УДК 025.4:004.912
DOI: 10.25559/SITITO.019.202303.684-695

Оригинальная статья

Комплексный сетевой алгоритм формирования глоссария контекстно-близких прогностических терминов

О. Р. Попов[1*], А. Гросу[2], С. О. Крамаров[2]

[1] МОО «Академия информатизации образования», г. Москва, Российская Федерация
Адрес: 109029, Российская Федерация, г. Москва, ул. Нижегородская, д. 32
[*] cs41825@aaanet.ru

[2] БУ ВО «Сургутский государственный университет», г. Сургут, Российская Федерация
Адрес: 628400, Российская Федерация, Ханты-Мансийский автономный округ – Югра, г. Сургут, пр. Ленина, д. 1

Аннотация

Сбор словаря терминов, составляющего ознакомительное проявление концепций предметной области, является одним из первых шагов к моделированию определенной области знаний. В условиях конвергентных тенденций «стыковых» междисциплинарных связей при развитии сложных систем приоритетное значение приобретает сфера моделирования информационно-коммуникационных технологий (ИКТ) и компьютерных наук. При формировании глоссария прогностических терминов применен комплексный алгоритмический подход, согласно которому интегрирован ряд условий, объединяющих возможности сетевого (графового) и семантического подходов: автоматическая генерация графов, учет ранжирования при оценке результатов поиска, контекстно-семантическая фильтрация. В результате разработан комплексный алгоритм и программный код, позволяющий формировать на базе сетевого сервиса «Википедия» глоссарий контекстно-близких специализированных терминов и тематических словосочетаний от изначально заданных терминов с ранжированием по средней арифметической оценке двух алгоритмов – PageRank и HITS. Визуализация работы алгоритма представлена на примере генерации графа от первичного термина «Quantum computing». Проанализированы данные, обосновывающие объективность представленного подхода к оценке веса термина, а также демонстрирующие результат работы алгоритма на примере расширения контекста прогностических терминов в рамках категории «Computing engineering». В качестве финальной демонстрации приведён вывод фрагмента глоссария, структурированного по категориям прогностических ИКТ. Результаты исследования будут использованы как базовый корпус знаний предметной области, необходимый для формирования обоснованных формул запросов при последующем анализе тематических статей, размещенных в библиографических базах данных и внешних сетевых ресурсах.

Ключевые слова: глоссарий, алгоритм, граф, семантика, прогностический термин, категория, информационно-коммуникационные технологии

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Для цитирования: Попов О. Р., Гросу А., Крамаров С. О. Комплексный сетевой алгоритм формирования глоссария контекстно-близких прогностических терминов // Современные информационные технологии и ИТ-образование. 2023. Т. 19, № 3. С. 684-695. https://doi.org/10.25559/SITITO.019.202303.684-695

Original article

Complex Network Algorithm for Glossary Formation Context-Related Predictive Terms

O. R. Popov[a*], A. Grosu[b], S. O. Kramarov[b]
[a] Academy of Informatization of Education, Moscow, Russian Federation
Address: 32 Nizhegorodskaya St., Moscow 109029, Russian Federation
[*] cs41825@aaanet.ru
[b] Surgut State University, Surgut, Russian Federation
Address: 1 Lenin Ave., Surgut 628400, Khanty-Mansi Autonomous Okrug – Ugra, Russian Federation

Abstract
This article describes the process of creating a glossary of terms for a specific domain, which is the initial step in knowledge modeling. In the context of converging trends and interdisciplinary connections in the development of complex systems, particular emphasis is placed on modeling information and communication technologies (ICT) and computer science. To form the glossary of prognostic terms, a comprehensive algorithmic approach was applied, integrating a range of conditions that combine the capabilities of network (graph-based) and semantic approaches. This approach includes automatic graph generation, considering ranking in the evaluation of search results, and context-semantic filtering. As a result, a comprehensive algorithm and software code were developed, allowing the creation of a glossary of contextually related specialized terms and thematic phrases based on the "Wikipedia" network service. These terms were ranked using the average score of two algorithms — PageRank and HITS. The algorithm's operation was visualized using the example of generating a graph from the primary term "Quantum computing". Data were analyzed to justify the objectivity of the proposed term weighting approach and to demonstrate the algorithm's results in expanding the context of prognostic terms within the category of "Computing engineering." A fragment of the structured glossary of ICT is presented as a final demonstration. The results of this research will be used as a foundational knowledge corpus necessary for formulating well-grounded queries when analyzing thematic articles located in bibliographic databases and external network resources.

Keywords: glossary, algorithm, graph, semantics, prognostic term, category, information and communication technologies

Conflict of interests: The authors declare no conflict of interest.

For citation: Popov O.R., Grosu A., Kramarov S.O. Complex Network Algorithm for Glossary Formation Context-Related Predictive Terms. *Modern Information Technologies and IT-Education.* 2023;19(3):684-695. https://doi.org/10.25559/SITITO.019.202303.684-695

Введение

Сбор словаря терминов, составляющего ознакомительное проявление концепций предметной области, является одним из первых шагов к моделированию определенной области знаний. Средства поддержки и создания таксономий (taxonomies), тезаурусов (thesauri) и глоссариев (glossary) выступают одним из элементов технологий Text Mining — интеллектуальным инструментом анализа неструктурированных текстов. В совокупности они представляют комплекс научно обоснованных решений, обеспечивающих создание программной платформы автоматического извлечения, анализа и обработки информации из сетей знаний.

Достоверная терминологическая база, охватывающая широкий спектр контекстно-связанных терминов, необходима для формирования обоснованных формул запросов при последующем анализе тематических статей, размещенных в библиографических базах данных и внешних сетевых ресурсах. Для расширения запроса при предметно-ориентированном информационном поиске используются различные варианты наименования соответствующего понятия или дескрипторы, построенные на основе изначально заданного термина.

В современном научном контексте проблема обусловлена несовершенством существующих методов составления тезаурусов узкоспециализированных терминов в аспекте моделирования и освоения новых предметных областей [1, 2], включая автоматизацию [3] и многоязычность [4]. Целевая разработка и апробация комплексного алгоритмического подхода к построению глоссария предметно-ориентированных терминов предоставляет инструмент, который позволит заполнить этот пробел и улучшить качество аналитических исследований в области технологического прогнозирования.

