В данном случае сложность определяется числом слов. При нескольких словах возникают комбинаторные задачи сочетаний и перестановок, которые увеличивают время анализа и поиска. Короткий запрос уменьшает время поиска, но увеличивает объем результатов поиска. Длинный запрос комбинаторно увеличивает время поиска, но сокращает объем результатов поиска. На практике средний размер веб-поиска составляет 2–4 слова.

Четвертая причина неточности информационного поиска является когнитивной. Она может быть обусловлена неуверенностью и недостаточной компетентностью пользователя. В этих условиях пользователь часто не уверен в том, что он ищет, пока не увидит результаты. Даже если пользователь знает, что он ищет, он не всегда знает, как правильно семантически составить соответствующий запрос.

Пятая причина заключается в плохом знании родного языка и, как следствие, некорректном использовании терминологических отношений [6] и плохом использовании вспомогательных терминов. Неправильное использование отношений искажает смысл и делает результаты информационного поиска неточными.

Шестая причина неадекватности поиска в том, что широко применяют информационный поиск, который является морфологическим [7]. Пользователь задает морфологический паттерн, смысл которого может отличаться от семантической формы запроса формы. В результате поиска формируется информационное множество по морфологическим, а не по смысловым признакам. Например, специалист в области геодезии задает поисковое слово «триангуляция», имея в виду методы построения геодезической сети. Однако триангуляция имеет множество значений:

– триангуляция в архитектуре есть метод нахождения пропорций и отношений частей здания с помощью треугольников;

– в геодезии триангуляцией называют метод создания геодезической сети на основе треугольников;

– в радиолокации триангуляция есть метод радиопеленгации;

– в математике триангуляция есть процесс разбиения сложной геометрической фигуры на треугольники (симплексы);

– триангуляция в общественных науках – применение трёх методов для измерения одного показателя для надежности результата;

– триангуляция в торговле – правило трех цен;

– триангуляция в психологии (политологии) есть метод манипулирования сознанием субъекта через промежуточного субъекта (авторитет).

В итоге информационный морфологический поиск такого «геодезиста» даст массу результатов, не соответствующих его информационной потребности. Поэтому логичней вместо одного слова задать сложный семантический паттерн «Метод построения геодезических сетей с использованием треугольников». В этом случае результаты поиска будут релевантными.

Можно продолжить перечисление причин неадекватности информационного поиска. Но следует вывод, что эти причины существуют и мотивируют совершенствование существующих методов поиска и разработку новых методов информационного поиска.

Информационный поиск (*information retrieval – IR*) не является однородной технологией [8]. Можно выделить разные виды поиска. Различают информационный поиск сведений (информации), который сводится к нахождению блоков информации или референций (ссылок). Этот поиск есть морфологический и семантический. Его применяют при сборе информации для выполнения аналитических исследований и обзоров. Этот поиск можно классифицировать как технологию «извлечения информации» [9]. Различают информационный поиск в научных исследованиях. Этот поиск направлен на нахождение новых знаний, научных решений или на поиск условий получения знаний и научных решений. Этот поиск является семантическим и онтологическим. Этот поиск можно классифицировать как технологию «извлечения знаний» [10]. Выделяют информационный поиск в патентных исследованиях. Этот поиск направлен на нахождение известных аналогов предлагаемого изобретения. Он направлен на нахождение блоков информации по паттерну. Этот поиск является семантическим.

Важным показателем IR является информационная потребность пользователя, выраженная в форме информационного запроса. Это тема, о которой пользователь хочет знать больше. Ее следует отличать от запроса, т.е. от того, что пользователь вводит в строку поиска информационно-поисковой системы (ИПС). Например, при написании диссертации у аспиранта существует информационная потребность в решении новой научной задачи. Если он на основе анализа найденных научных источников такое решение для себя нашел, то возникает дополнительная информационная потребность проверки его на новизну и на корректность методов решения. Информационные потребности различаются по типам перечисленных выше задач IR.

Технология поиска возникла в области прикладной информатики [11]. IR классифицируют как специализированную информационную технологию. Для реализации IR применяют специализированные ИПС. Преимущество использования глобальных ИПС заключается в том, что они индексируют контент сети Интернет. С другой стороны, базируясь на традиционной модели информационного поиска, они требуют формулирования информационной потребности в виде списка ключевых слов [12].

