Исследовательский поиск научных статей[1]

[1]*Я.Р. Недумов <yaroslav.nedumov@ispras.ru>*
[1,2,3,4] *С.Д. Кузнецов <kuzloc@ispras.ru>*

[1] *Институт системного программирования им. В.П. Иванникова РАН, 109004, Россия, г. Москва, ул. А. Солженицына, д. 25*
[2]*Московский Государственный университет имени М. В. Ломоносова Москва, 119991, ГСП-1, Ленинские горы, д. 1*
[3] *Московский физико-технический институт, 141700, Московская область, г. Долгопрудный, Институтский пер., 9*
[4] *НИУ "Высшая школа экономики", 101000, Россия, г. Москва, ул. Мясницкая, д. 20*

Аннотация. В этой статье мы анализируем современные работы, посвященные поисковому поведению ученых, и работы, посвященные методам исследовательского поиска. Мы показываем, что четыре вида поиска из шести, характерных для ученых (в том числе самый субъективно сложный – исследование новых направлений и самый часто возникающий – поиск для поддержания осведомленности), можно с достаточной степенью уверенности считать исследовательским поиском. Таким образом, поисковые системы, предназначенные для поиска научных данных, должны реализовывать специфичные для исследовательского поиска инструменты. Чтобы проверить это, мы анализируем семнадцать специализированных поисковых систем: от встроенных в электронные библиотеки Scopus и WoS до Google Scholar и социальных сетей для ученых, таких как ResearchGate и Academia.edu. Мы приходим к выводу, что степень их адаптации к специфике исследовательского поиска оставляет простор для улучшений и обсуждаем возможные направления их развития.

Ключевые слова: исследовательский поиск; поисковые системы для ученых; поисковое поведение

DOI: 10.15514/ISPRAS-2018-30(6)-10

Для цитирования: Недумов Я.Р., Кузнецов С.Д. Исследовательский поиск научных статей. Труды ИСП РАН, том 30, вып. 6, 2018 г., стр. 171-198. DOI: 10.15514/ISPRAS-2018-30(6)-10

[1] Эта работа поддержана грантом РФФИ 17-07-00978 А

1. Введение

Границы между традиционным поиском по ключевым словам и исследовательским поиском (exploratory search) довольно размыты. В своей классической работе 2006 года [1] Гари Марчионини определяет исследовательский (или разведочный) поиск от противного, сопоставляя его с традиционным поиском (lookup).

Задача традиционного поиска возникает, когда необходимо найти факты или ответы на четко сформулированные вопросы. Результаты традиционного поиска легко интерпретируемы и сравнительно компактны. Примеры запросов, характерных для традиционного поиска: сегодняшний курс валют, симптомы гриппа, в каком году была Куликовская битва и тому подобные.

Однако не все поисковые задачи можно свести к традиционному поиску. Часто пользователь поисковой системы до конца не понимает, что именно хочет найти, заранее не знает ключевых слов, изучает интересующую его тему уже в процессе поиска. Такого рода поиск может иметь сразу несколько целей, которые при этом могут изменяться по мере знакомства пользователя с предметной областью. Марчионини называет поиск такого рода *поиском для обучения* (searching to learn). Другая область исследовательского поиска связана с проведением исследований (investigation), когда пользователю требуется провести анализ, синтез, сопоставление данных из разных источников и т.д. Примером такого поиска может быть поиск с целью выбора товара для покупки или выбор места для отпуска.

В работе Уайта и Рота [2] делается попытка определить исследовательский поиск более формально. Для определения исследовательского поиска рассматриваются два аспекта: контекст проблемы (the problem context) и процесс поиска (the search process).

Для определения контекста проблемы, характерного для исследовательского поиска, авторы используют модель поисковой задачи, предложенную Марчионини в книге [3]. Марчионини рассматривает поисковую задачу в трех аспектах: специфичность цели (goal specifity), длительность (timeliness), объем результата (volume of the answer).

Специфичность цели по Марчионини характеризует степень ее определенности. Специфичность цели может быть высокой: например, если требуется найти актуальный курс валюты или прогноз погоды на завтра. А может быть низкой: например, если требуется разобраться в новой теории или выявить все точки зрения на некоторую проблему. В случае поиска фактов ищущий обычно достаточно уверен в том, нашел он уже искомый результат, или следует продолжать поиски. Напротив, когда специфичность цели низкая, трудно понять, следует ли уже закончить поиск или следует его продолжать, пытаясь получить более четкий и конкретный ответ.