В условиях конвергентных тенденций «стыковых» междисциплинарных связей при развитии сложных систем обосновано приоритетное значение информационно-коммуникационных технологий (ИКТ) и сферы компьютерных наук[1] [5]. Это положение предопределило основной выбор предметной области знаний для данного исследования.

Инструмент и методы

В литературе описаны методы автоматического извлечения технических и специализированных терминов из хранилищ документов [6, 7]. Отдельно следует выделить задачи выявления семантических связей и ассоциаций, новых терминов, возникшие в ходе научных исследований, для извлечения которых применяются методы эвристики. Используются различные подходы — статистический, вероятностный, семантический, графовый, нейросетевой и т. д., а также их комбинации, на основе которых реализуются алгоритмы извлечения и систематизации терминов и ключевых слов [8].

На практике достаточно часто применяются комбинированные методы, которые объединяют несколько теоретических подходов.

Классическими являются методы, основанные на статистико-вероятностном и лингвистическом анализе текста [9, 10]. Количественные методы базируются на статистических показателях для определения частоты/вероятности или закономерностей совпадения взаимосвязей между терминами. Лингвистические критерии в первую очередь учитывают грамматическую структуру терминов и коллокаций, которая может быть представлена в виде грамматического (синтаксического) образца, по которому распознаются извлекаемые терминологические кандидаты, т. е. синтаксически правдоподобные именные фразы. После фильтрации потенциальных терминов методами статистики, из-за их достаточно высокой специфичности они становятся информационным ресурсом для поддержки создания онтологии предметной области или терминологической базы.

При составлении словаря прогностических социальных терминов использован количественный анализ текста, позволяющий классифицировать документы на категории и определить, сколько документов в корпусе принадлежит каждой категории, представляющих архетипы сценариев будущего [11]. В результате был получен словарь из 300 слов и словосочетаний, представляющих списки терминов, которые репрезентативны соответствующему типу сценария. Однако отбор терминов-кандидатов из выбранных документов проводился не автоматизированным, а экспертным методом.

Несмотря на то что статистические методы легко вычислимы, они, как правило, требуют участия экспертов, поскольку выбор исходных данных в значительной степени зависит от предварительных знаний. Также акцент на статистике весьма ограничен, поскольку не учитывается семантическое значение терминов.

Семантическая маркировка (также известная как семантическая аннотация) — это процесс добавления семантики к терминам в тексте для облегчения автоматической интерпретации их значения. В процессе открытия знаний (извлечения из текста новых научных терминов) подходы, учитывающие контекстные отношения и другие смысловые связи между словами, как правило, производят более значимые ассоциации знаний и способствуют обнаружению предметно более релевантных слов[2] [12, 13].

При построении алгоритмической модели определения ключевых слов на базе корпуса «Википедии» предпринята попытка теоретической интеграции лингвистических, статистических и семантических методов [14]. Однако прикладные результаты работы не приводятся.

В рамках экспериментов по автоматизации обнаружения новых смысловых ассоциаций во Всемирной паутине комбинируются количественные статистические и экспертные методы, основанные на семантической ассоциативной системе [15]. Исследование, выявляя в режиме сетевого поиска подходящие неявные приложения от исходного запроса по компьютерной тематике, а именно «генетический алгоритм», показывает эффективность эвристики подхода. Валидация полученных смысловых ассоциаций происходит на основе анализа тестов

[1] Бодрунов С. Д. Ноономика : монография. Москва-Санкт-Петербург-Лондон : Культурная революция, 2018. 432 с. EDN: XQTTJZ

[2] Машина Е. А. Использование семантического анализа на начальном этапе информационного поиска в большом массиве источников научно-технической информации // XI Конгресс молодых учёных : Сб. науч. трудов. СПб : ИТМО, 2022. С. 418-425. EDN: LWUSRU

пересечений терминов в научной (WoS) и коммуникационной сети (UseNet). Однако, в данной работе отсутствует алгоритмический подход и не представлена полнота исходных данных. Сетевые (графовые) подходы используют свойства и теории графов для выявления новых понятий и семантических ассоциаций между понятиями [16]. Обычно они полагаются на модель обнаружения AnC и в основном выводят пути графа, не исключая ряд связующих терминов, соединяющих концепции начала (A) и цели (C). Однако пути графа могут включать ряд связующих (промежуточных) терминов (B) таким образом, что проявляется A→B→C (т. е. ABC — каноническая модель). Следовательно, результаты, основанные на графовых подходах, имеют существенное значение при создании более полных исследовательских гипотез.

Предложена модель, которая использует базу данных по медицине SemMedDB для извлечения семантических предикатов (бинарных отношений между двумя понятиями) для построения графа знаний [17]. Результат их работы — автоматическая генерация подграфов на основе контекста / тематического измерения путей.

Данный подход требует от пользователя только три элемента в качестве входных данных: (1) список меток понятий для источника (A) и цели (C), (2) максимальную длину пути k для генерируемых ABC-ассоциаций, (3) крайний срок dt для включения статей из научной литературы. Результатом подхода является ранжированный список подграфов S, т. е. создается функция F(q) = S, если q = {A;C;dt;k}.

Принципы данного подхода могут быть использованы при решении иных функциональных задач, в том числе в других предметных областях.

Наша базовая идея при формировании словаря прогностических терминов основана на интеграции трех условий, расширяющих возможности сетевого (графового) подхода:

1) автоматическая генерация графов (автоматизация);
2) учет ранжирования при оценке результатов поиска (ранжирование);
3) учет при генерации семантических контекстов (контекстно-семантическая фильтрация).

Для автоматизации поиска применяется несколько алгоритмов построителя графов, собирающего подходящие приложения для каждого исходного запроса по заданной тематике[3].

В процессе выполнения алгоритма поиска в ширину (BFS) запрос, отправленный из начального узла, направляется ко всем ближайшим соседям. Если узел-получатель обнаруживает запрос, выполняется поиск его локального индекса, и в случае успеха возвращается результат. В противном случае запрос направляется дальше по сети. При успешном завершении поиска формируется соответствующее сообщение.

