ИПМ им.М.В.Келдыша РАН

Абрау-2022 • Труды конференции

Д.А. Шачнев

Метод поиска и ранжирования связанных объектов в наукометрических системах

Рекомендуемая форма библиографической ссылки

Шачнев Д.А. Метод поиска и ранжирования связанных объектов в наукометрических системах // Научный сервис в сети Интернет: труды XXIV Всероссийской научной конференции (19-22 сентября 2022 г., онлайн). — М.: ИПМ им. М.В.Келдыша, 2022. — С. 526-538.
https://doi.org/10.20948/abrau-2022-32
https://keldysh.ru/abrau/2022/theses/32.pdf

Видеозапись выступления

Презентация к докладу

Метод поиска и ранжирования связанных объектов в наукометрических системах

Д.А. Шачнев

МГУ имени М.В. Ломоносова, НИИ механики

Аннотация. Современные информационные системы, работающие с научной информацией, как правило, имеют дело с большим количеством типов объектов и связей между ними. Например, объект «научная публикация» связан, с одной стороны, с одним или несколькими объектами «учёный», с другой стороны — с объектом «научный журнал» или «сборник», которые, в свою очередь, связаны с объектами «издательство» или «серия сборников». Наличие таких связей позволяет объединить все объекты системы в единый граф, а информацию о связях в таком графе использовать для различных задач, таких как: определение тематической близости двух объектов, отнесение объекта к той или иной предметной области, определение значимости объекта.

В работе описаны методы, применяемые в информационно-аналитической системе «ИСТИНА», разработанной в МГУ имени М.В. Ломоносова, для тематического поиска объектов (научных публикаций и экспертов) по ключевым словам и рубрикам классификаторов, а также ранжирования результатов поиска с учётом значимости каждого результата.

Ключевые слова: CRIS-системы, поиск экспертов, предметная область, тематический поиск, ранжирование, семантическая схожесть

Method for Search and Ranking of Related Objects in Scientometric Systems

D.A. Shachnev

Lomonosov Moscow State University, Institute of Mechanics

Abstract. Modern information systems used in scientific sphere usually deal with large numbers of object types and relationships between them. For example, a “scientific publication” object is linked, on one side, with one or multiple “scientist” objects, and on the other side with “scientific journal” or “collection” object, which in turn are linked to “publishing house” or “collection series” objects. Having such relationships allows us to represent all objects in a system as a graph, and to use information about relationships in this

graph to solve various problems, such as: determining the thematic similarity of two objects, assigning an object to a certain subject area, assessing the importance of an object.

The paper describes the methods used in the “ISTINA” information analysis system, developed in Lomonosov Moscow State University, for thematic search of objects (scientific publications and experts) by keywords and classification elements, as well as ranking search results, taking into account the significance of each result.

Keywords: CRIS, expert search, subject area, thematic search, semantic similarity

Введение

Будем предполагать, что наукометрическая информационная система использует систему управления базами данных (СУБД), основанную на реляционной модели данных. В такой модели данные хранятся в таблицах, где каждая таблица соответствует некоторому типу объектов (например, «автор» или «статья») либо типу связи между объектами (например, «авторство статьи»), каждая строка таблицы соответствует конкретному объекту, а столбец соответствует некоторому свойству объекта (например, для таблицы «статьи» столбцами могут быть название, DOI и год выхода).

Отношения между объектами будут определяться внешними ключами, то есть столбцами, содержащими первичный ключ строки в связанной таблице. Например, у таблицы «авторство статьи» будут внешние ключи на таблицы «автор» и «статья».

Такое устройство имеет информационно-аналитическая система «ИСТИНА» (Интеллектуальная Система Тематического Исследования НАукометрических данных), используемая в Московском государственном университете имени М.В. Ломоносова и ряде других вузов, институтов и центров Российской академии наук и Минздрава РФ. В этой системе собирается, систематизируется и анализируется информация о научной, педагогической, инновационной и иной деятельности сотрудников подключенных организаций.