Под длительностью Марчионини понимает ожидаемый необходимый объем времени для выполнения поиска. Для задач получения конкретных фактов

длительность обычно порядка нескольких секунд. Для задач изучения какой-то новой области длительность может варьироваться от месяцев и лет до бесконечности, в тех случаях, когда исследователь не рассчитывает полностью достигнуть цели. Наконец, объем характеризует сложность восприятия ответа поисковой системы. Объем может измеряться в битах информации или во времени, необходимом ищущему для его понимания. Объем может варьироваться от одиночных слов, чисел, изображений до коллекций документов или учебных курсов. Понять, что ответ поисковой системы содержит искомый курс валюты можно за несколько секунда, но если требовалось найти алгоритм решения некоторой задачи, то могут потребоваться часы, чтобы убедиться, что ответ поисковой системы релевантен запросу.

Если контекст проблемы является открытым, постоянным и многоаспектным, то есть если специфичность цели низкая, длительность поиска высокая, а объем результата большой, то такой контекст проблемы характерен для исследовательского поиска.

Второй аспект поиска, рассматриваемый Уайтом и Ротом, – процесс поиска. Для исследовательского поиска он должен быть оппортунистическим, итеративным и мультитактическим. К сожалению, они не дают более формального определения этим терминам, поэтому будем пользоваться ими в их словарных значениях. Таким образом, согласно Уайту и Роту, исследовательский поиск – это такой поиск, для которого контекст проблемы является открытым, постоянным и многоаспектным, а процесс поиска является оппортунистическим, итеративным и мультитактическим.

Интуитивно понятно, что поиск научных публикаций часто обладает многими характеристиками исследовательского поиска. Цель этой статьи – формализовать это интуитивное понимание, исследовать, какие поисковые задачи ученых можно отнести к исследовательскому поиску, какие существуют подходы к решению задачи исследовательского поиска вообще, и как они реализуются в специализированных поисковых системах для ученых.

Дальнейший текст статьи организован следующим образом. В следующем разделе мы более подробно рассмотрим, что и каким образом ищут ученые, а затем исследуем, насколько их поисковые задачи можно отнести к исследовательскому поиску. В четвертом разделе мы обсудим виды инструментов, предназначенных для ведения исследовательского поиска. В пятом разделе рассмотрим существующие поисковые системы для ученых и реализованные в них инструменты исследовательского поиска.

2. Поисковые задачи ученых

Начало исследований поискового поведения ученых прослеживают по крайней мере до сороковых годов прошлого века [4]. В классических работах Эллиса и др. [4, 5] авторы выделили восемь видов поискового поведения ученых: начало исследования (starting), анализ ссылок (chaining), просмотр (browsing),

дифференцирование (differentiating), мониторинг (monitoring), извлечение (extracting), верификация (verifying) и окончание исследования (ending).

- Начало исследования включает в себя активности по предварительному сбору информации, знакомству с предметной областью. Часто включает общение с коллегами или научным руководителем.
- Анализ ссылок – перемещение между статьями по библиографическим ссылкам, как назад, к статьям, процитированным в данной статье, так и вперед, к статьям, цитирующим данную.
- Просмотр – полунаправленное, полуструктурированное блуждание по коллекциям аннотаций с целью найти релевантные. Используется как при начальном знакомстве с областью, так и для поддержания осведомленности.
- Дифференцирование – неявное разделение публикаций на качественные и некачественные. Может базироваться как на метаданных – месте публикации, авторах, так и на анализе текста статей, использованных методах.
- Мониторинг – процесс поддержки осведомленности о происходящем в области научных интересов.
- Извлечение – систематическая проработка некоторого источника с целью извлечения релевантной информации.
- Верификация – проверка найденных фактов.
- Окончание исследования во многом повторяет начало. Хотя большая часть работы с литературой приходится на начало исследования, многие корреспонденты Эллиса возвращались к этой работе при написании отчета или статьи по завершающемуся проекту.

Авторы [4, 5] сопоставляли интервью физиков, химиков и социологов и пришли к выводу, что различия в поисковом поведении ученых разных специальностей несущественны. В более свежих работах [6, 7] тоже не обнаруживается существенная корреляция между областью научных интересов и поисковым поведением. Основные черты модели поиска [4] были подтверждены и дополнены в работе [8] десять лет спустя, после появления и широкого распространения универсальных поисковых движков, таких как Google.