При использовании алгоритма поиска в глубину (DFS) построитель графов оперирует генератором путей для извлечения всех путей между узлами (A, C) с заданной длиной k. Выбор DFS также может оказаться эффективным при обходе графа.

Определение авторитетности источника может быть облегчено анализом топологии ссылок между документами, основанным на взаимосвязях алгоритма ранжирования результатов сетевого поиска. Ранжирование представляет собой процесс, в ходе которого поисковая система упорядочивает результаты поиска в соответствии с наилучшим соответствием конкретному запросу. Визуализация результатов поиска тесно связана с выбранным алгоритмом ранжирования.

В 1998 и 1999 годах появились два наиболее известных алгоритма ранжирования веб-страниц, основанных на сетевых связях: PageRank, разработанный в Стэнфордском университете С. Брином и Л. Пейджем [18], и HITS (Hyperlink Induced Topic Search) — Дж. Клейнбергом в IBM [19].

В алгоритме учета популярности HITS выделяются два вида узлов: «авторы» — авторитетные страницы-первоисточники, на которые ссылаются, и страницы-посредники («хабы»), которые содержат множество ссылок на страницы, являющихся ценными первоисточниками. Алгоритм HITS заключается в выборе подграфа из гипертекстовой сети на основе запроса и определении лучших авторов и посредников по результатам анализа этого подмножества.

Для каждого документа d_j из информационного массива D вычисляется его важность как автора $a(d_j)$ и посредника $h(d_j)$ согласно формулам:

$$a(d_j) = \sum_{i=1,\, i \neq j}^{|D|} h(d_i),\ h(d_j) = \sum_{i=1,\, i \neq j}^{|D|} a(d_i) \qquad (1)$$

Алгоритм PageRank является одним из наиболее известных расширений индекса цитирования в информационных сетях. Он определяет важность некоей веб-страницы A на основе информации о страницах массива D, ссылающихся на нее.

Допустим, имеется n страниц $\{d_1, \ldots, d_n\}$, которые ссылаются на данную веб-страницу A, а C (A) обозначает общее количество ссылок с веб-страницы A на другие документы.

Алгоритм PageRank рассматривает каждую страницу как состояние в марковской модели. Далее он устанавливает положительную вероятность перехода между двумя страницами, если они имеют общую ссылку [20]. Оценивая вероятность того, что пользователь, просматривая страницу из множества D, перейдет на страницу A по ссылке, вводится фиксированное значение δ (коэффициент затухания). Обычно значение δ принимается близко к 0,85.

Для того чтобы устранить проблему перехода с одной страницы на любую другую, например, с которой нет общих ссылок, в алгоритме прибавляется небольшая случайная вероятность перехода с одной страницы на любую другую. Индекс PageRank RP(A) для страницы A интерпретируется как вероятность того, что пользователь окажется на этой странице в какой-то случайный момент времени:

$$PR(A) = \frac{(1-\delta)}{N} + \delta \sum_{i=1}^{n} \frac{PR(d_i)}{C(d_i)} \qquad (2)$$

где d_i (i = 1, 2..., n) представляет другие страницы, которые ссылаются на страницу A; $C(d_j)$ — количество исходящих ссылок на странице d_i; $PR(d_j) / C(d_j)$ — значение индекса PR, которое страница d_i вносит в страницу A; N — общее количество страниц в коллекции.

[3] Левитин А. В. Алгоритмы. Введение в разработку и анализ. М. : Вильямс, 2006. 576 с.

По приведенной выше формуле первоначальное значение PR для каждой страницы устанавливается равным 1, а затем рекурсивно рассчитывается простым итерационным алгоритмом, пока не будет достигнуто стабильное значение. Этот алгоритм основан на модели, в которой доступ пользователя к сети совершенно случайный, а саму модель PageRank можно рассматривать как сочетание модели случайного блуждания и модели Маркова[4] [21].

В работе [22] для извлечения текстовых ключевых слов применен концептуально аналогичный модели PageRank алгоритм ранжирования на основе графов TextRank, используя «Википедию» в качестве внешней базы знаний.

Очевидно, что сетевые концепции могут быть достаточно обширными, если они не ограничиваются определенной темой, соответствующей предметной области. Чтобы преодолеть этот эффект, алгоритм должен обеспечивать семантическую фильтрацию контента — предметом анализа являются только те документы, которые содержат контекстно-близкие термины по отношению к изначально заданному в запросе. Соответствие этим требованиям ограничивает размер (длину) формируемых сетей — моделей предметных областей, а также динамику их формирования [23].

Среди рассмотренных методов подход[5] представляется наиболее подходящим. Учитывая характер нашего исследования, где актуальность и разнообразие данных являются ключевыми факторами, использование обширной базы знаний «Википедии» предоставляет нам значительное преимущество [24].

Исходя из предложенного подхода к формированию предметной области разработан модифицированный алгоритм, применимый к задаче формирования словаря прогностических терминов. Данный алгоритм позволяет на базе сервиса «Википедия» формировать словарь контекстно-близких терминов по отношению к изначально заданному термину. При этом задаётся набор изначальных терминов и по результатам их обработки формируется полноценный словарь. Алгоритм оценивает «вес» близости терминов по отношению к изначальному термину. Ранжирование терминов выполняется по средней арифметической сумме оценок алгоритмов PageRank и HITS. Таким образом, разработанный алгоритм формирования словаря прогностических терминов основан на комплексном использовании подходов семантического и сетевого анализа.

Алгоритм написан на языке Python. Используемые библиотеки:

1) Requests — для выполнения запросов в сеть Интернет;
2) BeautifulSoup — для расшифровки полученных ответов на запросы;
3) Networkx — работа с графами, PageRank, HITS, степень узла;
4) Counter(collections) — для работы со словарями;
5) Concurrent.futures — для асинхронного выполнения алгоритма.