Типы объектов, хранящихся в ИАС «ИСТИНА», включают в себя:

- статьи в журналах и сборниках, тезисы докладов (по состоянию на июль 2022 года — более 1 миллиона);
- доклады на конференциях (более 300 тысяч);
- книги (монографии, учебные пособия, научно-популярная литература и т.п.);
- кандидатские и докторские диссертации;
- патенты и свидетельства о регистрации прав на ПО;
- участие в научно-исследовательских, опытно-конструкторских и технологических работах (НИОКТР);

- членство в редколлегиях журналов и сборников, в оргкомитетах и программных комитетах конференций;
- научное руководство студентами и аспирантами;
- читаемые учебные курсы и авторства курсов (более 110 тысяч курсов);
- членство в диссертационных советах.

1. Тематические портреты объектов и их построение

В данном разделе будет описана структура, позволяющая компактно хранить информацию для идентификации тематической принадлежности объекта. Такую структуру будем называть *тематическим портретом*. Также будет представлена функция для определения схожести двух тематических портретов в численном виде. В случае, если поисковый запрос будет представлен в виде такой же структуры, данная функция позволит определять релевантность объектов из базы поисковому запросу.

Исходя из данных, которые имеются в системе (в основном это мета-данные, полные тексты статей доступны лишь в некоторых случаях, и формат файлов не стандартизирован), для построения тематических портретов была выбрана модель, основанная на ключевых словах и элементах рубрикаторов, таких как ГРНТИ и УДК.

Тематический портрет является множеством ключевых слов и рубрик с присвоенными им весовыми коэффициентами: $\{(w_1, c_1), \dots, (w_n, c_n)\}$, где $c_i > 0$. Для каждого такого портрета p через $W(p)$ обозначим множество $\{w_1, \dots, w_n\}$. В дальнейшем для упрощения будем говорить только о ключевых словах (рубрики учитываются аналогично им).

Такие множества могут быть сформированы автоматизированным способом для объектов различных типов. Например, для построения тематического портрета научного журнала можно взять ключевые слова статей, опубликованных в нём. Весовые коэффициенты в этом случае можно установить таким образом, что каждая статья даёт суммарный вес 1 всем ключевым словам, которые в ней используются (то есть к весу каждого ключевого слова в портрете эта статья добавляет $\frac{1}{N}$, где N — число ключевых слов в статье). Такая схема позволяет равномерно учитывать вклад статей как с большим, так и с маленьким количеством ключевых слов.

Метод оценки тематической схожести двух портретов был впервые представлен в работе [1] и получил развитие в настоящей статье. В основе метода лежит дистрибутивная гипотеза, которая связывает семантическую близость двух ключевых слов с числом общих контекстов, в которых они встречаются. В рамках наукометрических систем такими контекстами могут быть, например, тематические портреты статей. Будем предполагать, что заранее задано некоторое множество P таких портретов, которые будут использоваться при расчёте схожести. Для каждого ключевого слова w

через $P(w)$ обозначим множество портретов из P, содержащих w: $P(w) = \{p \in P: w \in W(p)\}$.

Сначала определим вес связи между ключевыми словами, которые встречаются вместе хотя бы в одном портрете из P:

$$\omega(w_k, w_l) = \sum_{p \in P(w_k) \cap P(w_l)} \frac{c_l}{\sum_{w_i \in W(p)} c_i}.$$

Далее будет удобно рассматривать вспомогательный граф, вершинами которого являются отдельные ключевые слова, и две вершины соединены ребром в том случае, если соответствующие им ключевые слова встречаются совместно в одном из портретов множества P. В основе функции схожести двух ключевых слов w_i и w_j лежит множество общих «соседей» двух этих слов и веса связей каждого из слов-соседей с первым и со вторым словом. Множество общих соседей определим следующим образом:

$$W_{i,j} = \{w: |P(w_i) \cap P(w)| \geq 1, |P(w_j) \cap P(w)| \geq 1\}.$$

Функцией, позволяющей в первом приближении оценить схожесть двух ключевых слов или рубрик, является ненормированная функция схожести. Она определяется при помощи суммирования по всем цепочкам из двух рёбер, соединяющих w_i и w_j, весов связей, соответствующим обеим рёбрам:

$$s_0(w_i, w_j) = \sum_{w_k \in W_{i,j}} \left(\omega(w_i, w_k) + \omega(w_j, w_k)\right).$$

Чтобы частотность встречаемости ключевого слова не повышала коэффициенты его схожести с другими словами, проведём нормировку по обратной частоте его встречаемости. Нормированная функция схожести двух ключевых слов будет выглядеть так:

$$s(w_i, w_j) = \frac{s_0(w_i, w_j)}{|P(w_i)| + |P(w_j)|}.$$

Можно показать (см. работу [2]), что областью значений такой функции является диапазон $[0,1]$, и значение равно 1 тогда и только тогда, когда множества соседей обоих ключевых слов совпадают.