В работе [9] модель поискового поведения Эллиса рассматривается в качестве формализации задачи исследовательского поиска наравне с моделью Марчионини.

Авторы более поздней статьи Атукорала и др. [10] исследовали поисковое поведение ученых в области компьютерных наук и то, как они используют те средства поиска, которые предоставляют им поисковые системы, как универсальные, так и специализированные. Компьютерные науки были выбраны в предположении, что исследователи в этой области раньше начали пользоваться электронными поисковыми системами и обладают более развитыми навыками их использования. В ходе анализа авторы выделили

четыре цели поиска при проведении научной работы: поддержка осведомленности, исследование новых направлений, обзор литературы и поиск коллег для совместной работы, а также две цели поиска при проведении образовательной работы: подготовка лекций и рекомендация статей студентам. Авторы не претендуют на то, что список целей является исчерпывающим или может быть обобщен на другие науки, но считают, что выделенные ими цели оказывают наибольшее влияние на поисковое поведение ученых.

В следующем разделе мы обсудим, насколько можно считать исследовательским каждый из видов поиска, выявленных Атукоралой с соавторами.

3. Относится ли поисковое поведение ученых к исследовательскому поиску?

В этом разделе для каждой из целей поиска ученых, выявленных в исследовании Атукоралы и др. [10], мы рассмотрим соответствующую поисковую задачу с точки зрения контекста проблемы и процесса поиска и таким образом определим, можно ли отнести ее к исследовательскому поиску или нет. Напомним, что всего были выделены шесть целей: поддержание осведомленности, исследование новых направлений, обзор литературы, подготовка лекций, рекомендация статей студентам и поиск потенциальных коллег для совместного проведения исследований. Далее мы рассмотрим каждую из них, опираясь на описания задачи и процесса поиска из исследования Атукоралы и др.

Участники исследования, осуществлявшие поиск с целью поддержки осведомленности, делали это, время от времени просматривая сайты профильных конференций, сайты издательств и личные страницы известных авторов и исследовательских групп в поисках новых статей. Оценим такой поиск с точки зрения контекста проблемы и процесса поиска.

Контекст проблемы в аспекте специфичности является открытым, потому что никогда нельзя быть уверенным, что охвачены все релевантные работы, всегда остается вероятность, что какая-то новая работа не попала в результаты поиска. Контекст является постоянным, потому что новые работы появляются все время и нельзя рассчитывать ознакомиться с ними со всеми и закончить поиск раз и навсегда. Контекст мультиаспектен, потому что результатом поиска является нечто априори неизвестное, мы не знаем, что содержится в новой статье: это может быть новый факт, лишь дополняющий существующую картину мира, а может быть новая теория, на осмысление которой уйдут годы.

Поиск с целью поддержки осведомленности по описанию авторов [10] выполняется время от времени, когда представится возможность, с использованием как систем оповещений, так и ручного просмотра личных страниц избранных авторов. Процесс поиска можно считать оппортунистическим, итеративным и мультитактическим.

Таким образом, поиск с целью поддержания осведомленности в аспекте контекста проблемы является открытым, постоянным и мультиаспектным, а в аспекте процесса поиска – оппортунистическим, итеративным и мультитактическим.

Покажем, что поиск с целью исследования новых направлений также является исследовательским поиском. Рассмотрим характеристики контекста проблемы и процесса поиска. Контекст проблемы является открытым, так как трудно определить границу, при достижении которой мы можем быть уверены, что новая область полностью изучена, а поиск окончен. Ожидаемая длительность поиска с целью исследования новых направлений сравнительно велика и может достигать нескольких лет у начинающих ученых. Хотя и нельзя сказать, что такой поиск является постоянным, все же он не является традиционным с точки зрения ожидаемой длительности. После окончания поиска с целью исследования новых направлений ученый вероятнее всего продолжит исследования в области и будет осуществлять поиск с целью поддержки осведомленности. В этом смысле эти две цели поиска можно считать продолжением одной другой. Контекст проблемы является мультиаспектным, так же, как и для поиска с целью поддержания осведомленности. Априори мы не знаем границ и объема изучаемой области.

Процесс поиска с целью исследования новых направлений также является характерным для исследовательского поиска. Он итеративен, потому что мы постепенно уточняем свое понимание новой предметной области. Он оппортунистичен, потому что у ученого обычно нет готового учебного плана для изучения новой области. Он является мультитактическим, потому что на разных этапах используются разные источники и поисковые сервисы. Многие начинают с общения с коллегами, использования Википедии и Google, только затем переходя к более специализированным источникам [10].