Алгоритм формирует словарь в асинхронном режиме с использованием распараллеливания основного алгоритма к каждому из первичных терминов. Алгоритм состоит из следующих итераций:

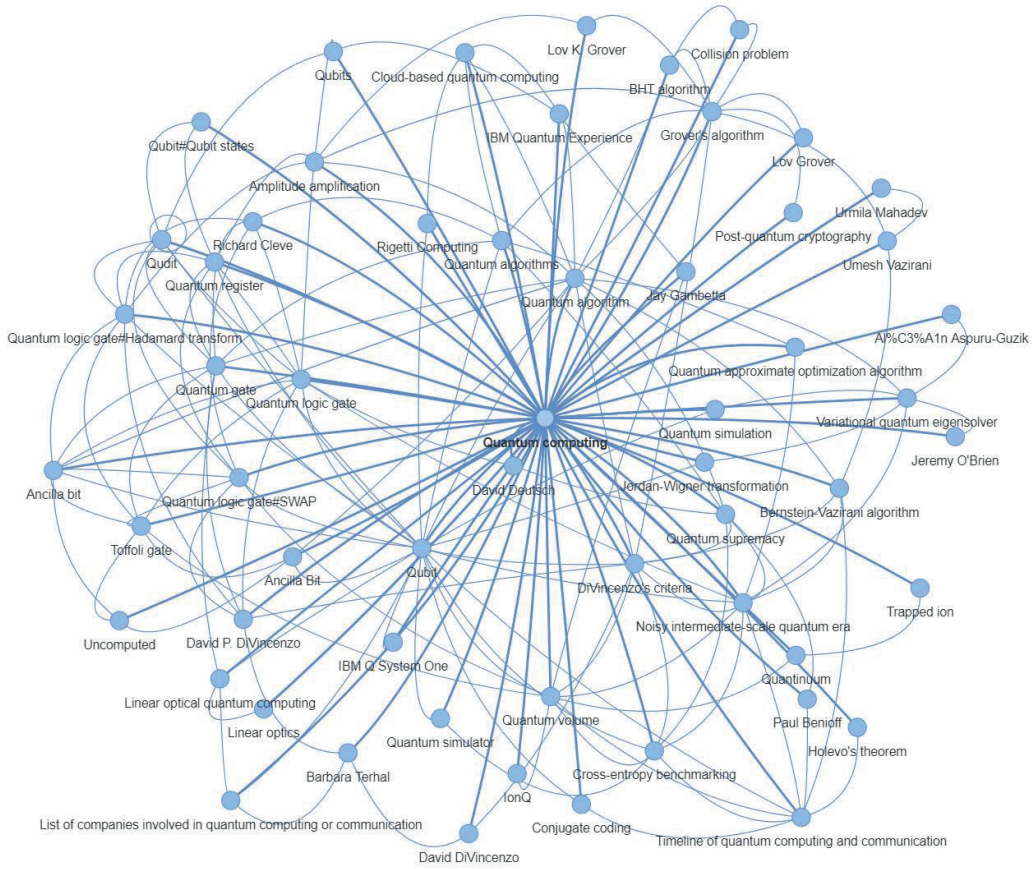
1) выбирается первичный термин/словосочетание и добавляется в граф как узел;
2) алгоритм находит одноименную страницу в сервисе «Википедия», посвященную данному термину;
3) анализируется страница в поисках всех ссылок на другие термины/словосочетания, которые ведут на другие страницы «Википедии», не учитываются разделы «Содержание», «Примечания», «Литература»;
4) по очереди для каждого найденного термина/словосочетания выполняются шаги 2–3;
5) выполняется проверка: если первичный термин/словосочетание был найден среди терминов на страницах, полученных при выполнении шагов 2–3 к первичному термину, то он добавляется в граф с установкой связи к узлу, от которого начался анализ;
6) рекурсивно выполняются шаги 2–6 к добавленному новому термину/словосочетанию (узлу), но важно, что среди его терминов алгоритм продолжает искать первичный термин/словосочетание;
7) рекурсия завершения в тот момент, когда ветвь будет сформирована;
8) выполняется переход на следующей термин из очереди, полученной при выполнении шага 3 к первичному алгоритму;
9) основной алгоритм завершается после того, как все ветви графа будут сформированы;
10) после завершения основной части алгоритма, выбираются первые 10 терминов/словосочетаний, полученных в ходе анализа, контекстно близких к первоначальному термину, и к ним применяются шаги 1–9;
11) все полученные термины ранжируются с использованием среднеарифметического значения результатов PageRank и HITS (authorities и hubs);
12) на выходе: ранжированный список из собранных терминов/словосочетаний по приближенности к первичному термину. Аналогичные списки для каждого из первичных терминов.

Для запуска работы алгоритма набор изначальных терминов необходимо сформировать по категориям технологий, заданных программой исследования [25]. Анализ данных должен быть сгруппирован по категориям заданной предметной области соответственно.

Для интерпретации результатов Text Mining особое значение имеет визуализация, что подразумевает специальную обработку структурированных данных, выводимых алгоритмом. Визуализация работы алгоритма на примере первичного термина «Quantum computing» представлена на рис.1.

[4] Page S. E. The Model Thinker: What You Need to Know to Make Data Work for You. New York: Basic Books, 2018. 427 p.

[5] Lande D. V., Andrushchenko V. B., Balagura I. V. Formation of the Subject Area on the Base of Wikipedia // Open Semantic Technologies for Intelligent Systems (OSTIS-2017). Minsk: BSUIR, 2017. С. 211-214. URL: https://libeldoc.bsuir.by/bitstream/123456789/12059/1/Lande_Formation.PDF (дата обращения: 23.08.2023).



Р и с. 1. Визуализация графа терминов/словосочетаний, полученных на основе входного термина «Quantum computing»
F i g. 1. Visualization of a graph of terms/phrases obtained based on the input term “Quantum computing”

Источник: здесь и далее в статье все таблицы и рисунки составлены авторами.
Source: Hereinafter in this article all tables and figures were made by the authors.

Для анализа полученных данных необходим вывод сводной таблицы с результатами ранжирования по алгоритмам PageRank и HITS. Вывод таблицы с раскладкой по данным позволит оценить объективность подхода к формированию глоссария.

Результаты и их обсуждение

В рамках выполнения задачи формирования глоссария прогностических терминов были проанализированы с использованием описанного выше алгоритма 4 категории терминов [25] по направлению ИКТ (Таблица 1).

