МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М. В. ЛОМОНОСОВА

На правах рукописи

Шачнев Дмитрий Алексеевич

Семантические модели классификации и анализа данных в больших информационно-аналитических системах

05.13.17 – Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва – 2021

Работа выполнена на кафедре вычислительной математики механико-математического факультета ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова».

Научный руководитель: **Васенин Валерий Александрович**
д. ф.-м. н., профессор

Официальные оппоненты: **Новиков Борис Асенович**
д. ф.-м. н., профессор,
НИУ ВШЭ, Санкт-Петербургская школа физико-математических и компьютерных наук, профессор

Захаров Владимир Анатольевич
д. ф.-м. н.,
МГУ имени М. В. Ломоносова, факультет вычислительной математики и кибернетики, профессор

Горбунов-Посадов Михаил Михайлович
д. ф.-м. н., старший научный сотрудник,
ИПМ им. М.В. Келдыша РАН, заведующий отделом

Защита состоится 27 октября 2021 г. в 16 часов 45 минут на заседании диссертационного совета МГУ.05.01 при ФГБОУ ВО «Московский государственный университет имени М. В. Ломоносова» по адресу: 119991, Москва, ГСП-1, Ленинские горы, д. 1, Главное здание МГУ, аудитория 14-08.

E-mail: vasenin@msu.ru

С диссертацией можно ознакомиться в отделе диссертаций научной библиотеки МГУ имени М. В. Ломоносова (Ломоносовский просп., д. 27) и на сайте ИАС «ИСТИНА»: `https://istina.msu.ru/dissertations/384304186/`.

Автореферат разослан 24 сентября 2021 г.

Учёный секретарь
диссертационного совета
МГУ.05.01, к.ф.-м.н. М. Кривчиков Кривчиков
Максим Александрович

Общая характеристика работы

Актуальность темы исследования. В настоящее время одним из важных направлений исследований в мире и в России является разработка математических моделей, алгоритмов и программных средств для поиска и систематизации информации о результатах деятельности субъектов (работников, коллективов) в разных предметных областях. Результаты таких исследований нужны для оценки её эффективности на основе анализа данных, которые хранятся в больших, перманентно изменяющихся базах, содержащих сведения о деятельности субъектов. Задачи в подобной постановке актуальны во всех секторах экономической и хозяйственной деятельности. На их решение направлен Национальный проект «Цифровая экономика Российской Федерации».

Эффективность механизмов управления во всех секторах национальной экономики зависит от точности и оперативности данных, на основе анализа которых управленческие решения готовятся и принимаются. Важное место в таком анализе, как правило, отводится специалистам, имеющим хорошие знания и богатый практический опыт в отдельной предметной области. Таких специалистов в контексте данной диссертации будем именовать *экспертами*. Их знания, навыки используются не только для получения адекватной оценки текущего состояния сферы деятельности, но и для формирования планов, определяющих её будущее.

С этих позиций решение задачи анализа данных, аккумулируемых в отмеченных базах, с целью поиска и ранжирования результатов работ, выполняемых специалистами — экспертами, является актуальным и в экономическом, и в социальном плане. Методы и средства такого ранжирования зависят от конкретной отрасли и установленных в ней приоритетов, а также изменяются со временем.

Результаты исследований будут изложены применительно к сфере научно-исследовательской, опытно-конструкторской и технологической деятельности. Такая сфера деятельности является одной из самых сложных с точки зрения бизнес-логики и многопараметричности её описания. Это обстоятельство позволяет с вы-

сокой степенью вероятности утверждать, что полученные в этой области решения могут с успехом использоваться в других областях деятельности.

Большой объём данных в рассматриваемой сфере деятельности связан с тем, что в научных организациях в последние годы стали широко использоваться системы анализа текущих исследований (Current Research Information Systems, *CRIS-системы*), в которых хранятся метаданные о результатах деятельности специалистов: научной, инновационной, преподавательской, публицистической и т.п. Примерами таких результатов деятельности являются научные публикации, выступления на конференциях, учебные курсы, научное руководство, членство в редакционных коллегиях научно-технических журналов и сборников.

Важное место в процессе анализа научно-исследовательских, опытно-конструкторских и технологических работ (НИОКТР) имеют индикаторы и показатели их значимости в той или иной системе оценивания. Например, одним из механизмов определения значимости научно-технической публикации является число её цитирований. На цитируемости основаны более сложные показатели, например, импакт-фактор журнала и индекс Хирша. В большинстве CRIS-систем подобные показатели доступны, и, как следствие, возникает возможность *поиска экспертов в заданной предметной области по их результатам деятельности и с учётом показателей этих результатов*.

Исследования, результаты которых представлены в диссертации, предполагают соблюдение принципов построения CRIS-систем, описанных в Лейденском манифесте для наукометрии[1] и нацеленных на рациональное применение индикаторов с учётом их недостатков. Создание новых наукометрических показателей выходит за рамки диссертации.

В случае, если доступных показателей для оценки значимости несколько, возникает вопрос, какие именно показатели целесообразно использовать, и как их необходимо учитывать при вычислении коэффициента значимости результата деятельности. Поскольку требования к показателям результатов деятельности

[1] The Leiden Manifesto for research metrics / D. Hicks [et al.] // Nature. 2015. Vol. 520. P. 429–431. ISSN 0028-0836.

могут меняться, возникает необходимость разработать модель, которая позволила бы пользователям системы описывать *сложные критерии отбора и оценки результатов* в формализованном виде, и механизмы для учёта этих критериев.

Таким образом, в диссертации представлены модели, алгоритмы и программные механизмы анализа данных в больших информационно-аналитических системах с целью выявления экспертов, активно работающих в заданной предметной области, и их ранжирования с учётом динамически изменяемых критериев.