На основе функции схожести для отдельных ключевых слов можно определить функцию схожести для тематических портретов. При сравнении векторных представлений текста, в качестве метрики схожести обычно используется косинус угла между соответствующими векторами в R^n [3]. Однако, использование такой метрики предполагает, что все ключевые слова, являющиеся компонентами векторов, независимы друг от друга и образуют ортогональный базис. Это не так, если между какими-то из слов есть ненулевая схожесть. Для таких случаев разработано обобщение косинусной меры близости, которое известно как мягкая косинусная мера [4].

Применительно к тематическим портретам эту меру можно определить следующим образом. Для каждого слова w_i, пусть c_i — его вес в p (или 0, если $w_i \notin W(p)$), а c'_i — его вес в p'. Введём следующие вспомогательные коэффициенты:

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

Функцию схожести двух тематических портретов определим следующим образом:

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

Аналогично функции схожести ключевых слов, можно показать, что эта функция принимает значения в диапазоне $[0,1]$.

Для оценки качества работы разработанной функции схожести ключевых слов и рубрик было проведено сравнение с естественной метрикой, определённой для рубрик внутри одного классификатора — длиной пути между рубриками в дереве классификатора. В качестве такого классификатора был использован ГРНТИ (Государственный рубрикатор научно-технической информации). Версия рубрикатора, используемая в ИАС «ИСТИНА», насчитывает 7795 рубрик, однако лишь 3628 из них используются.

Все возможные пары рубрик были разбиты на группы в зависимости от длины пути между ними. ГРНТИ имеет три уровня, поэтому с добавлением корневой вершины максимальная длина пути равна 6. Внутри каждой группы были рассчитаны коэффициенты схожести между парой рубрик (там, где достаточно данных для такого расчёта) и взято их среднее арифметическое. Результаты показаны на рис. 1. Из графика следует, что среднее значение схожести убывает практически равномерно с увеличением длины пути.

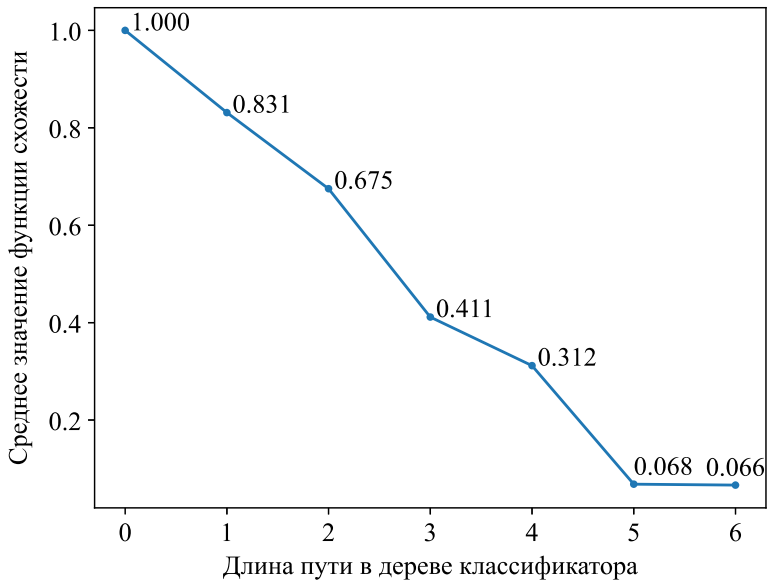

Рис. 1. Сравнение коэффициента схожести с длиной пути

2. Вычисление значимости объектов

В наукометрических системах распространены различные показатели значимости объектов. Это может быть число цитирований для статьи, импакт-фактор или квартиль для журнала, объём финансирования для НИОКТР и так далее.