Поиск с целью обзора литературы возникает в момент написания обзора существующих работ. В целом он похож на поиск с целью поддержки осведомленности, но более формализован. В случае же написания систематических обзоров последовательность действий задается совсем жестко. Специфичность задана выбранной методологией, поэтому можно точно сказать, когда поиск завершен, время поиска ограничено и объем результатов тоже, хотя и может быть достаточно велик. Таким образом, хотя поиск с целью составления обзора литературы в значительно меньшей степени является исследовательским, чем поиск с целью поддержки осведомленности, он все-таки далек от традиционного. Время поиска существенно больше нескольких секунд, объем результатов велик; кроме того, после анализа отобранных по методологии статей цель поиска может быть уточнена, и тогда потребуется еще одна итерация. Таким образом, можно сделать вывод, что поиск с целью обзора литературы является исследовательским.

Поиск как при подготовке лекций, так и для рекомендации статей студентам во многом заключается в поиске уже известных ученому статей [10].

Специфичность цели здесь высока: пользователь уверен в том, что именно он хочет найти. Ожидаемое время поиска невелико (а в случае, если ищущий помнит название работы, задача сводится к традиционному поиску). Объем результата тоже невелик. Процесс поиска больше похож на характерный для исследовательского поиска. Некоторые респонденты просматривали результаты Google Scholar в поисках ссылок, по которым они уже переходили, некоторые использовали специальный BibTex файл, в котором сохраняли ссылки на все релевантные статьи. Процесс поиска не является итеративным. Таким образом, поиск с целью обучения как при подготовке лекций, так и для рекомендации статей для студентов можно считать исследовательским только с некоторой натяжкой.

Табл. 1. Сводная таблица целей использования электронных библиотек в контексте исследовательского поиска
Table 1. Summary table of the use of digital libraries in the context of exploratory search

Цель поиска	**Контекст проблемы**			**Процесс поиска**
	Специфичность	Длительность	Объем результатов	
Поиск с целью исследования новых направлений	низкая	высокая	большой	итеративный, оппортунистический, мультитактический
Поиск с целью поддержания осведомленности	низкая	высокая	большой	итеративный, оппортунистический, мультитактический
Поиск с целью обзора литературы	высокая	высокая	большой	задан используемой методологией
Поиск с целью сотрудничества	высокая	средняя	большой	итеративный, оппортунистический, мультитактический
Поиск с целью обучения (подготовки лекций, рекомендации статей студентам)	высокая	низкая	небольшой	оппортунистический, мультитактический

Наконец, последний вид поиска: поиск рецензентов или потенциальных коллег для совместной работы. В этом случае специфичность цели высока: пользователь ожидает увидеть в результатах поиска профили ученых, и поисковая система вернет ему профили ученых. Оценка релевантности возвращенных результатов может оказаться сложнее. С одной стороны, респонденты Атокаралы упоминали, что одной из задач было нахождение профиля ученого, который приходил к ним на семинар или встретился на конференции. В этом случае можно быстро оценить результат, и объем можно считать небольшим.

С другой стороны, поиск рецензента для статьи может оказаться куда более сложной задачей (и в некоторых журналах авторов статей просят самим

предоставить список потенциальных рецензентов). Понять по доступной информации, сможет ли найденный кандидат справиться с написанием рецензии на статью, может оказаться вообще невозможно, придется отправлять текст нескольким кандидатам, чтобы они самостоятельно приняли решение. Процесс поиска также зависит от варианта задачи и для поиска рецензента вероятно будет итеративным, мультитактическим и оппортунистическим.

Таким образом, можно прийти к выводу, что этот последний вид поиска может быть как исследовательским, если мы не знаем, какого именно человека хотим найти (в случае поиска рецензента по не очень знакомой области), так и вполне традиционным (если мы ищем знакомого). Поэтому можно сделать вывод, что и наиболее трудоемкий вид поиска с целью исследования новых направлений, и чаще всего возникающий вид поиска с целью поддержки осведомленности являются по своей природе видами исследовательского поиска, и для их поддержки требуются специальные инструменты. Остальные результаты систематизированы в табл. 1. В следующем разделе мы рассмотрим, какие существуют инструменты для проведения исследовательского поиска.