Степень разработанности темы исследования. Подходы к поиску экспертов, средства и системы, реализующие эти подходы, исследуются с конца 1980-х годов. В ранних исследованиях на этом направлении, которые, например, представлены в работе М. Марона и др.[2], предполагается, что имеется конечное множество экспертов, для каждого из которых указаны области его интересов, и по этим областям выполняется поиск. Недостатком такой архитектуры является необходимость вручную поддерживать профили экспертов в актуальном состоянии, и отсутствие возможности установить, является ли заданная предметная область основной сферой интересов эксперта или второстепенной. В некоторых работах предполагается определять компетенции экспертов по косвенным данным, таким как их переписка по электронной почте (как, например, в работе К. Кэмпбелла и др.[3]), данные документов внутренней сети организации (как в системе P@NOPTIC Expert, описанной Н. Красвеллом и др.[4]) или по совокупности данных, например информации на домашних страницах работников, почтовой переписке и по информации о совместном участии в проектах (как в работе Р. Д'Амора[5]). В

[2] *Maron M. E.*, *Curry S.*, *Thompson P.* An Inductive Search System: Theory, Design, and Implementation // IEEE Transactions on Systems, Man, and Cybernetics. 1986. Vol. 16, no. 1. P. 21–28.

[3] Expertise Identification Using Email Communications / C. S. Campbell [et al.] // Proceedings of the Twelfth International Conference on Information and Knowledge Management (New Orleans, LA, USA). New York, NY, USA : Association for Computing Machinery, 2003. P. 528–531. (CIKM '03). ISBN 978-1-58113-723-1.

[4] P@NOPTIC Expert: Searching for Experts not just for Documents / N. Craswell [et al.] // In Ausweb. 2001. P. 21–25.

[5] *D'Amore R.* Expertise Community Detection // Proceedings of the 27th Annual International ACM SIGIR Confer-

работе Б. Алеман-Мезы и др.[6] предлагается добавлять на персональные страницы экспертов структурированные данные с использованием технологий семантической паутины, которые потом смогут быть использованы для поиска. В работе Д. Имам-Сейда и А. Кобсы[7] проводится обзор некоторых существующих информационных систем, реализующих поиск экспертов, и предлагается модульная архитектура, к которой можно динамически подключать новые источники данных.

В новых работах, в дополнение к анализу ключевых слов и текстовому поиску по документам, предлагается использовать методы машинного обучения для классификации документов по предметным областям. В работах К. Балога с соавторами[8] для этой цели используются генеративные модели линейного классификатора, в работе И. Фанга и др.[9] — дискриминативные модели. В работе Дж. Танга и др.[10] описывается система AMiner, извлекающая данные об учёных из разнородных данных в сети Интернет (в первую очередь, из их персональных страниц) и использующая генеративную вероятностную модель для реализации различных механизмов, в том числе поиска экспертов. В работе А. Синхи и др.[11] описывается система Microsoft Academic Graph. Данные в ней получены из интернета и из базы

---

ence on Research and Development in Information Retrieval (Sheffield, United Kingdom). New York, NY, USA : Association for Computing Machinery, 2004. P. 498–499. (SIGIR '04). ISBN 978-1-58113-881-8.

[6] Combining RDF Vocabularies for Expert Finding / B. Aleman-Meza [et al.] // The Semantic Web: Research and Applications / ed. by E. Franconi, M. Kifer, W. May. Berlin, Heidelberg : Springer, 2007. P. 235–250. ISBN 978-3-540-72667-8.

[7] *Yimam-Seid D.*, *Kobsa A.* Expert-finding systems for organizations: Problem and domain analysis and the DEMOIR approach // Journal of Organizational Computing and Electronic Commerce. 2003. Vol. 13, no. 1. P. 1–24.

[8] *Balog K.*, *Azzopardi L.*, *de Rijke M.* Formal Models for Expert Finding in Enterprise Corpora // Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Seattle, Washington, USA). New York, NY, USA : Association for Computing Machinery, 2006. P. 43–50. (SIGIR '06). ISBN 978-1-59593-369-0.

[9] *Fang Y.*, *Si L.*, *Mathur A. P.* Discriminative Models of Integrating Document Evidence and Document-Candidate Associations for Expert Search // Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (Geneva, Switzerland). New York, NY, USA : Association for Computing Machinery, 2010. P. 683–690. (SIGIR '10). ISBN 978-1-4503-0153-4.

[10] AMiner: Search and Mining of Academic Social Networks / H. Wan [et al.] // Data Intelligence. 2019. Vol. 1, no. 1. P. 58–76.

[11] An Overview of Microsoft Academic Service (MAS) and Applications / A. Sinha [et al.] // Proceedings of the 24th International Conference on World Wide Web (Florence, Italy). New York, NY, USA : Association for Computing Machinery, 2015. P. 243–246. (WWW '15 Companion). ISBN 978-1-4503-3473-0.

поискового сервиса Bing, формирование списка предметных областей и классификация данных выполнялись с использованием методов машинного обучения.

В упомянутых выше публикациях предлагаются методы и средства, предназначенные для анализа полуструктурированных данных, которые не содержат представления результатов деятельности работников в виде, удобном для их целевого анализа. Таким недостатком обладают, например, корпоративные информационные системы. Кроме того, многие из предлагаемых методов не используют в полной мере информацию о связях между объектами системы. Одним из основных отличий подхода, применяемого в диссертационном исследовании, является использование данных CRIS-систем, которые обеспечивают целевой сбор и хранение информации о результатах деятельности исследователей в структурированном виде. Это позволяет учесть как тематику результатов, так и их значимость.

Цели и задачи диссертационной работы. Целью диссертационной работы является исследование и разработка математических моделей, алгоритмов и программных средств для поиска экспертов в предметной области путём анализа данных в информационно-аналитических системах о результатах деятельности, и ранжирования экспертов с учётом показателей значимости результатов.

Для достижения поставленной цели решаются следующие задачи:

1. создание модели тематического портрета, который может быть использован для описания предметной области, результата деятельности или эксперта;
2. создание метрики схожести тематических портретов и отдельных составляющих их элементов;
3. разработка алгоритмов для составления тематических портретов объектов в информационно-аналитических системах по связям этих объектов с другими;
4. создание модели, позволяющей динамически определять критерии для отбора результатов деятельности и назначения им коэффициентов значимости в соответствии с их показателями, а также алгоритма для формирования

запросов к реляционным СУБД на основе этой модели;

5. разработка алгоритма поиска экспертов по данным об их результатах деятельности, с учётом релевантности результатов запросу и их значимости.