Для реализации потребностей разных групп пользователей ИАС «ИСТИНА», основной целью при разработке являлось не создание новых показателей и не использование какого-то конкретного показателя, а предоставление пользователям возможности использовать различные удобные для них наукометрические показатели и их комбинации. Такой принцип применяется и при расчёте персонального рейтинга [5], и при поиске результатов деятельности, и при поиске экспертов в заданной предметной области.

Доступные для использования показатели определяются отдельно для каждого типа объектов. Значения показателей могут быть численные, булевы или относиться к перечислимым типам.

В качестве примера приведём доступные для использования показатели статей в научных журналах:

- вид статьи: научная, научно-популярная, рецензия, краткое сообщение, тезисы (перечислимый тип);
- количество страниц (численный);
- количество печатных листов (численный);
- количество авторов (численный);
- порядковый номер текущего автора (численный);
- является ли статья оригинальной или переводной версией (булев);

- наличие индексации статьи в системах WoS, Scopus, РИНЦ (булевы);
- вхождение журнала в список рекомендованных ВАК (булев);
- импакт-факторы журнала WoS, SJR, РИНЦ (численные);
- позиции журнала внутри своих рубрик WoS, SJR (численные);
- количество цитирований статьи по данным WoS, Scopus (численные);
- наличие у статьи DOI (булев);
- наличие у журнала ISSN (булев).

Такой спектр показателей открывает богатые возможности для их комбинации. Кроме того, у пользователей системы может возникнуть необходимость использовать различные правила для расчёта показателя значимости. Например, в зависимости от вхождения журнала в различные индексирующие системы использовать показатели именно этих систем.

Для реализации такой возможности была разработана модель *наборов правил*. Каждое правило определяет тип объектов, дополнительные требования к объекту (в примере выше — вхождение в индексирующие системы) и формулу расчёта коэффициента значимости.

Дополнительные требования представляют собой логическое выражение, в котором можно использовать как булевы и перечислимые показатели, так и численные (например, «число страниц > 10»). Формула расчёта представляет собой арифметическое выражение, в котором можно использовать численные показатели данного типа объектов. Пример такой формулы: $100 \cdot N_{\text{cit}} \div N_{\text{aut}}$, где N_{cit} — количество цитирований, N_{aut} — количество авторов.

В наборе может быть одно или несколько правил для каждого типа объектов. При этом один объект может удовлетворять дополнительным требованиям сразу нескольких правил. В таком случае будет использоваться то правило, у которого значение по формуле расчёта наибольшее. Например, в наборе может быть правило для статей в высокорейтинговых журналах и более общее правило для всех статей. В этом случае если статья опубликована в высокорейтинговом журнале, то для неё будет использоваться первое правило, а если нет, то второе.

Опишем, как устроена техническая реализация механизма расчёта коэффициента значимости по набору правил. Разработчиками системы определяется список доступных для использования типов объектов и показателей для каждого типа, а также возможных значений показателей перечислимых типов. В некоторых случаях объекты разных типов могут физически храниться в одной таблице в СУБД. В этих случаях должно быть определено условие, определяющее, является ли запись в таблице объектом данного типа. Для каждого показателя указывается его название, тип и соответствующее ему SQL-выражение. Это может быть название столбца в таблице или более сложное выражение, например, вложенный запрос. В качестве альтернативы языку SQL можно описывать такие выражения в виде конструкций языков программирования общего

назначения. Например, для языка Python существуют библиотеки SQLAlchemy [6] и Django ORM, которые преобразовывают такие конструкции в выражения SQL.

Набор правил описывается структурой в формате JSON. Он представляет собой список, элементами которого являются записи, соответствующие правилам. У каждой записи есть следующие ключи:

- `label` — название правила, определённое пользователем;
- `type` — кодовое название типа объектов;
- `restrictions` — список дополнительных требований;
- `value` — формула расчёта коэффициента значимости.

Ограничения являются записями JSON, которые имеют следующие ключи.

- Для всех типов показателей: `property` и `value` для определения, какой показатель следует ограничить, и его желаемое значение (при преобразовании в SQL используется оператор `=`).
- Для численных показателей: `property`, `lower_bound` и/или `upper_bound` для установки нижней и/или верхней границы допустимых значений показателя (используются операторы `>=` и `<=`).
- Для показателей перечислимых типов: `property`, `allowed_values` или `disallowed_values` для определения списка допустимых либо недопустимых значений (используются операторы `IN` и `NOT IN`).