Анализ публикаций в области информационного поиска свидетельствует о тенденции диверсификации методов поиска информации. Однако значительная часть работ не является целостной. Это обусловлено тем, что в работах отсутствуют четкие требования к выявлению доказательства истинности найденной информации. По существу большинство технологий IR сводится к формальной схеме совпадений морфологических форм. Поэтому обобщение методов и развитие подходов к информационному поиску в настоящее время остаются актуальными задачами.

Целью работы является формальное описание онтологического информационного поиска на основе выделенных критериев наличия неявного знания в исходной информации запросов и таксономии причин появления неявного знания в результатах поиска.

Теоретико-множественный подход к описанию моделей информационного поиска

Можно описать модели поиска с использованием теории множеств. На рисунке 1 дана принципиальная схема информационного поиска в обозначениях теории множеств.

Основой поиска служит информационная потребность, которая на схеме не показана. На основе информационной потребности формируют поисковое множество или в простейшей интерпретации поисковый паттерн (*pat*). Поиск производят в исходном информационном множестве (*information set – IS*). Результатом поиска является множество результатов поиска RIR. Поисковое множество вступает в

информационное взаимодействие (*II1*) с исходным информационным множеством. Информационное множество находится во взаимодействии (*II2*) с RIR. Первое информационное взаимодействие (*II1*) обусловлено и протекает по условиям организации поискового множества. Второе информационное взаимодействие (*II2*) обусловлено и протекает по алгоритмам ИПС [13]. Это два качественно разных взаимодействия в информационном поиске. Во множестве результатов поиска можно выделить три варианта. Они приведены на рисунке 2.

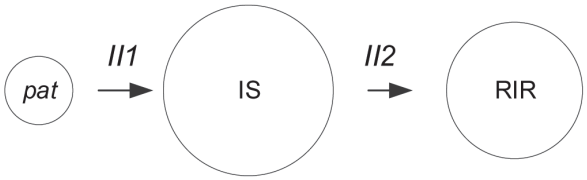

Рисунок 1 – Теоретико-множественная схема IR

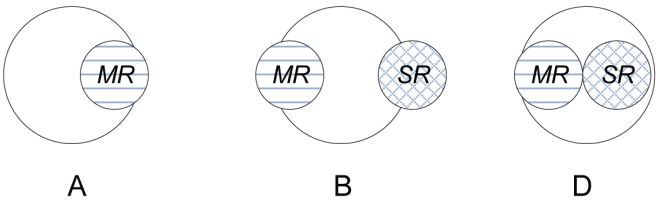

Рисунок 2 – Три варианта результатов IR

Результат поиска, не содержащий релевантные данные, не приведен на рисунке 2. На рисунке 2A показан результат поиска, который называют формальной релевантностью или морфологической релевантностью. На рисунке 2B приведен результат поиска, который называют смысловой релевантностью или частичной семантической релевантностью. На рисунке 2D приведен результат поиска, который называют полной релевантностью, или пертинентностью [11]. Символами обозначены MR – морфологическое соответствие (множество) и SR – семантическое соответствие (множество).

Для формальной релевантности RIRA (рисунок 2A) имеют место отношения

$$RIRA = RIR \cup MR;\ RIRA \cap MR \rightarrow MR. \quad (1)$$

Для частичной смысловой релевантности RIRB с учетом (1) имеют место отношения

$$RIRB = RIR \cup MR \cup SR;\ RIRA \cap SR \neq \varnothing. \quad (2)$$

Из выражения (2) следует, что существует частичное семантическое соответствие между результатами поиска и семантическим множеством, которое удовлетворяет информационную потребность пользователя. Расхождение между ними не является пустым множеством, а реально существует.

Для полной релевантности RIRD имеют место отношения

$$RIRD = RIR \cup MR \cup SR;\ RIRA \cap SR = \varnothing. \quad (3)$$

Из выражения (3) следует, что существует полное семантическое соответствие между результатами поиска и семантическим множеством, которое удовлетворяет информационную потребность пользователя.

Неявные знания и онтологический поиск

Необходимо разграничивать неявные знания и неявную информацию [14]. Например, при статистических исследованиях тренд, который можно вычислить по статистическому ряду, является неявной информацией. Объяснение тренда и закономерность, которую оно отражает, есть неявное знание. Неявная информация может содержать неявное знание.

Причины появления неявных знаний в IR разнообразны. Основная причина в том, что исходное информационное множество содержит неявные знания. IR в сфере научных исследований направлен

на нахождение неявных знаний. Причина появления неявных знаний состоит в неинформированности пользователя и организации некорректного поискового запроса. Причина появления неявных знаний состоит в незнании пользователем области, в которой он ведет информационный поиск. Причина появления неявных знаний состоит также в неумении интерпретировать результаты поиска, протекании скрытых (латентных) процессов, которые меняют ситуацию исследования. В этом аспекте информационный поиск есть инструмент выявления неявных знаний.