Соответствие паспорту специальности. Диссертация соответствует паспорту специальности 05.13.17 — «Теоретические основы информатики» в части п. 1 «Исследование, в том числе с помощью средств вычислительной техники, информационных процессов, информационных потребностей коллективных и индивидуальных пользователей», п. 2 «Исследование информационных структур, разработка и анализ моделей информационных процессов и структур», п. 5 «Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечениях, разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений» и п. 9 «Разработка новых интернет-технологий, включая средства поиска, анализа и фильтрации информации, средства приобретения знаний и создания онтологии, средства интеллектуализации бизнес-процессов».

Научная новизна диссертационной работы характеризуется следующими результатами.

- Разработаны метрики схожести тематических портретов и составляющих их элементов, учитывающие весовые коэффициенты в портретах и иерархическую структуру рубрикаторов. Сформулированы и доказаны теоремы о свойствах этих метрик. Экспериментально подтверждена их эффективность.
- Предложена модель, позволяющая динамически задавать критерии для отбора результатов деятельности и назначать им коэффициенты значимости с учётом их показателей.
- Разработан алгоритм поиска экспертов на основе данных об отдельных результатах их деятельности.

Теоретическая и практическая значимость. Изложенные в диссертации модели и алгоритмы востребованы в информационно-аналитических системах в

различных сферах деятельности. В научно-технической сфере результаты диссертации могут быть использованы для решения следующих задач:

- поиск специалистов для экспертизы заявок на проведение НИОКТР и их результатов;
- определение тематики деятельности отдельных исследователей или коллективов исполнителей, структурных подразделений и организаций в целом;
- выявление научных коллективов и связей между различными подразделениями и организациями при проведении междисциплинарных исследований;
- определение наиболее активно развивающихся научных областей;
- формирование рейтинговых показателей отдельных учёных, структурных подразделений или организаций целиком, в том числе, с привязкой к заданной предметной области;
- формирование отчётных и статистических материалов.

Теоретическая значимость диссертации заключается в разработке метрики схожести тематических портретов с учётом взаимных семантических связей между составляющими их элементами, алгоритма для построения тематического портрета вершины в графе информационно-аналитической системы по цепочкам связей с другими вершинами, а также модели для определения правил отбора результатов и вычисления коэффициентов значимости.

Методы исследования. Аппарат и методы теории графов, алгебры, комбинаторики, элементы онтологии.

Положения, выносимые на защиту. На защиту выносятся: обоснование актуальности, научная новизна, теоретическая и практическая значимость работы, а также следующие положения, которые подтверждаются результатами исследования, представленными далее в разделе Заключение.

1. Модель тематических портретов для описания заданной предметной области, алгоритм построения тематических портретов результатов деятельности и их авторов на основе совокупности информации о ключевых словах и

элементах рубрикаторов, связанных с ними в графе информационно-аналитической системы.

2. Метрики для определения степени схожести тематических портретов и составляющих их элементов, на основе анализа данных об их совместном использовании.
3. Модель и алгоритм, позволяющие динамически определять критерии для отбора результатов деятельности и назначать таким результатам коэффициенты значимости в соответствии с их наукометрическими показателями.
4. Программная реализация системы интеллектуального анализа данных для поиска и ранжирования экспертов.

Степень достоверности и апробация результатов. Материалы диссертации докладывались на всероссийских конференциях «Научный сервис в сети Интернет» (Новороссийск, 21.09.2016) и «Знания – Онтологии – Теории» (Новосибирск, 5.10.2017), международной Ершовской конференции по информатике (Москва, 29.06.2017), международной конференции «Актуальные проблемы системной и программной инженерии» (Москва, 12.11.2019), конференции «Ломоносовские чтения» (Москва, 18.04.2019, 29.10.2020 и 22.04.2021).

Кроме того, материалы были доложены на семинарах «Проблемы современных информационно-вычислительных систем» под руководством проф. В. А. Васенина (Москва, 13.12.2016, 17.04.2018 и 13.11.2018) и «Теоретические проблемы программирования» под руководством проф. В. А. Захарова (Москва, 19.05.2021).

Апробация результатов была проведена в информационно-аналитической системе «ИСТИНА», в разработке которой автор принимает участие. Система «ИСТИНА» используется в МГУ имени М. В. Ломоносова, ряде других вузов и институтов Российской академии наук, научно-технических центрах Минздрава РФ. Алгоритмы, изложенные в диссертации, легли в основу отдельного модуля поиска экспертов в этой ИАС, а также нашли применение в подсистемах расчёта рейтингов, проведения конкурсов, формирования отчётных материалов. По результатам апробации модели и реализующих её программных средств получен акт

о внедрении.

Публикации. Материалы диссертации опубликованы в 10 печатных работах, из них 2 статьи в журнале, входящем в индекс RSCI и в перечень ВАК [1 ; 2], 1 статья в журнале, индексируемом в системах WoS и Scopus [3], 4 статьи в сборниках трудов конференций [4 ; 5 ; 6 ; 7] и 3 тезиса докладов [8 ; 9 ; 10]. Список работ представлен в заключительной части автореферата.

Личный вклад автора. Все результаты, представленные в диссертации, получены лично автором. Предшествующие научные работы, использованные в диссертации, отмечены библиографическими ссылками.

Структура и объем диссертации. Диссертация состоит из введения, 5 глав и заключения. Полный объём диссертации составляет 112 страниц, включая 7 рисунков и 7 таблиц. Список литературы содержит 54 наименования.

Содержание работы

Во Введении обоснована актуальность диссертационной работы, сформулирована цель и аргументирована научная новизна исследований, показана практическая значимость полученных результатов, представлены выносимые на защиту научные положения.

Первая глава имеет вспомогательный характер. В разделах 1.1–1.6 представлены требования к моделям, алгоритмам и их программной реализации, которые принимались во внимание при их разработке. В разделе 1.7 приведены используемые термины и обозначения.

Во второй главе приведён обзор известных моделей для описания предметных областей, соответствующих различным объектам, и методов их сопоставления. Наиболее распространёнными являются методы машинного обучения, но их недостатками являются необходимость наличия полных текстов, которые не всегда можно получить в CRIS-системах, и недостаточная наглядность. Другое направление связано с построением онтологий, но на данный момент методы автома-

тизированного построения онтологий на основе текстов далеки от совершенства. Обоснован выбор модели, основанной на множестве ключевых слов и рубрик с присвоенными им вещественными коэффициентами. К достоинствам такой модели относится её наглядность, а также наличие ключевых слов и рубрик в научных публикациях, являющихся основным объектом анализа в CRIS-системах.