Т а б л и ц а 1. **Категории перспективных направлений развития ИКТ, соответствующих четвертому уровню SML-матрицы**
T a b l e 1. **Categories of promising areas of ICT development corresponding to the fourth level of the SML matrix**

I	**Human–computer interfaces**	
1	aambient intelligence	интеллектуальная среда
2	brain–computer interface	интерфейс мозг–компьютер
3	city brain	городской мозг
4	semantic web	семантический веб
5	smart city	умный город

II	**Computing engineering**	
1	exascale computing	масштабные вычисления
2	neuromorphic engineering	нейроморфная инженерия
3	optical computing	оптические (фотонные) вычисления
4	quantum computing	квантовые вычисления
III	**Memory and data storage technologies**	
1	3D optical data storage	3D оптическое хранение данных
2	DNA digital data storage	цифровое хранилище данных ДНК
3	holographic data storage	голографическое хранилище данных
4	patterned media	узорчатые носители
5	phase-change memory	память с фазовым переходом
6	quantum memory	квантовая память
IV	**Electronics and communications**	
1	atomtronics	атомтроника
2	carbon nanotube field-effect transistor	полевой транзистор из углеродных нанотрубок
3	Li-Fi (Light Fidelity)	Li-Fi
4	memristor	мемристор, мемтранзистор, мемистор, транситор
6	software-defined radio	программно-определяемое радио
7	spintronics	спинтроника, твистроника, валлейтроника

По результатам работы алгоритма в отношении каждого термина/словосочетания из представленных в Таблице 1 были получены ранжированные по релевантности списки терминов. Для оценки релевантности к анализируемому термину/словосочетанию использованы результаты двух алгоритмов PageRank и HITS. Формула средней арифметической суммы, по которой выполнялась оценка веса полученного термина:

$$S_{cp} = \frac{PageRank + HITS}{2} \quad (3)$$

где *PageRank* — вес термина по оценке алгоритма PageRank, *HITS* — вес термина по оценке алгоритма HITS.

Вес термина HITS, в свою очередь, определялся по формуле:

$$HITS = \frac{HITSauthorities + HITShubs}{2} \quad (4)$$

где *HITSauthorities* – вес термина по оценке алгоритма *HITSauthorities, HITShubs* – вес термина по оценке алгоритма HITShubs.

Для оценки репрезентативности подхода проведен сравнительный анализ терминов по каждому перечню, ранжированному по весам PageRank, HITSauthorities, HITShubs, а также по параметру степени узла в сети поочередно. Наиболее высокоранговые результаты приведены на примере исходных терминов «Atomtronics» из категории «Electronics and communications» (Таблица 2), «Quantum computing» из категории «Computing engineering» (Таблица 3).

Таблица 2. **Сравнение результатов алгоритмов PageRank и HITS по термину «Atomtronics»**
Table 2. **Comparison of results of PageRank and HITS algorithms for the term "Atomtronics"**

Atomtronics										
№	PageRank		HITSauthorities		HITShubs		(PageRank+HITS)/2		Степень узла	
1	Unconventional computing	0,14899	Phase shift gates	0,00121	Reversible computing	0,14989	Unconventional computing	0,08602	Reversible computing	33
2	Carbon nanotube field-effect transistor	0,10494	Universal quantum gates	0,00121	Toffoli gate	0,11164	Reversible computing	0,05390	Unconventional computing	27
3	Beyond CMOS	0,04662	Deutsch gate	0,00121	Logic gate	0,07539	Carbon nanotube field-effect transistor	0,05250	Toffoli gate	22

4	Biocomputer	0,04502	Quantum logic gate	0,00123	Norman Margolus	0,07230	Toffoli gate	0,03975	Tangible user interface	22
5	Schottky barrier	0,04460	Quantum gate	0,00112	Quantum circuit	0,06541	Quantum circuit	0,02789	Billiard-ball computer	20
6	Wetware computer	0,03302	Toffoli gate	0,00175	Ancilla bit	0,06294	Biocomputer	0,02424	TUIO	16
7	Chemical computing	0,02959	Reversible computing	0,01193	List of quantum logic gates	0,06216	Ancilla bit	0,02394	Quantum circuit	14
8	Peptide computing	0,02914	Quantum circuit	0,00218	Billiard-ball computer	0,04065	Logic gate	0,02392	Logic gate	13
9	Chemical computer	0,02454	Billiard ball computer	0,00877	Unconventional computing	0,03993	Beyond CMOS	0,02334	Ancilla bit	11
10	Biocomputers	0,02239	Ancilla bit	0,00093	Controlled NOT gate	0,03878	Schottky barrier	0,02230	Quantum logic gate	11

Таблица 3. **Сравнение результатов алгоритмов PageRank и HITS по термину «Quantum computing»**

Table 3. **Comparison of the results of the PageRank and HITS algorithms for the term "Quantum computing"**

Quantum computing										
№	PageRank		HITSauthorities		HITShubs		(PageRank+HITS)/2		Степень узла	
1	Quantum counting	0,03026	Quantum logic gates	0,01495	Qubit	0,11698	Qubit	0,03275	Qubit	86
2	Quantum phase estimation algorithm	0,03010	Controlled gates	0,01491	Quantum register	0,04902	Quantum phase estimation algorithm	0,01981	Quantum phase estimation algorithm	38
3	DiVincenzo's criteria	0,02175	Phase shift gate	0,01491	Qudit	0,04842	DiVincenzo's criteria	0,01891	DiVincenzo's criteria	32
4	Bernstein-Vazirani algorithm	0,01464	Quantum gate	0,01485	Toffoli gate	0,04353	Quantum counting	0,01742	Quantum counting	28
5	Magic state distillation	0,01452	Quantum logic gate	0,01483	Ancilla bit	0,03983	Qudit	0,01665	Qudit	27
6	Quantum phase estimation algorithm#Analysis	0,01398	Hadamard transform	0.01482	Quantum information	0,03919	Ancilla bit	0,01628	Toffoli gate	23
7	IBM Quantum Experience	0,01389	Hadamard gate	0.01464	Timeline of quantum computing and communication	0,02559	Quantum register	0,01558	Ancilla bit	23
8	Hidden Linear Function problem	0,01235	SWAP	0, 01464	Quantum algorithm	0,02445	Toffoli gate	0,01425	Quantum register	23
9	Geometric proof of correctness	0,01133	Rotation operator gates	0.01464	DiVincenzo's criteria	0,02313	Quantum algorithm	0,01324	Quantum algorithm	22
10	Gottesman-Knill theorem	0,01029	Universal quantum gates	0.01459	Noisy intermediate-scale quantum era	0,02069	Quantum information	0,01315	Quantum information	22