Формула расчёта коэффициента значимости представляется в виде древовидной структуры, где листовыми вершинами являются кодовые названия показателей или числа, а остальными вершинами — арифметические и логические операторы.

Формулу из примера выше можно представить следующим образом:

```
{
  "type": "Divide",
  "left": {
    "type": "Multiply",
    "left": 100,
    "right": "num_citations"
  },
  "right": "num_authors"
}
```

На основе каждого правила из набора формируется SQL-запрос, результатом выполнения которого является список объектов, соответствующих правилу. Столбцами, которые возвращает запрос, являются поля, идентифицирующие объект (его тип, название, авторы, дата), а также рассчитанный по формуле коэффициент значимости. Дополнительные требования учитываются в поле `WHERE` запроса.

Для каждого типа объектов в структуре, которая его описывает, можно определить функцию, вносящую дополнительные изменения в запрос. Например, такая функция может добавлять дополнительные поля

из таблицы или информацию из других таблиц для показа пользователю расширенной информации в карточке объекта.

Для удобства отладки программной реализации и самих наборов правил, был также разработан вспомогательный инструмент для проверки того, каким из дополнительных требований соответствует данный объект, а каким нет. В этом инструменте используется модификация SQL-запроса, которая выбирает те столбцы, которые используются в требованиях. После обработки результатов запроса инструмент формирует для пользователей список требований, которым объект удовлетворяет, а для тех, которым он не удовлетворяет — указывает причину. Приведём примеры таких сообщений:

- значение показателя «число страниц» меньше нижней границы 10;
- отсутствует показатель «наличие у статьи DOI»;
- показатель «вид статьи» имеет значение «тезисы», которое исключено правилом.

3. Использование алгоритмов в наукометрической системе

Описанные выше механизмы позволяют формировать списки объектов, удовлетворяющих определённым требованиям, а также вычислять для каждого объекта два коэффициента: коэффициент релевантности запросу и коэффициент значимости. Эти механизмы и их комбинации внедрены в производственную работу ИАС «ИСТИНА» и используются в перечисленных ниже модулях.

В профиле пользователя показывается список результатов деятельности, принадлежащих пользователю, с разбивкой по типам (например, статьи разбиваются на статьи в журналах, статьи в сборниках и тезисы конференций). Каждый пользователь может также указать свои области интересов и результаты деятельности, относящиеся к каждой области (рис. 2).

В персональном рейтинге пользователя и в рейтингах структурных подразделений и организаций используются наборы правил, в соответствии с которыми вычисляется коэффициент значимости каждого объекта и суммарный рейтинговый балл учёного, подразделения или организации.

Ответственным пользователям системы от подразделений и организаций предоставляется раздел «Список публикаций» (рис. 3), в котором можно получить список статей или книг подразделения, с возможностью гибкой настройки требований и большим количеством столбцов, которые можно включить в таблицу. В данном разделе предусмотрена настройка «Установить опции для отчёта в Минобрнауки», которой активно пользуются некоторые научные организации, например, Институт проблем механики имени А.Ю. Ишлинского РАН.