В разделе 2.1 приведена математически формализованная модель тематического портрета и представлено его определение.

Определение 1. *Ключевым словом* называется одно или несколько слов на естественном языке, описывающие тематику объекта. Множество возможных ключевых слов будем обозначать W.

Рубрикой называется вершина древовидного классификатора (рубрикатора). Каждый классификатор R имеет корневую вершину $r_0 \in R$ и не образующую циклов функцию, определённую для рубрик и сопоставляющую им родительские рубрики: $\mathrm{par}\colon R \setminus \{r_0\} \to R$. Примерами таких классификаторов являются УДК, ГРНТИ и классификатор РФФИ.

Определение 2. *Тематическим портретом* или далее по тексту — *портретом* будем называть набор ключевых слов и рубрик, каждой из которых присвоен некоторый вещественный коэффициент от 0 до 1:

$$p = \{(w_1, c_{w_1}), \ldots, (w_n, c_{w_n}), (r_1, c_{r_1}), \ldots, (r_m, c_{r_m})\}, \tag{1}$$

где $n + m > 0,\ w_i \in W,\ r_j \in R,\ 0 < c_{w_i}, c_{r_j} \leq 1$.

Для удобства ключевые слова и рубрики будем называть *токенами*. Через T обозначим множество всех токенов $W \cup R$.

Определение 3. Для портрета p из (1) множество $\{w_1, \ldots, w_n, r_1, \ldots, r_m\}$ будем называть *множеством токенов портрета* p и обозначать T_p.

Целью разделов 2.2 и 2.3 является разработка функций схожести: двух ключевых слов, двух произвольных токенов и двух тематических портретов. Для разработки таких функций будет необходимо иметь заранее заданное множество

тематических портретов. Такое множество будем называть *обучающей выборкой* и обозначать P_{train}. Через $P(t)$ будем обозначать множество портретов из обучающей выборки, содержащих токен t: $P(t) = \{p \in P_{\text{train}} : t \in T_p\}$. Через $P(t_k, t_l)$ обозначим пересечение $P(t_k) \cap P(t_l)$.

Функция схожести двух ключевых слов основана на функции контекстной близости, описанной в работе К. В. Лунёва[12], с изменениями, позволяющими учесть коэффициенты c_{w_i} и добиться коэффициента схожести 1 ключевого слова с самим собой. Этот подход, в свою очередь, основан на дистрибутивной гипотезе. Определим вес связи между t_k и t_l следующей формулой:

$$\omega(t_k, t_l) = \sum_{p \in P(t_k, t_l)} \frac{c_{t_l}}{\sum_{t \in T_p} c_t}. \tag{2}$$

Пусть теперь w_i и w_j — два ключевых слова, $T_{i,j} = \{t \in T : |P(w_i, t)| \geq 1, |P(w_j, t)| \geq 1\}$ — множество токенов, встречающихся совместно с обоими словами в обучающей выборке. За коэффициент схожести между w_i и w_j примем следующее значение:

$$s(w_i, w_j) = \frac{\sum_{t \in T_{i,j}} (\omega(w_i, t) + \omega(w_j, t))}{|P(w_i)| + |P(w_j)|}. \tag{3}$$

Знаменатель учитывает частоту встречаемости слов в обучающей выборке, что позволяет компенсировать высокие значения числителя для частотных слов.

Теорема 1. *Значение коэффициента схожести $s(w_i, w_j)$ равно 1 тогда и только тогда, когда $\bigcup_{p \in P(w_i)} T_p = \bigcup_{p \in P(w_j)} T_p$, то есть все токены, являющиеся соседними для одного слова, являются также соседними и для другого. В частности, коэффициент схожести слова с самим собой равен 1. В остальных случаях $s(w_i, w_j) < 1$.*

В подразделе 2.2.2 вводятся функции схожести ключевого слова и рубрики, а также двух рубрик. Эти функции основаны на (3) с изменениями, позволяющими учесть древовидную структуру рубрикатора.

[12] *Лунев К. В.* Графовые методы определения семантической близости пары ключевых слов и их применения к задаче кластеризации ключевых слов // Программная инженерия. Москва, 2018. т. 9, № 6. с. 262—271. ISSN 2220-3397.

Определение 4. *Предком n-го уровня* рубрики r будем называть рубрику

$$\mathrm{par}^n(r) = \begin{cases} r & \text{если } n = 0, \\ \mathrm{par}\left(\mathrm{par}^{n-1}(r)\right) & \text{если } n \in \mathbb{N}. \end{cases}$$

Определение 5. *Высотой* рубрики r будем называть высоту поддерева, соответствующего этой рубрике: $h(r) = \max\{n \in \mathbb{Z}_{\geq 0} : \exists r' : r = \mathrm{par}^n(r')\}$.

Введём понятие *значимости рубрики* — величины, которая будет равномерно убывать от 1 для рубрик-листьев до 0 для корня дерева:

$$I(r) = \frac{h(r_0) - h(r)}{h(r_0)}.$$

Через $\hat{s}$ будем обозначать функцию схожести (3), применённую к произвольным токенам как к ключевым словам. Функцию схожести ключевого слова и рубрики определим следующим образом:

$$s(w_k, r_l) = \max\left\{ \hat{s}(w_k, r_l),\ \frac{I(\mathrm{par}(r_l))}{|R(\mathrm{par}(r_l))|}\hat{s}(w_k, \mathrm{par}(r_l)),\ \frac{I(r_l)}{|R(r_l)|}\sum_{r' \in R(r_l)} \hat{s}(w_k, r') \right\}.$$

Деление на $|R(\mathrm{par}(r_l))|$ и $|R(r_l)|$ позволяет уменьшить вклад связи между родительской рубрикой и дочерней пропорционально количеству других дочерних рубрик. Кроме того, благодаря умножению на значимость рубрики, связи, расположенные ближе к корню дерева, учитываются с меньшим весом. Аналогично определим функцию схожести для двух рубрик:

$$s(r_k, r_l) = \max\left\{ \hat{s}(r_k, r_l),\ \frac{I(\mathrm{par}(r_k))}{|R(\mathrm{par}(r_k))|}\hat{s}(\mathrm{par}(r_k), r_l),\ \frac{I(r_k)}{|R(r_k)|}\sum_{r' \in R(r_k)} \hat{s}(r', r_l), \right.$$
$$\left. \frac{I(\mathrm{par}(r_l))}{|R(\mathrm{par}(r_l))|}\hat{s}(r_k, \mathrm{par}(r_l)),\ \frac{I(r_l)}{|R(r_l)|}\sum_{r' \in R(r_l)} \hat{s}(r_k, r') \right\}.$$

Утверждение 1. *Теорема 1 остаётся действительной для произвольных токенов.*

В разделе 2.3 вводится функция схожести двух тематических портретов. Эта функция основана на известной метрике — «мягкой» косинусной мере близости[13], которая является обобщением косинусной меры для случаев отсутствия ортогонального базиса.