Рассмотрим понятие «онтологический информационный поиск» (*OIR*) [15; 16]. Этот поиск направлен на поиск знаний с применением инструментов онтологии. Применение инструментов онтологии позволяет реализовать имитацию логических рассуждений и возможность автоматически классифицировать и связывать информацию. Это позволяет сузить область поиска решения и соответственно уменьшить время поиска. На рисунке 3 приведена ситуация онтологического информационного поиска.

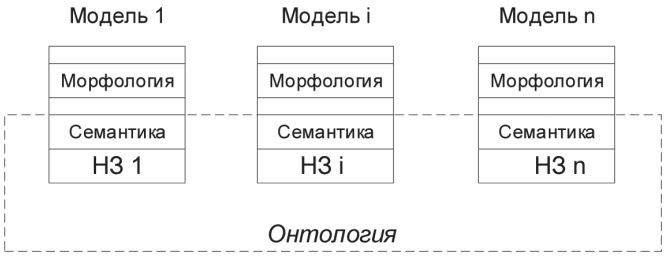

Рисунок 3 – Ситуация онтологического информационного поиска

Принципиальным отличием от IR в OIR является то, что в первом случае ищется некая модель, удовлетворяющая определенным требованиям и информационным потребностям пользователя. При онтологическом поиске частности (морфология) исключаются. При онтологическом поиске исследуется ряд моделей и находится некое обобщение. Для этой цели наиболее подходящей процедурой является метамоделирование [17] как обобщение.

При отсутствии неявных знаний существуют две цепочки отношений для моделей (рисунок 3), которые получены в IR. Они связаны с причиной (*cause*), семантикой (*semantics*), следствием (*effect*). Для каждой модели на рисунке 3 при отсутствии неявных знаний имеет место прямая импликация

$$\text{Cause} \rightarrow \text{semantics} \rightarrow \text{effect} \quad (4)$$

и обратная импликация

$$\text{Cause} \leftarrow \text{semantics} \leftarrow \text{effect}. \quad (5)$$

Выражение (4) есть выдвижение гипотезы. Выражение (5) есть проверка гипотезы. Выражения (4) и (5) справедливы для всех моделей (рисунок 3) при отсутствии неявного знания.

При наличии неявного знания фактически будет действовать цепочка, включающая неявное знание (*tacit knowledge*). В этом случае в реальности будут действовать отношения

$$\text{Cause}_i \rightarrow \text{semantics} \wedge \text{tacit knowledge} \rightarrow \text{effect}. \quad (6)$$

Не имея информации о неявном знании, пользователь будет моделировать ее по схеме (4) вместо схемы (6) для каждой модели. Соответственно проверку он попытается сделать по схеме (5). Но при этом будет получаться результат

$$\text{Cause}_i^* \leftarrow \text{semantics} \leftarrow \text{effect}. \quad (7)$$

В выражении (7) Cause_i^* есть «обратный результат» для каждой модели i и для текущей модели. Для такой ситуации будет иметь место следующий эффект:

$$\text{Cause}_i^* \neq \text{Cause}_i. \quad (8)$$

Выражение (8) имеет место для одной i-ой модели. Оно означает, что исходная причина процесса не совпадает с моделируемой причиной, т.е.

$$\text{Cause}_i^* \neq \text{Cause}_j^*. \quad (9)$$

Выражение (9) имеет место для разных i-ой и j-ой моделей. Оно означает, что исходные одинаковые причины разных процессов не совпадают с моделируемыми причинами разных процессов.

Выражения (8) и (9) являются одними из признаков наличия неявных знаний. Онтологический поиск выделяет общность [15; 16]. Он производится не по одной модели, а по нескольким. Если в результате поиска выявляется, что в результатах поиска получают две части: RIR(*int*) – интерпретируемая и RIR(*nint*) – не интерпретируемая, то не интерпретируемая часть является индикатором наличия неявных знаний и требует дальнейшего анализа по их выявлению.

Таким образом, традиционный IR ищет информацию или знания как объект или феномен. Онтологический поиск осуществляет обобщение объектов и отношений между ними или формирует спецификацию концептуализации [18]. Онтологический поиск основан на семантическом поиске, соответствии смысла и теории соответствия [19]. В силу этого онтологический информационный поиск позволяет выявлять неявные знания.