Анализ данных вышеприведённых таблиц подтверждает репрезентативность подхода, основанного на учете ранжирования данных, полученных с помощью двух алгоритмов PageRank и HITS. Кроме того, сведение данных к среднеарифметической оценке веса полученных терминов по формуле (3) демонстрирует следующие преимущества в сравнении с применением показателей, выбранных в рамках отдельных алгоритмов:

- учет различных аспектов влияния. Среднеарифметическое ранжирование учитывает как PageRank, так и HITS, что дает более полное представление о влиянии каждого элемента. При этом оно устраняет возможные искажения, которые могут возникнуть при использовании только одного алгоритма;
- сглаживание экстремальных значений. В некоторых случаях PageRank и HITS могут давать сильно разные результаты. Среднеарифметическое значение сглаживает эти различия, предоставляя более устойчивую оценку важности;

- улучшенная устойчивость к выбросам. Если один из алгоритмов даёт неверную оценку для какого-то элемента, среднеарифметическое ранжирование смягчает этот эффект и делает рейтинг более устойчивым к подобным ошибкам;
- более надежное среднее значение. Среднеарифметическое значение уменьшает вероятность переоценки или недооценки важности элементов, что делает его более надежным средством оценки;
- сохранение важности высокорейтинговых элементов. При среднеарифметическом ранжировании элементы, оцененные высоко как в PageRank, так и в HITS, остаются на высоких позициях, что отражает их действительную значимость.

Для детализации работы алгоритма при формировании глоссария в рамках заданных перспективных направлений развития ИКТ были рассмотрены результаты в отношении семантической контекстной близости к изначальному термину. В Таблице 4 приведены списки полученных терминов/словосочетаний применительно к четырем исходным базовым направлениям по категории «Computing engineering».

Таблица 4. **Демонстрация результата работы алгоритма по категории «Computing engineering»**
Table 4. **Demonstration of the result of the algorithm in the "Computing engineering" category**

Exascale computing	Neuromorphic engineering	Optical computing	Quantum computing
Petascale computing	Retinomorphic sensor	Optical transistor	Qubit
Zettascale computing	MOSIS	Optical switch	Quantum phase estimation algorithm
Human Brain Project	Silicon retina	Exciton-polaritons	DiVincenzo's criteria
LINPACK benchmarks	Electronic design automation	Exciton-polariton	Quantum counting
Traversed edges per second	Mead and Conway revolution	Quantum vortices	Qudit
LINPACK benchmarks HPLinpack	VLSI Project	Polariton laser	Ancilla bit
Computer performance by orders of magnitude	Computation and Neural Systems	Polariton superfluid	Quantum register
Cajal Blue Brain	Mead-Conway VLSI chip design revolution	Vertical-external-cavity surface-emitting-laser	Toffoli gate
Blue Brain	Event camera	Hybrid silicon laser	Quantum algorithm
Brain simulation	Silicon compiler	Distributed Bragg reflector laser	Quantum information

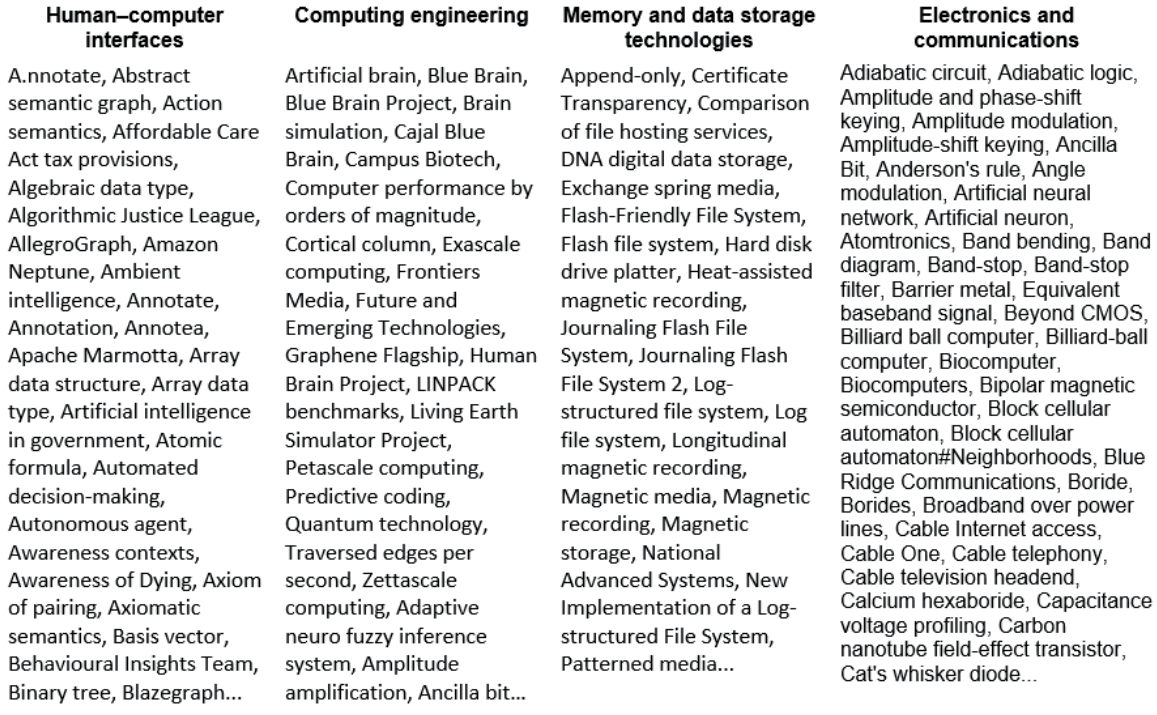


Рис. 2. Усечённый вывод структурированного по категориям глоссария
Fig. 2. Truncated output of a category-structured glossary

Каждый столбец Таблицы 4 демонстрирует ранжированный по среднеарифметическому показателю алгоритмов PageRank и HITS список из 10 терминов / тематических словосочетаний, стоящих на высоких позициях и при этом являющихся контекстно-близкими по отношению к изначальному термину. Качественная оценка контекстной близости следует из экспертного анализа смысловых значений терминов / тематических словосочетаний, представленных в Таблице 4.