Пусть p и p' — два тематических портрета. Дополним каждый портрет недостающими токенами из обучающей выборки с нулевыми весами. Тогда эти портреты можно представить в виде $\{(t_i, c_i) \mid 1 \le i \le n\}$ и $\{(t_i, c'_i) \mid 1 \le i \le n\}$. Введём следующие вспомогательные величины:

$$c_{ij} = \sqrt{s(t_i, t_j)} \cdot \frac{c_i + c_j}{2}, \quad c'_{ij} = \sqrt{s(t_i, t_j)} \cdot \frac{c'_i + c'_j}{2}.$$

Коэффициент схожести двух портретов определяется следующим образом:

$$s(p, p') = \frac{\sum_{i,j} c_{ij} c'_{ij}}{\sqrt{\sum_{i,j} c_{ij}^2} \sqrt{\sum_{i,j} c'^2_{ij}}}. \tag{4}$$

Утверждение 2. *Коэффициент схожести двух портретов равен 1 тогда и только тогда, когда они состоят из одних и тех же токенов, с точностью до замены на токены, имеющие те же коэффициенты схожести с другими токенами обоих портретов, и веса токенов во втором портрете пропорциональны весам в первом портрете ($\forall i \; c'_i = \alpha c_i$). В остальных случаях коэффициент схожести меньше 1.*

В разделе 2.5 рассматривается вычислительная сложность функций схожести. Пусть в информационно-аналитической системе всего k различных токенов, m тематических портретов, в среднем каждый портрет состоит из n токенов ($n \ll k$). Как правило, величина m растёт быстрее всего, а порядок величины n не меняется. Автором доказана справедливость следующей оценки.

Утверждение 3. *Для вычисления коэффициента схожести двух ключевых слов требуется в среднем $O(\frac{mn^2}{k})$ операций. Для вычисления коэффициента схожести двух тематических портретов — $O(\frac{mn^4}{k})$.*

[13] Soft similarity and soft cosine measure: Similarity of features in vector space model / G. Sidorov [et al.] // Computación y Sistemas. 2014. Vol. 18, no. 3. P. 491–504. ISSN 1405-5546.

Для поиска портретов, имеющих наибольшие коэффициенты схожести с заданным портретом p, необходимо выполнить в среднем $O(\frac{m^2n^4}{k})$ операций. Предложен способ ускоренного поиска с потерей точности.

В разделе 2.6 приводится сравнение разработанной функции схожести с известными метриками, использующими другие источники данных. Для ключевых слов такой метрикой является мера Жаккара для множеств авторов публикаций, использующих каждое из двух слов, а для рубрик — длина пути между ними в дереве рубрикатора. Результаты сравнения показывают высокую степень соответствия между разработанной функцией схожести и другими метриками. В подразделе 2.6.3 описан проведённый эксперимент, показавший возможность выявлять переводы ключевых слов с помощью разработанной метрики.

В третьей главе рассматривается вопрос построения тематического портрета объекта в информационно-аналитической системе. В простейшем случае токены (ключевые слова или рубрики) могут быть сопоставлены объекту непосредственно. Однако в системах с большим количеством типов объектов и связей, для некоторых типов токены можно получить только при помощи информации из других объектов, связанных с данным. Например, тематический портрет журнала можно построить по данным о публикуемых в нём статьях, а тематический портрет учёного — по данным о результатах его деятельности.

В разделе 3.1 приводится формализация понятия информационно-аналитической системы на основе графовой модели данных. Другие распространённые модели могут быть к ней сведены. В такой модели данные представляются в виде ориентированного графа, в котором вершины соответствуют объектам, а рёбра — свойствам объектов и связям между ними. Значениями свойств могут быть числа, строки и другие значения; множество таких значений для удобства определим следующим образом.

Определение 6. *Множеством литеральных значений L* будем называть некоторое заранее заданное множество (возможно, бесконечное), не привязанное к

конкретной информационно-аналитической системе.

В частности, в L входят ключевые слова и рубрики.

Определение 7. *Информационно-аналитической системой* в контексте её описания в диссертационной работе будем называть следующую пару множеств.

- Множество N основных вершин графа (nodes).
- Множество E рёбер графа (edges), где каждое ребро является триплетом (тройкой) $e = (subj, pred, obj)$, где $subj, pred \in N$, $obj \in N \cup L$. Элементы триплетов будем называть *субъект*, *предикат* и *объект*, соответственно.

Предикат показывает тип отношения между субъектом и объектом. Например, профиль человека может быть соединён с книгой следующими предикатами: «является автором», «является редактором», «является переводчиком».

В разделе 3.2 представлен алгоритм построения тематического портрета вершины. Чтобы построить такой портрет, необходима информация о токенах, соединённых в графе с данной вершиной цепочкой из одного или нескольких рёбер. Токены, соединённые одним ребром, являются наиболее ценными, но такие токены могут существовать не всегда.

Для того, чтобы все рёбра имели единое направление, для каждого предиката $pred$ дополним множество N обратным предикатом $pred^{-1}$ и для каждого триплета $(subj, pred, obj)$, такого что $obj \in N$, дополним множество E триплетом $(obj, pred^{-1}, subj)$. После этого любую такую цепочку, если в ней для части рёбер заменить $pred_i$ на $pred_i^{-1}$, можно представить в виде последовательности рёбер, направленных от рассматриваемой вершины к токену:

$$subj \overset{pred_1}{\rightarrow} obj_1 \overset{pred_2}{\rightarrow} \ldots \overset{pred_m}{\rightarrow} t.$$

Входными данными алгоритма составления тематического портрета является множество упорядоченных наборов предикатов, где каждому набору присвоен некоторый положительный коэффициент, и сумма коэффициентов всех наборов

равна 1. Наборы будем обозначать $(pred_1, \ldots, pred_m)$, а коэффициент набора — $\sigma(pred_1, \ldots, pred_m)$.