4. Инструменты исследовательского поиска

По определению исследовательского поиска контекст проблемы характеризуется низкой специфичностью, высокой длительностью и большим объемом ответа, а процесс поиска является итеративным, оппортунистическим и мультитактическим. Такой поиск требует специальных инструментов, отличных от инструментов для традиционного поиска.

Авторы [2] предлагают восемь видов инструментов, характерных для систем исследовательского поиска:

- поддержка при составлении запроса и быстрое уточнение запроса;
- фасеты для фильтрации результатов по метаданным;
- использование контекста;
- интуитивно понятные визуализации, помогающие при принятии решений;
- поддержка обучения и понимания;
- поддержка совместной работы;
- истории поиска, рабочие места и отслеживание прогресса;
- поддержка управления задачами.

Многие из этих возможностей в том или ином виде уже внедрены в универсальные поисковые системы и не требуют представления, но многие не получили распространения несмотря на прошедшие со времени написания статьи годы. Мы кратко опишем каждую из возможностей, но за более подробным и исчерпывающим описанием отсылаем к первоисточнику [2].

Существуют и другие классификации инструментов, предназначенных для исследовательского поиска. Автор одного из самых свежих обзоров по исследовательскому поиску [11] особенно отмечает важность инструментов обзора больших объемов результатов поиска и рассматривает четыре основных

подхода: иерархическую классификацию, фасетную классификацию, динамическую кластеризацию и социальную классификацию (классификацию по социальным, экономическим, демографическим и другим критериям).

Сравнительно новым подходом к обработке больших коллекций результатов поиска является тематическое моделирование.

Кроме того, специально для анализа структуры научных областей разработан целый ряд специализированных инструментов, способных обрабатывать большие коллекции статей.

В ходе анализа всех этих инструментов мы разделили их на три группы, отражающие специфику контекста проблемы исследовательского поиска:

1. инструменты помощи при формулировании запроса.
2. инструменты помощи при анализе результатов.
3. инструменты поддержки длительного поиска.

Далее мы рассмотрим каждую из трех групп по очереди.

4.1. Инструменты помощи при формулировании запроса

Инструменты помощи при формулировании запроса помогают справиться с низкой специфичностью: так как пользователь не уверен в том, что именно он хочет найти, то он не уверен и в том, как сформулировать поисковый запрос. Системы исследовательского поиска должны помогать пользователю формулировать запрос. Суть быстрого уточнения запроса (rapid query refinement) [2] состоит в автоматическом расширении запроса за счет анализа близких запросов, сделанных ранее, возможно другими пользователями системы. Пользователь вводит (или начинает вводить) первое ключевое слово, а поисковая система предлагает ему несколько возможных продолжений запроса, из которых он может выбрать одно, после чего процесс может повторяться и для последующих слов запроса.

Интуиция, стоящая за этим методом, заключается в том, что разные люди могут искать одно и то же, и что, скорее всего, кто-то уже искал раньше то, что нужно пользователю сейчас. Таким образом, можно использовать историю близких поисковых запросов, чтобы вместо формулировки запроса с нуля, пользователь мог выбрать подходящий вариант из списка известных поисковой системе запросов.

Так как обычный запрос состоит всего из нескольких слов, зачастую многозначных, для повышения качества результатов необходимо дополнительно его расширять. Одним из способов такого расширения запроса является использование контекста. Контекст может быть получен или явно, за счет уточняющих вопросов пользователю, какие документы или фрагменты текста он считает релевантными, или неявно, за счет анализа пользовательского поведения.

Хотя большинство поисковых систем предполагают запрос в виде нескольких ключевых слов, возможны и более сложные запросы в виде фрагментов текста,

картинок, аудио файлов и т.д. Например, авторы [12] использовали в качестве запросов фрагменты текста объемом около страницы и затем осуществляли поиск релевантных документов в коллекции с помощью тематического моделирования.

К этой же группе мы отнесем рекомендации, которые можно рассматривать в качестве неявных запросов, когда пользователю не нужно использовать ни ключевые слова, ни какой-то другой вид входных данных, а вместо этого система использует в качестве запроса уже накопленную в ней информацию о пользователе. Например, может использоваться информация о поисковых запросах, сделанных ранее, об отобранных пользователем статьях, о статьях, написанных самим пользователем, и так далее. В одном из последних обзоров методов рекомендации статей рассматривается более двухсот работ [13].