Заключение

Проблемой всех видов поиска является сложность поиска, время поиска и объем результатов поиска. Для онтологического поиска эти факторы возрастают, что является препятствием его применения. В то же время как инструмент обобщения и поиска знаний он остается наиболее эффективным методом. Результаты онтологического информационного поиска и морфологического информационного поиска различаются. В морфологическом и семантическом поиске находят объекты и модели. В онтологическом информационном поиске находят обобщения, метамодели, концепции, знания, отношения. Все обобщения описывают группу моделей, процессов или объектов. Онтологический поиск можно рассматривать как один из инструментов нахождения неявных знаний. Подводя итог, следует констатировать: онтологический информационный поиск является перспективным направлением в теории информационного поиска и требует дальнейших исследований в части разработки критериев оценки результатов онтологического поиска. Пока они оцениваются только на когнитивном уровне.

Список литературы

1. *Azad H.K., Deepak A.* Query Expansion Techniques for Information Retrieval: a Survey // Information Processing & Management. – 2019. – T. 56, No. 5. – C. 1698–1735.
2. *Лёвин Б.А., Цветков В.Я.* Информационные процессы в пространстве «больших данных» // Мир транспорта. – 2017. – Т. 15, № 6 (73). – С. 20–30.
3. *Hariri R.H., Fredericks E.M., Bowers K.M.* Uncertainty in Big Data Analytics: Survey, Opportunities, and Challenges // Journal of Big Data. – 2019. – Vol. 6, No. 1. – P. 1–16.
4. *Guo J., Fan Y., Pang L., Yang L., Ai Q., Zamani H. & Cheng X.* A Deep Look into Neural Ranking Models for Information Retrieval // Information Processing & Management. – 2020. – No. 57 (6). – DOI 10.1016/j.ipm.2019.102067.
5. *Номоконова О.Ю.* Информационная неопределенность в информационном взаимодействии // Славянский форум. – 2017. – № 1 (15). – С. 104–110.
6. *Тихонов А.Н., Иванников А.Д., Цветков В.Я.* Терминологические отношения // Фундаментальные исследования. – 2009. – № 5. – С. 146–148.
7. *Курдюков Н.С.* Семантика и морфология информационного поиска // Славянский форум. – 2023. – № 4 (42). – С. 46–56.
8. *Zamani H., Dumais S.T., Craswell N., Bennett P.N., and Lueck G.* Generating Clarifying Questions for Information Retrieval // Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan. ACM, New York. – New York, 2020. – 11 p. – DOI 10.1145/ 3366423.338012.
9. *Landolsi M.Y., Hlaoua L., Ben Romdhane L.* Information Extraction from Electronic Medical Documents: State of the Art and Future Research Directions // Knowledge and Information Systems. – 2023. – Vol. 65, No. 2. – P. 463–516.
10. *Zhang Y.* et al. Unleashing the Power of Knowledge Extraction from Scientific Literature in Catalysis // Journal of Chemical Information and Modeling. – 2022. – Vol. 62, No. 14. – P. 3316–3330.
11. *Поляков А.А., Цветков В.Я.* Прикладная информатика. – Москва: Янус-К, 2002. – 392 с.

12. *Ахмадеева И.Р.* Использование онтологии для автоматизации поиска научной информации в сети Интернет // Информационные и математические технологии в науке и управлении. – 2018. – № 4 (12). – С. 42–49. – DOI 10.25729/2413-0133-2018-4-04.
13. *Курдюков Н.С.* Алгоритмическая логистика // Славянский форум. – 2024. – № 1 (43). – С. 154–164.
14. *Цветков В.Я.* Неявное знание и его разновидности // Вестник Мордовского университета. – 2014. – Т. 24, № 3. – С. 199–205.
15. *Sharma A., Kumar S.* Machine Learning and Ontology-Based Novel Semantic Document Indexing for Information Retrieval // Computers & Industrial Engineering. – 2023. – Vol. 176. – P. 108940.
16. *Kurdukov N.S.* Ontologies in Information Retrieval // European Journal of Technology and Design. – 2023. – No. 11 (1). – P. 9–14. – DOI 10.13187/ejtd.2023.1.9.
17. *Цветков В.Я., Булгаков С.В., Титов Е.К., Рогов И.Е.* Метамоделирование в геоинформатике // Информация и космос. – 2020. – № 1. – С. 112–119.
18. *Голомазов Д.Д.* Методы и средства управления научной информацией с использованием онтологий: дис. … канд. физ.-мат. наук: 05.13.17. – Москва, 2012. – 183 с.
19. *O'Connor D.J.* The Correspondence Theory of Truth. – Routledge, Hutchinson: University Library; London, 2021. – 146 p.