Первым шагом алгоритма является определение для каждого из наборов всех возможных цепочек вершин $obj_1, \ldots, obj_m$, где $obj_i \in pred_i(obj_{i-1})$ для $1 \leq i \leq m$, $obj_0 = subj$, $obj_m \in L$ — некоторый токен. Будем проходить по каждой такой цепочке от последнего её элемента до первого, и на каждом шаге будем присваивать очередной вершине некоторый вес. Изначально вес токена obj_m положим равным 1. Вес obj_i должен зависеть от количества исходящих из obj_i рёбер с предикатом $pred_{i+1}$ и от количества входящих в obj_{i+1} рёбер с данным предикатом.

Определение 8. *Образом вершины $subj \in N$ при предикате $pred$* будем называть множество $pred(subj) = \{obj \in N \cup L : (subj, pred, obj) \in E\}$.

Прообразом вершины $obj \in N \cup L$ при предикате $pred$ будем называть множество $pred^{-1}(obj) = \{subj \in N : (subj, pred, obj) \in E\}$.

Вес ребра $subj \overset{pred}{\to} obj$ определим следующим образом:

$$\Omega(subj, pred, obj) = \frac{1}{|pred(subj)| \cdot |pred^{-1}(obj)|}. \tag{5}$$

Далее, если $l : obj_0 \overset{pred_1}{\to} obj_1 \overset{pred_2}{\to} \ldots \overset{pred_m}{\to} obj_m$ — цепочка рёбер, определим её вклад в вес токена obj_m как коэффициент соответствующего набора предикатов, умноженный на произведение весов всех входящих в цепочку рёбер:

$$\Omega(l) = \sigma(pred_1, \ldots, pred_m) \cdot \prod_{i=1}^{m} \Omega(obj_{i-1}, pred_i, obj_i).$$

Для каждого токена t_j, который является последним элементом хотя бы в одной цепочке, определим его вес как сумму вкладов всех таких цепочек: $c_j = \sum_{l:\ldots\to t_j} \Omega(l)$. В качестве тематического портрета $subj$ возьмём множество

$$p(subj) = \{(t_1, c_1), \ldots, (t_n, c_n)\}. \tag{6}$$

Для того, чтобы построенный портрет удовлетворял определению 2, требуется, чтобы все c_j были не больше 1. В диссертации доказано более сильное утверждение.

Теорема 2. *Пусть тематический портрет $p(subj)$ построен в соответствии с алгоритмом, описанным выше. Тогда в формуле* (6) $\sum_{j=1}^{n} c_j \leq 1$.

Справедливо также следующее симметричное утверждение.

Теорема 3. *Сумма вкладов произвольного токена в тематические портреты, построенные по общему множеству наборов предикатов, не превосходит 1.*

В четвёртой главе разработана модель, позволяющая составлять правила, в соответствии с которыми для каждой вершины может быть вычислен вещественный *коэффициент значимости*, учитывающий значения предикатов для самой вершины и для связанных с ней вершин. В применении к задаче поиска экспертов такие коэффициенты будут определять, должен ли учитываться соответствующий вершине результат деятельности при поиске, а также влиять на ранг его авторов в поисковой выдаче. Коэффициент значимости может учитывать различные показатели результата. Например, для статьи это может быть число цитирований, для книг — число печатных листов, для участия в НИР — объём финансирования.

В разделе 4.1 представлена математическая формализация модели для вычисления коэффициентов значимости. Эта модель основана на правилах, которые могут быть объединены в наборы правил.

Будем считать, что каждая вершина относится к некоторому типу. Для таких отношений будем использовать обозначения вида $obj \in \tau$. Через 2^M будем обозначать множество подмножеств множества M.

Одновременно введём два вида выражений, тесно связанных друг с другом — пути доступа и вычислимые выражения.

Определение 9. *Путями доступа* будем называть выражения, задающие функции, сопоставляющие вершине типа τ либо конечное множество других вершин из N ($f\colon \tau \to 2^N$), либо некоторое (также конечное) множество литеральных значений ($f\colon \tau \to 2^L$). Такие выражения должны быть построены одним из следующих способов.

1. Предикат, определённый для вершин типа τ — путь доступа.
2. Пусть f_1 и f_2 — два пути доступа. Тогда их композиция $f' : f'(subj) = \bigcup_{obj \in f_1(subj)} f_2(obj)$ — путь доступа.
3. Пусть $f_1 \colon \tau \to 2^M$ — путь доступа, $M \subseteq L$, $f_2 \colon M \to L$ — функция. Тогда их композиция $f' : f'(subj) = \{f_2(obj) \mid obj \in f_1(subj)\}$ — путь доступа.
4. Пусть $f_1 \colon \tau_1 \to 2^{\tau_2}$ — путь доступа, $f_2 \colon \tau_2 \to \{0, 1\}$ — вычислимое выражение, имеющее булев тип. Тогда выражение $f' : f'(subj) = \{obj \in f_1(subj) \mid f_2(obj) = 1\}$ — путь доступа (выражение f_2 в этом случае будем называть *выражением фильтрации*).

Путь доступа f будем называть *однозначным*, если его значения состоят не более, чем из одного элемента: $\forall obj \in \tau \; |f(obj)| \leq 1$.

Определение 10. *Вычислимыми выражениями* будем называть выражения, задающие функции, сопоставляющие вершине типа τ некоторое вещественное число: $f \colon \tau \to \mathbb{R}$. Такие выражения должны быть построены одним из следующих способов.