По результатам работы алгоритма получены два сводных глоссария — общий и структурированный. Общий глоссарий содержит список из 1279 терминов прогностических ИКТ-технологий, отсортированных по алфавиту.

Для составления структурированного словаря в рамках категорий ИКТ, описанных в Таблице 1, сформированы алфавитные списки четырех глоссариев по тематикам. Словарный состав в глоссариях прогностических терминов, выводимый в приложении, находится в диапазоне от 66 до 155. Чтобы компенсировать более частое появление слов в отдельных категориях, данный состав терминов получен отсечением слов по среднему арифметическому показателю в значении менее 0,01. Для категории «Memory and data storage technologies» эти слова не ограничены. При процедуре анализа списков слов также исключены именованные сущности. Рисунок 2 демонстрирует полученный усечённый, структурированный глоссарий по категориям ИКТ.

Таким образом, в результате исследования достигнута первичная цель по формированию глоссария контекстно-близких прогностических терминов. На основе комплексного использования подходов семантического и сетевого анализа разработаны алгоритм и программный код, позволяющий осуществлять автоматическую генерацию графов от изначально заданных терминов. Результаты проведенного исследования коррелируют с результатами, полученными в ранних моделях использования обширной базы знаний «Википедии» [23]. Однако в предложенном варианте при оценке веса выводимых терминов данные распределяются с учетом ранжирования по средней арифметической оценке комбинации двух алгоритмов — PageRank и HITS.

В прикладном аспекте проанализированы и отражены таблицы, показывающие объективность представленного подхода к совокупной оценке веса термина. Выполнен вывод таблицы, демонстрирующей расширение контекста прогностических терминов по отношению к изначально заданному. Получены два сводных глоссария — общий и структурированный по категориям для дальнейшего развития исследования.

Заключение

В представляемом текущем этапе исследования ставилась задача формирования семантического ядра прогностических информационных технологий, которое будет использовано как базовый корпус знаний предметной области при последующем анализе библиографических баз данных тематических статей и дальнейшей кластеризации.

Качество структурирования информации в значительной степени определяет дальнейшие результаты современных технологий работы с текстами, что в конечном итоге сказывается на общих результатах анализа. В связи с этим в работе уделяется большое внимание вопросам предварительной подготовки исходных данных.

Для выполнения дальнейших исследований будет использован расширенный алгоритм, требующий реализации на основе программной платформы, которая включает несколько блоков обработки информации:

1) систему поиска в библиографических базах данных, электронных энциклопедиях, специализированных сайтах сети Интернет, сбора и хранения темпоральной коллекции текстовых документов, сформированных по категориям заданной предметной области;

2) систему препроцессинга текстовых данных, создание расширенных словарей терминов, векторизация текстовых документов. Каждый документ представляется вектором, элементами которого являются значения частот входящих в документ терминов, определенных на основе словаря всей коллекции;

3) систему ассоциативно-семантической кластеризации текстовых документов. С целью выбора оптимальных алгоритмов кластеризации векторов по смыслам будет проведено исследование не только различных вычислительных методов (например, итеративный алгоритм K-means, плотностный алгоритм DBSCAN, иерархические алгоритмы), но и используемых для их создания семантических технологий.

References

[1] Ataeva O.M., Serebryakov V.A., Tuchkova N.P. On Synonyms Search Model. *CEUR Workshop Proceedings.* 2022;3066:13-22. Available at: https://ceur-ws.org/Vol-3066/paper2.pdf (accessed 23.08.2023).

[2] Lanza C., Hazem A., Daille B. Towards Automatic Thesaurus Construction and Enrichment. In: Proceedings of the 6th International Workshop on Computational Terminology (COMPUTERM 2020). Language Resources and Evaluation Conference (LREC 2020). Marseille: European Language Resources Association; 2020. p. 62-71. Available at: https://aclanthology.org/2020.computerm-1.9.pdf (accessed 23.08.2023).

[3] Koutsomitropoulos D.A., Andriopoulos A.D. Thesaurus-based word embeddings for automated biomedical literature classification. *Neural Computing and Applications*. 2022;34(2):937-950. https://doi.org/10.1007/s00521-021-06053-z

[4] Vakaliuk T., Chernysh O., Babenko V. The Algorithm of Electronic Multilingual Terminological Dictionary Compilation. In: Proceedings of the 1st Symposium on Advances in Educational Technology. Vol. 2: AET. SciTePress; 2022. p. 323-331. https://doi.org/10.5220/0010931400003364

[5] Popov O.R., Kramarov S.O. The Study of Information Dissemination in Networks Arranged from a Set of Forecasting Terms. *Proceedings in Cybernetics*. 2022;(1):38-45. (In Russ., abstract in Eng.) https://doi.org/10.34822/1999-7604-2022-1-38-45

[6] Kozakov L., Park Y., Fin T., Drissi Y., Doganata Y., Cofino T. Glossary extraction and utilization in the information search and delivery system for IBM Technical Support. *IBM Systems Journal*. 2004;43(3):546-563. https://doi.org/10.1147/sj.433.0546