References

1. *Azad H.K., Deepak A.* Query Expansion Techniques for Information Retrieval: a Survey // Information Processing & Management. – 2019. – T. 56, No. 5. – S. 1698–1735.
2. *Lyovin B.A., Cvetkov V.Ya.* Informacionnye processy v prostranstve «bol'shih dannyh» // Mir transporta. – 2017. – T. 15, № 6 (73). – S. 20–30.
3. *Hariri R.H., Fredericks E.M., Bowers K.M.* Uncertainty in Big Data Analytics: Survey, Opportunities, and Challenges // Journal of Big Data. – 2019. – Vol. 6, No. 1. – P. 1–16.
4. *Guo J., Fan Y., Pang L., Yang L., Ai Q., Zamani H. & Cheng X.* A Deep Look into Neural Ranking Models for Information Retrieval // Information Processing & Management. – 2020. – No. 57 (6). – DOI 10.1016/j.ipm.2019.102067.
5. *Nomokonova O.Yu.* Informacionnaya neopredelennost' v informacionnom vzaimodejstvii // Slavyanskij forum. – 2017. – № 1 (15). – S. 104–110.
6. *Tihonov A.N., Ivannikov A.D., Cvetkov V.Ya.* Terminologicheskie otnosheniya // Fundamental'nye issledovaniya. – 2009. – № 5. – S. 146–148.
7. *Kurdyukov N.S.* Semantika i morfologiya informacionnogo poiska // Slavyanskij forum. – 2023. – № 4 (42). – S. 46–56.
8. *Zamani H., Dumais S.T., Craswell N., Bennett P.N., and Lueck G.* Generating Clarifying Questions for Information Retrieval // Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan. ACM, New York. – New York, 2020. – 11 p. – DOI 10.1145/ 3366423.338012.
9. *Landolsi M.Y., Hlaoua L., Ben Romdhane L.* Information Extraction from Electronic Medical Documents: State of the Art and Future Research Directions // Knowledge and Information Systems. – 2023. – Vol. 65, No. 2. – P. 463–516.
10. *Zhang Y.* et al. Unleashing the Power of Knowledge Extraction from Scientific Literature in Catalysis // Journal of Chemical Information and Modeling. – 2022. – Vol. 62, No. 14. – P. 3316–3330.
11. *Polyakov A.A., Cvetkov V.Ya.* Prikladnaya informatika. – Moskva: Yanus-K, 2002. – 392 s.
12. *Ahmadeeva I.R.* Ispol'zovanie ontologii dlya avtomatizacii poiska nauchnoj informacii v seti Internet // Informacionnye i matematicheskie tekhnologii v nauke i upravlenii. – 2018. – № 4 (12). – S. 42–49. – DOI 10.25729/2413-0133-2018-4-04.
13. *Kurdyukov N.S.* Algoritmicheskaya logistika // Slavyanskij forum. – 2024. – № 1 (43). – S. 154–164.
14. *Cvetkov V.Ya.* Neyavnoe znanie i ego raznovidnosti // Vestnik Mordovskogo universiteta. – 2014. – T. 24, № 3. – S. 199–205.
15. *Sharma A., Kumar S.* Machine Learning and Ontology-Based Novel Semantic Document Indexing for Information Retrieval // Computers & Industrial Engineering. – 2023. – Vol. 176. – P. 108940.
16. *Kurdukov N.S.* Ontologies in Information Retrieval // European Journal of Technology and Design. – 2023. – No. 11 (1). – P. 9–14. – DOI 10.13187/ejtd.2023.1.9.

17. *Cvetkov V.Ya., Bulgakov S.V., Titov E.K., Rogov I.E.* Metamodelirovanie v geoinformatike // Informaciya i kosmos. – 2020. – № 1. – S. 112–119.
18. *Golomazov D.D.* Metody i sredstva upravleniya nauchnoj informaciej s ispol'zovaniem ontologij: dis. … kand. fiz.-mat. nauk: 05.13.17. – Moskva, 2012. – 183 s.
19. *O'Connor D.J.* The Correspondence Theory of Truth. – Routledge, Hutchinson: University Library; London, 2021. – 146 r.