1. Константа $a \in \mathbb{R}$ — вычислимое выражение.
2. Пусть $f_1 \colon \tau \to 2^{\mathbb{R}}$ — путь доступа со значениями в $\mathbb{R}$, $f_2 \colon 2^{\mathbb{R}} \to \mathbb{R}$ — функция, определённая для конечных подмножеств. Тогда композиция $f' = f_2 \circ f_1$ является вычислимым выражением для типа τ. Функцию f_2 в этом случае будем называть *агрегатной функцией*. Примеры агрегатных функций: число элементов в множестве; сумма, максимум или минимум элементов. В частности, однозначный путь доступа со значениями в $\mathbb{R}$ сам по себе является вычислимым выражением.
3. Пусть f_1 и f_2 — два вычислимых выражения для одного типа, g — некоторый бинарный оператор ($g \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$). Тогда функция $f' : f'(subj) = g(f_1(subj), f_2(subj))$ — вычислимое выражение. Примеры бинарных операторов:
   - численные операторы: $+$, $-$, $\times$, $\div$, max, min;

- логические операторы, применимые к вычислимым выражениям, имеющим булев тип: $\wedge$, $\vee$;
- операторы сравнения (сами имеют булев тип): $=$, $<$, $\leq$, $>$, $\geq$.

4. Пусть f_1 — вычислимое выражение, f_2 — путь доступа для того же типа со значениями в $2^{\mathbb{R}}$. Тогда следующая функция f' является вычислимым выражением (имеющим булев тип):

$$f'(subj) = \begin{cases} 1 & \text{если } f_1(subj) \in f_2(subj), \\ 0 & \text{иначе.} \end{cases}$$

Определение 11. *Правилом* будем называть пару (τ, f), где:

- τ — некоторый тип вершин информационно-аналитической системы;
- $f : \tau \to \mathbb{R}$ — вычислимое выражение, значения которого будут коэффициентами значимости вершин.

Будем называть правило (τ, f) *применимым к вершине obj*, если $obj \in \tau$. *Результатом применения* правила к вершине будем называть число $f(obj)$.

В подразделе 4.1.1 рассматривается вычислительная сложность применения правила к вершине. Сформулировано утверждение, оценивающее количество операций взятия образа вершины при предикате.

В подразделе 4.1.2 предлагается схема построения вычислимых выражений для расчёта коэффициентов значимости результатов деятельности, которые можно использовать при поиске экспертов.

1. Взятие в качестве начального коэффициента значимости некоторой вещественной константы.
2. Умножение на вычислимое выражение, соответствующее индикатору, который необходимо учесть.
3. Умножение на одно или несколько вычислимых выражений, принимающих значения 0 или 1, для исключения нежелательных результатов. Такие выражения будем называть *ограничивающими*.

4. Учёт типа авторства и вклада автора для объектов, имеющих несколько авторов. В простых случаях это может быть деление на число соавторов. Если информация об авторстве хранится в промежуточной вершине графа, то вычислимые выражения можно применять к такой вершине.
5. Если требуется, чтобы коэффициент значимости был ограничен сверху, это реализуется при помощи оператора min и заданной константы.

Приведём пример вычислимого выражения для статей в журналах, которое построено по описанной схеме: «импакт-фактор журнала за 2020 год, разделённый на число авторов статьи, при условии, что объём статьи не менее 5 страниц».

В разделе 4.2 представлен подход, позволяющий использовать вычислимые выражения на практике в информационно-аналитических системах, использующих реляционную модель данных. В рамках такого подхода отношения соответствуют типам в графовой модели, каждый кортеж отношения соответствует некоторой вершине, а каждый атрибут отношения — некоторому предикату. Если типы образуют иерархию с несколькими уровнями, то для любого пути в дереве типов от корня к листовой вершине ровно одному элементу пути должно соответствовать отношение.

Пути доступа могут быть представлены в виде запросов к СУБД на языке SQL, а вычислимые выражения — в виде SQL-выражений. Вместо генерации непосредственно текста SQL-запроса, можно использовать промежуточное представление запроса в виде абстрактного синтаксического дерева или в виде конструкций различных языков программирования. Такой подход используется и в предлагаемой автором программной реализации.

В подразделе 4.2.1 решается вспомогательная задача — по заданному набору правил определить, какие из правил применимы к вершине, и для каждого правила вычислить значения всех ограничивающих выражений, входящих в состав этого правила. Наличие в информационно-аналитической системе такого сервиса позволит пользователю понять, почему конкретный результат деятельности по-

лучает тот или иной коэффициент значимости, и можно ли его улучшить, внеся недостающие данные.

В подразделе 4.2.2 описывается схема, при помощи которой можно представить правила и их наборы в структурированном виде, для хранения наборов и их передачи по сети. Схема основана на представлении правил как части графа информационно-аналитической системы и использовании модели RDF, которая позволяет сериализовать данные в XML, JSON, Turtle и других форматах. Приведён пример представления правила в формате JSON.

В пятой главе описывается алгоритм поиска результатов деятельности и экспертов. Входными данными для алгоритма являются поисковый запрос, заданный тематическим портретом q, и набор правил. Портрет q при необходимости может быть автоматизированно построен по заданному объекту.

В разделе 5.1 описываются шаги алгоритма поиска. Сначала составляется список результатов, имеющих ненулевой коэффициент схожести с q. Это результаты, в портретах которых есть токены, имеющие общих соседей с токенами из q. Далее для каждого результата проверяется его соответствие набору правил. Пусть $(\tau_1, f_1), \ldots, (\tau_k, f_k)$ — такой набор, f_0 — общее для набора ограничивающее выражение. Введём следующую вспомогательную функцию:

$$f_i'(obj) = \begin{cases} f_i(obj) \cdot f_0(obj) & \text{если } obj \in \tau_i, \\ 0 & \text{если } obj \notin \tau_i. \end{cases}$$

Вклад результата *obj* в ранг автора *subj* определим как

$$\lambda(obj, subj) = s(p(obj), q) \cdot \max_{1 \le i \le k} f_i'(obj). \tag{7}$$

Далее составляется список потенциальных экспертов, которыми являются авторы результатов из списка, описанного выше. Ранг эксперта будем вычислять следующим образом (*authorOf* — предикат «является автором»):

$$\lambda(subj) = \sum_{obj \in authorOf(subj)} \lambda(obj, subj). \tag{8}$$

В разделе 5.2 рассматривается задача выявления конфликта интересов. Например, в случае экспертизы темы исследований в поисковую выдачу могут попасть сами авторы заявки, их коллеги или соавторы. Предложен механизм для выявления таких ситуаций на основе путей доступа. В случае, если в выдачу попал эксперт, соответствующий пути доступа, его ранг может быть понижен умножением на заданный коэффициент, или он может быть исключён из выдачи.