[7] Velardi P., Navigli R., D'Amadio P. Mining the Web to Create Specialized Glossaries. *IEEE Intelligent Systems*. 2008;23(5):18-25. https://doi.org/10.1109/MIS.2008.88
[8] Dogra V., Verma S., Kavita, Chatterjee P., Shafi J., Choi J., Ijaz M.F. A Complete Process of Text Classification System Using State-of-the-Art NLP Models. *Computational Intelligence and Neuroscience.* 2022:1883698. https://doi.org/10.1155/2022/1883698
[9] Soliman A. An unsupervised linguistic-based model for automatic glossary term extraction from a single PDF textbook. *Education and Information Technologies*. 2023;28:16089-16125. https://doi.org/10.1007/s10639-023-11818-1
[10] Van S. *Semanticheskij i strukturnyj analiz tekstov v seti Internet* [Semantic and structural analysis of texts on the Internet]. *E-Scio*. 2020;(4):51-60. (In Russ., abstract in Eng.) EDN: PBIGEH
[11] Fergnani A., Jackson M. Extracting scenario archetypes: A quantitative text analysis of documents about the future. *Futures & Foresight Science.* 2019;1(2):e17. https://doi.org/10.1002/ffo2.17
[12] Altınel B., Can Ganiz M. Semantic text classification*: A survey of past and rece*nt advances. *Information Processing & Management*. 2018;54(6):1129-1153. https://doi.org/10.1016/j.ipm.2018.08.001
[13] Jia C., Carson M.B., Wang X., Yu J. Concept decompositions for short text clustering by identifying word communities. *Pattern Recognition*. 2018;76:691-703. https://doi.org/10.1016/j.patcog.2017.09.045
[14] Kogay V.N., Pak V.S. *Algoritmicheskaya model' komp'yuternoj sistemy vydeleniya klyuchevyh slov iz teksta na baze ontologij* [Algorithmic model of computerized system of keywords extracting from text based on ontology]. *Problemy sovremennoj nauki i obrazovanija* = Problems of modern science and education. 2016;(16):33-40. (In Russ., abstract in Eng.) EDN: WFGOIT
[15] Gordon M., Lindsay R., Fan W. Literature-Based Discovery on the World Wide Web. *ACM Transactions on Internet Technology*. 2002;2(4):261-275. https://doi.org/10.1145/604596.604597
[16] Veremyev A., Semenov A., Pasiliao E., Boginski V. Graph-based exploration and clustering analysis of semantic spaces. *Applied Network Science.* 2019;(4):109. https://doi.org/10.1007/s41109-019-0228-y
[17] Cameron D., Kavuluru R., Rindflesch Th., Sheth A., Thirunarayan K., Bodenreider O. Context-Driven Automatic Subgraph Creation for Literature-Based Discovery. *Journal of Biomedical Informatics*. 2015;54:141-157. https://doi.org/10.1016/j.jbi.2015.01.014
[18] Brin S., Page L. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*. 1998;30(1-7):107-117. https://doi.org/10.1016/S0169-7552(98)00110-X
[19] Kleinberg J.M. Authoritative sources in a hyperlinked environment. *Journal of the ACM*. 1999;46(5):604-632. https://doi.org/10.1145/324133.324140
[20] Patchmuthu R.K., Goh K.L.A., Singh A.K. Application of Markov Chain in the PageRank Algorithm. *Pertanika Journal of Science & Technology*. 2013;21(2):541-554. Available at: http://www.pertanika.upm.edu.my/pjst/browse/regular-issue?article=JST-0397-2012 (accessed 23.08.2023).
[21] Li H. The PageRank Algorithm. In: Machine Learning Methods. Singapore: Springer; 2024. p. 473-492. https://doi.org/10.1007/978-981-99-3917-6_21
[22] Li W., Zhao J. TextRank Algorithm by Exploiting Wikipedia for Short Text Keywords Extraction. In: 2016 3rd International Conference on Information Science and Control Engineering (ICISCE). Beijing, China: IEEE Computer Society; 2016. p. 683-686. https://doi.org/10.1109/ICISCE.2016.151
[23] Pech F., Martinez A., Estrada H., Hernandez Y. Semantic Annotation of Unstructured Documents Using Concepts Similarity. *Scientific Programming.* 2017;2017(1):7831897. https://doi.org/10.1155/2017/7831897
[24] Heist N., Heiko P. Entity Extraction from Wikipedia List Pages. *The Semantic Web*. 2020;(12123):327-342. https://doi.org/10.1007/978-3-030-49461-2_19
[25] Kramarov S.O., Popov O.R., Dzhariev I.E., Petrov E.A. Dynamics of link formation in networks structured on the basis of predictive terms. *Russian Technological Journal*. 2023;11(3):17-29. https://doi.org/10.32362/2500-316X-2023-11-3-17-29

Поступила 23.08.2023; одобрена после рецензирования 07.10.2023; принята к публикации 10.10.2023.
Submitted 23.08.2023; approved after reviewing 07.10.2023; accepted for publication 10.10.2023.

Об авторах:

Попов Олег Русланович, ведущий научный сотрудник Южного отделения, МОО «Академия информатизации образования» (109029, Российская Федерация, г. Москва, ул. Нижегородская, д. 32), кандидат технических наук, доцент, член-корреспондент АИО, **ORCID: https://orcid.org/0000-0001-6209-3554,** cs41825@aaanet.ru
Гросу Адриан, аспирант кафедры автоматики и компьютерных систем, БУ ВО «Сургутский государственный университет» (628400, Российская Федерация, Ханты-Мансийский автономный округ – Югра, г. Сургут, пр. Ленина, д. 1), **ORCID: https://orcid.org/0009-0004-5520-6708**, grosu_a@surgu.ru
Крамаров Сергей Олегович, главный научный сотрудник, БУ ВО «Сургутский государственный университет» (628400, Российская Федерация, Ханты-Мансийский автономный округ – Югра, г. Сургут, пр. Ленина, д. 1), доктор физико-математических наук, профессор, **ORCID: https://orcid.org/0000-0003-3743-6513**, maoovo@yandex.ru

Все авторы прочитали и одобрили окончательный вариант рукописи.

About the authors:

Oleg R. Popov, Leading Researcher of the Southern Branch of the Academy of Informatization of Education (32 Nizhegorodskaya St., Moscow 109029, Russian Federation), Cand. Sci. (Tech.), Associate Professor, Corresponding member of the Academy of Informatization of Education, **ORCID: https://orcid.org/0000-0001-6209-3554,** cs41825@aaanet.ru

Adrian Grosu, Postgraduate Student of the Department of Automation and Computer Systems, Surgut State University (1 Lenin Ave., Surgut 628400, Khanty-Mansi Autonomous Okrug – Ugra, Russian Federation), **ORCID: https://orcid.org/0009-0004-5520-6708**, grosu_a@surgu.ru

Sergey O. Kramarov, Chief Researcher, Surgut State University (1 Lenin Ave., Surgut 628400, Khanty-Mansi Autonomous Okrug – Ugra, Russian Federation), Dr. Sci. (Phys.-Math.), Professor, **ORCID: https://orcid.org/0000-0003-3743-6513**, maoovo@yandex.ru

All authors have read and approved the final manuscript.