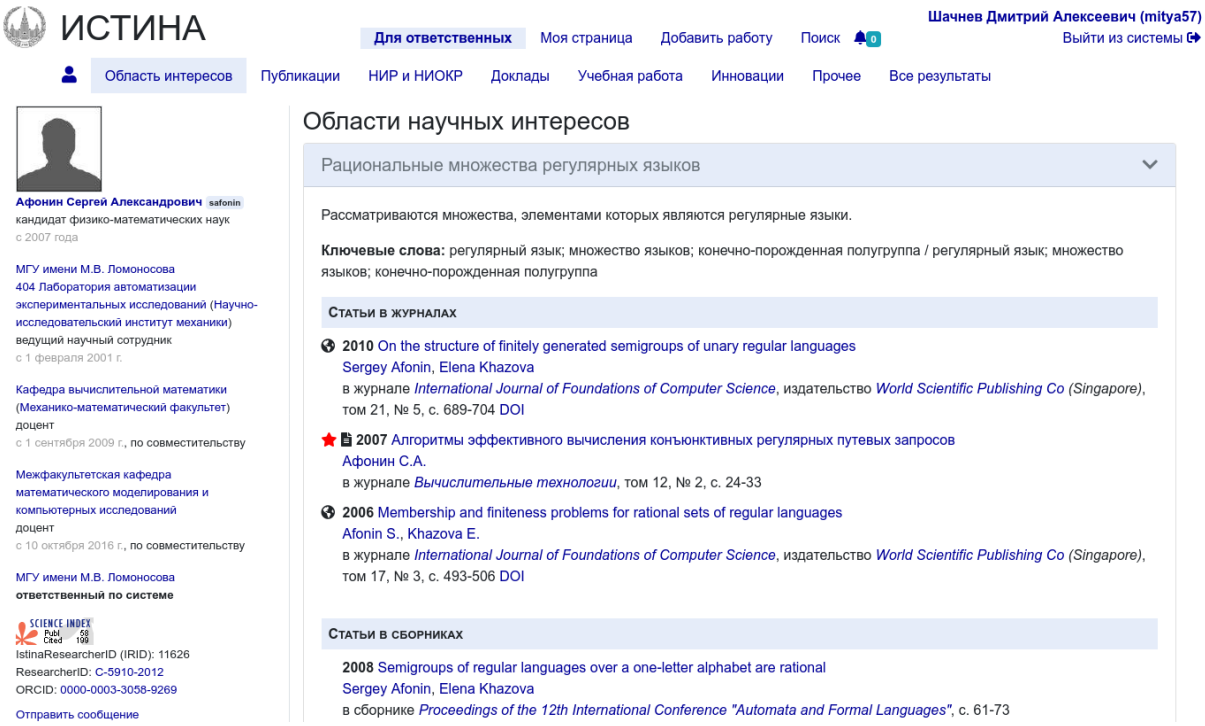

Рис. 2. Области интересов в профиле пользователя

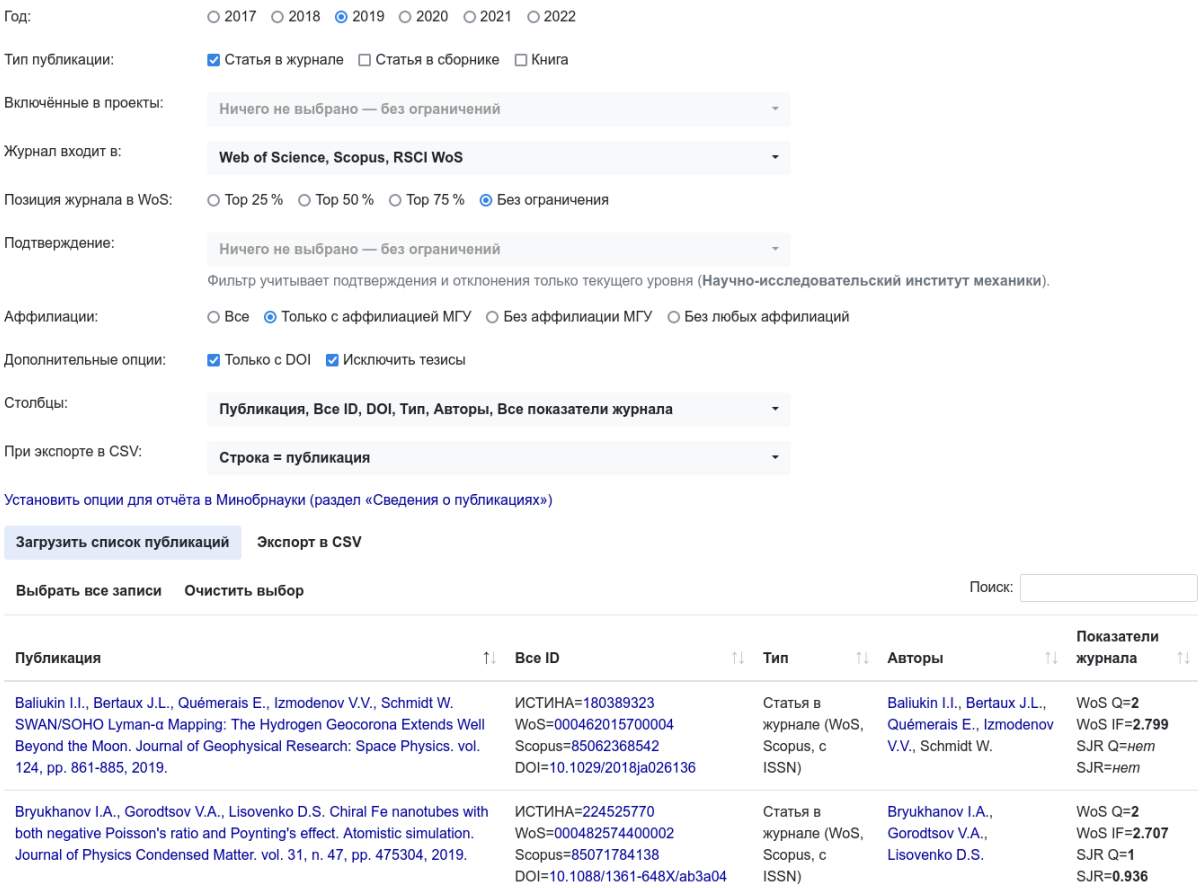

Рис. 3. Список публикаций подразделения

Относительно новой разработкой является интерфейс поиска экспертов в предметной области, заданной ключевыми словами и рубриками классификаторов ГРНТИ и ОЭСР (рис. 4). Данный интерфейс пока не открыт для всех пользователей ИАС «ИСТИНА», однако он уже применяется, в частности, для поиска рецензентов статей, поступающих в журнал «Программная инженерия».

Поиск происходит в два шага. На первом шаге выполняется поиск объектов, коэффициент схожести тематических портретов которых с портретом поискового запроса наибольший (поддерживаются публикации, диссертации и НИОКТР), и для каждого объекта вычисляется ранг, равный произведению коэффициентов релевантности и значимости.

На втором шаге составляется множество учёных, являющихся авторами объектов, и для каждого учёного вычисляется его ранг в результатах поиска. Доступны возможные варианты распределения ранга объекта между его авторами: равномерное распределение или распределение, предоставляющее больший вес первому автору.

Ключевые слова, введённые на русском языке, автоматически переводятся на английский язык при помощи программных интерфейсов двух внешних сервисов: словарей Abbyy Lingvo и Википедии. Если в русскоязычном разделе Википедии есть статья с таким же названием, как введённое слово, и для этой статьи есть ссылка на соответствующую статью в англоязычном разделе, то название последней статьи считается переводом ключевого слова на английский язык.