В разделе 5.3 описана разработанная автором программная реализация алгоритма поиска экспертов. Сервис поиска реализован в виде модуля информационно-аналитической системы «ИСТИНА». Программный код написан на языке Python с использованием инструментального средства Django. При поиске учитываются: статьи в журналах и сборниках, участие в научно-исследовательских работах, защищённые кандидатские и докторские диссертации. Проведено исследование производительности реализации.

В Заключении изложены основные результаты диссертационной работы и сформулированы выводы.

Заключение

В диссертации получены следующие основные результаты.

- В рамках модели тематических портретов, основанных на токенах — ключевых словах и рубриках, введены функции схожести отдельных токенов и основанная на них функция схожести самих портретов. Приведены оценки вычислительной сложности, проведено сравнение с другими метриками.
- Разработан алгоритм, позволяющий при наличии заранее заданных наборов предикатов составить тематический портрет вершины по ключевым словам и рубрикам, связанным с ней цепочками рёбер с заданными предикатами. Доказательно описаны свойства построенных таким образом портретов.
- Предложена модель, позволяющая динамически определять правила для отбора вершин и сопоставления им коэффициентов значимости, учитывающих

значения предикатов для самих вершин и для вершин, связанных с ними путями заданного вида. Описан метод построения запросов к реляционным СУБД на основе таких правил.

- Описан алгоритм, учитывающий тематическую схожесть и коэффициенты значимости результатов для составления списка потенциальных экспертов и присвоения им рангов. Создана программная реализация этого алгоритма.

Благодарность

Автор благодарит своего научного руководителя — профессора, доктора физико-математических наук В. А. Васенина, а также доцента, кандидата физико-математических наук С. А. Афонина за постановку задач и внимание к работе на всех этапах её выполнения.

Основные публикации по теме диссертации

Научные статьи, опубликованные в рецензируемых научных изданиях, рекомендованных для защиты в диссертационном совете МГУ по специальности 05.13.17 «Теоретические основы информатики»:

1. *Афонин С. А.*, *Козицын А. С.*, *Шачнев Д. А.* Программные механизмы агрегации данных, основанные на онтологическом представлении структуры реляционной базы наукометрических данных // Программная инженерия. — Москва, 2016. — т. 7, № 9. — с. 408—413. — ISSN 2220-3397. — DOI: 10.17587/prin.7.408-413. — RSCI (импакт-фактор РИНЦ 2020: 0,459).
2. *Шачнев Д. А.* Searching for Activity Results and Experts in a Given Subject Area, Taking Results Significance into Account // Программная инженерия. — Moscow, 2021. — т. 12, № 5. — с. 260—266. — ISSN 2220-3397. — DOI: 10.17587/prin.12.260-266. — RSCI (импакт-фактор РИНЦ 2020: 0,459).

3. *Shachnev D.*, *Karpenko D.* Using Subject Area Ontology for Automating Processes in Sphere of Scientific Investigation and Education // Programming and Computer Software. — Road Town, United Kingdom, 2018. — Vol. 44, no. 1. — P. 15–22. — ISSN 0361-7688. — DOI: 10.1134/S0361768818010061. — Web of Science (Impact Factor 2020: 0,936).

Иные публикации:

4. *Шачнев Д. А.*, *Афонин С. А.*, *Козицын А. С.* Использование онтологического представления структуры реляционной базы для агрегации наукометрических данных // Научный сервис в сети Интернет: труды XVIII Всероссийской научной конференции (19–24 сентября 2016 г., г. Новороссийск). — Москва : ИПМ им. М.В. Келдыша, 2016. — с. 58—63. — ISBN 978-5-98354-027-9. — DOI: 10.20948/abrau-2016-1.

5. *Шачнев Д. А.* Онтология предметной области для научных исследований и автоматизации учебного процесса: методы реализации и алгоритмы использования // Материалы Всероссийской конференции с международным участием «Знания-Онтологии-Теории» (ЗОНТ-2017). т. 2. — Новосибирск : ООО «Дигит Про», 2017. — с. 148—155.

6. *Shachnev D.* Web ontology editor: architecture and applications // 5th International Conference on Actual Problems of System and Software Engineering, APSSE 2017. Vol. 1989. — CEUR Workshop Proceedings (CEUR-WS.org), 2017. — P. 342–350.

7. Methods for Intelligent Data Analysis Based on Keywords and Implicit Relations: The Case of “ISTINA” Data Analysis System / D. Shachnev [et al.] // 2019 Actual Problems of Systems and Software Engineering (APSSE). — IEEE, 2019. — P. 157–161. — ISBN 978-1-7281-6061-0. — DOI: 10.1109/APSSE47353.2019.00027.

8. *Шачнев Д. А.* Программные механизмы автоматической генерации SQL–запросов в ИАС «ИСТИНА»: особенности и варианты использования // Ломоносовские чтения. Секция механики. Тезисы докладов. — Москва : Издательство Московского университета, 2018. — с. 196—197. — ISBN 978-5-19-011337-2.

9. Методы и средства тематического анализа данных в больших системах на основе ключевых слов и косвенных связей между ними / Д. А. Шачнев [и др.] // Ломоносовские чтения. Секция механики. Тезисы докладов. — Москва : Издательство Московского университета, 2019. — с. 51—51. — ISBN 978-5-19-011444-7.

10. *Шачнев Д. А.* Новая версия pmodel — генератора SQL-запросов, основанного на онтологическом представлении структуры базы данных информационной системы, с использованием Django ORM // Ломоносовские чтения. Секция механики. Тезисы докладов. — Москва : Издательство Московского университета, 2020. — с. 207—208. — ISBN 978-5-19-011565-9.

В работах [1 ; 4] С. А. Афониным сформулирована задача, начальный алгоритм её решения создан авторами совместно, далее был существенно доработан и описан Д. А. Шачневым. В работе [3] Д. А. Шачневым разработаны механизмы для хранения и редактирования онтологических данных в системах, использующих реляционные СУБД; Д. С. Карпенко описаны приложения в образовательной сфере. В работах [7 ; 9] К. В. Лунёвым разработаны методы анализа ключевых слов; Д. А. Шачневым представлен метод использования косвенных связей между объектами и ключевыми словами и описаны возможности применения методов в информационно-аналитической системе; В. А. Васениным и С. А. Афониным были сформулированы задачи и внесены улучшения в текст.