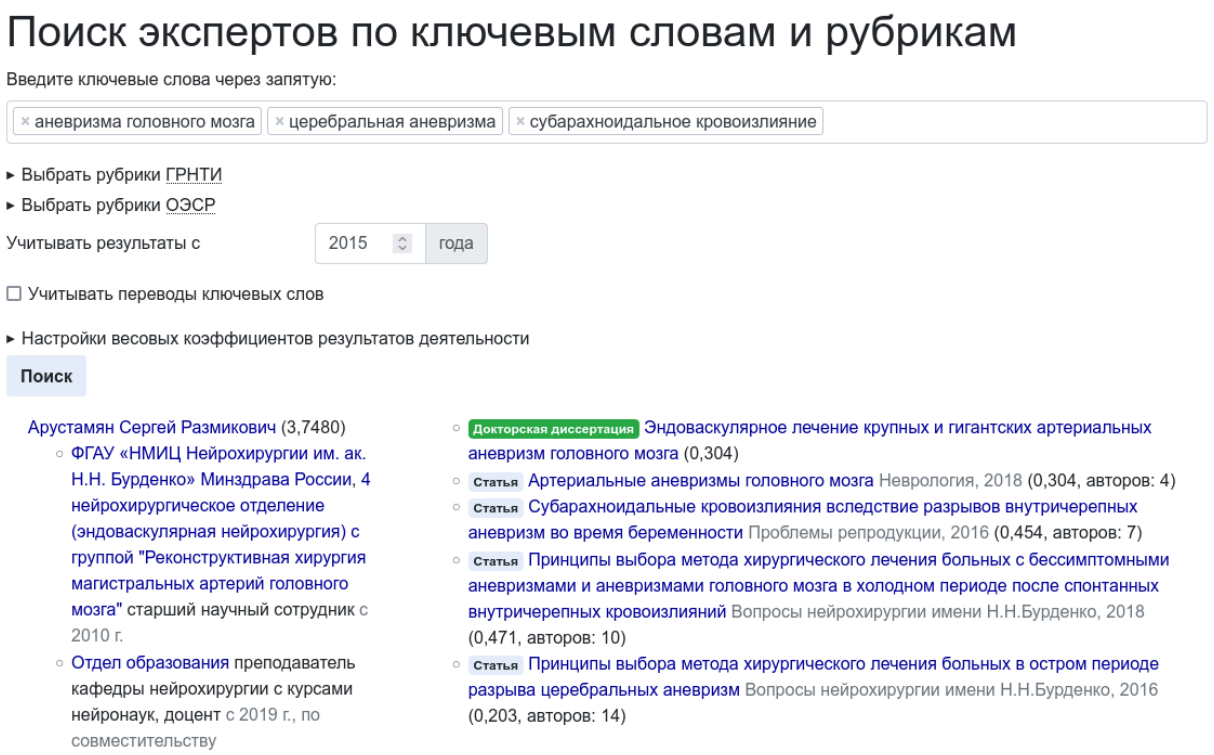

Рис. 4. Интерфейс поиска экспертов

Заключение

В работе описаны теоретические и практические методы отбора и ранжирования объектов в наукометрических информационных системах, позволяющие решать множество востребованных на практике задач. Такими задачами являются поиск результатов деятельности, формирование списков результатов по заданным критериям, поиск экспертов в предметной области. При поиске и отборе может учитываться как тематическая принадлежность объектов, так и их наукометрические показатели.

Литература

1. Лунев К.В. Теоретико-графовые алгоритмы выявления семантической близости между понятиями на основе анализа наборов ключевых слов взаимосвязанных объектов: кандидатская диссертация. Москва: МГУ имени М.В. Ломоносова, 2021. 195 с.
2. Шачнев Д.А. Семантические модели классификации и анализа данных в больших информационно-аналитических системах: кандидатская диссертация. Москва: МГУ имени М.В. Ломоносова, 2021. 112 с.
3. Singhal A. Modern Information Retrieval: A Brief Overview // IEEE Data Engineering Bulletin. 2001. Т. 24. Сс. 35–43.
4. Sidorov G. и др. Soft similarity and soft cosine measure: Similarity of features in vector space model // Computación y Sistemas. Centro de Investigación en Computación, IPN, 2014. Т. 18, № 3. Сс. 491–504.
5. 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. Сс. 58–63.
6. Bayer M. SQLAlchemy // The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few More Fearless Hacks / под ред. Brown A., Wilson G. aosabook.org, 2012.

References

1. Lunev K.V. Graph Graph-Theoretic Algorithms for Revealing Semantic Proximity Between Concepts Based on the Analysis of Keyword Sets of Interrelated Objects: PhD thesis. Moscow: Lomonosov Moscow State University, 2021. 195 pp.
2. Shachnev D.A. Semantic Models for Data Classification and Analysis in Large Information Analysis Systems. Moscow: PhD thesis, Lomonosov Moscow State University, 2021. 112 pp.

3. Singhal A. Modern Information Retrieval: A Brief Overview // IEEE Data Engineering Bulletin. 2001. Vol. 24. Pp. 35–43.
4. Sidorov G. и др. Soft similarity and soft cosine measure: Similarity of features in vector space model // Computación y Sistemas. Centro de Investigación en Computación, IPN, 2014. Vol. 18, No. 3. Pp. 491–504.
5. Shachnev D.A., Afonin S.A., Kozitsyn A.S. Using the ontological representation of the relational database structure for the aggregation of scientometric data // Scientific service on the Internet: Proceedings of the XVIII All-Russian Scientific Conference (September 19–24, 2016, Novorossiysk). Moscow: Keldysh Institute of Applied Mathematics, 2016. Pp. 58–63.
6. Bayer M. SQLAlchemy // The Architecture of Open Source Applications Volume II: Structure, Scale, and a Few More Fearless Hacks / ed. Brown A., Wilson G. aosabook.org, 2012.