После того, как запрос сформулирован и отправлен, необходимо отобразить его результаты. Причем, поскольку для исследовательского поиска высокая точность формулировки запроса маловероятна, даже релевантные запросу результаты поиска могут оказаться нерелевантными действительной поисковой потребности и наоборот. По этой причине нельзя рассчитывать на то, что пользователю окажется достаточно одного или нескольких релевантных запросу результатов. Скорее всего, их потребуется значительно больше, причем на этом этапе необходимо предоставить пользователю средства для упрощения восприятия больших объемов результатов. Для этого служат инструменты из второй группы.

4.2. Инструменты помощи при анализе результатов

Количество результатов поискового запроса может быть очень велико. Классические поисковые системы взвешивают все результаты по релевантности запросу и возвращают их в виде списка, отсортированного от более релевантных результатов к менее релевантным. Однако для исследовательского поиска на первых итерациях большая часть даже релевантных введенному запросу результатов может оказаться нерелевантной поисковым потребностям пользователя из-за недостаточно хорошо сформулированного запроса. Из-за этого нельзя больше рассчитывать, что удовлетворяющий пользователя результат будет найден на одной из первых страниц. Пользователь должен будет так или иначе проанализировать больший объем результатов, и простой сортированный список не очень для этого подходит.

Чтобы дать пользователю возможность приблизительно оценить структуру результатов поискового запроса и затем сфокусироваться на более релевантных частях, необходимо предоставлять пользователю средства обобщения, реферирования результатов поиска. Результат может быть рассеян по целой коллекции документов, и полнота поиска может оказаться значительно важнее точности. Хорошей иллюстрацией здесь может служить задача проведения систематического обзора: пропущенные релевантные результаты могут

привести к тому, что из обзора будут сделаны неверные выводы. Для решения этой задачи разработан целый ряд методов. Первый, который мы рассмотрим, – это фасеты.

Фасеты позволяют ограничивать результаты поиска с помощью указания допустимых значений атрибутов каждого результата. Каждый фасет – это четко определенный аспект, характеристика или свойство объекта или класса объектов. Значения фасета разбивают множество объектов на непересекающиеся подмножества, объединение которых равно этому множеству объектов. Как правило, фасеты создаются вручную, и для каждого конкретного домена они свои. Например, для домена научных публикаций примерами фасетов могут служить год публикации, место публикации (название журнала или конференции), имя одного из авторов.

Для решения задачи обобщения результатов поиска фасеты применяются естественным образом. Все результаты поиска группируются в зависимости от значений каждого фасета, и затем пользователь может сужать пространство поиска, отбирая те или иные значения фасетов. Применительно к информационному поиску предполагается, что каждый фасет плоский, его значения не образуют иерархии. Некоторым обобщением фасетов можно считать иерархическую классификацию.

Иерархическая классификация предполагает разбиение результатов поиска по фиксированному множеству вложенных и непересекающихся в рамках одного уровня вложенности классов (фактически, дереву). Примерами подобных иерархических классификаторов являются каталоги веб-страниц, широко применявшиеся до появления поисковых систем, или, например, классификатор научных областей ACM[2]. Один раз разобравшись со структурой классификатора, пользователь сможет быстро оценивать множество результатов поиска и отбирать из него только нужные ему подмножества. Основным недостатком этого подхода является сложность создания и поддержания в актуальном состоянии подобных классификаторов. Для домена научных данных в бурно развивающихся научных областях классификаторы всегда оказываются на шаг позади.

В тех случаях, когда применение классификатора по тем или иным причинам невозможно, применяют альтернативный инструмент: кластеризацию [11]. С помощью кластеризации можно автоматически выделить группы близких документов среди всех результатов поиска и позволить пользователю сэкономить усилия, просмотрев лишь по несколько документов из каждого кластера. Основной проблемой при реализации кластеризации является выбор хорошей функции расстояния между документами. Для перевода текстовых документов в векторное пространство могут применяться различные подходы [14-16].

[2] https://www.acm.org/about-acm/class

Еще один способ неформальной классификации, применимой для исследовательского поиска, – это фолксономия, или народная классификация [11]. При использовании этого инструмента каждый класс задается соответствующим тегом, а каждый документ может быть отнесен к одному или нескольким классам путем ручного указания соответствующих тегов. Теги может проставлять каждый пользователь, формируя таким образом общую картину.

Кроме группировки результатов поиска самой по себе, в системах исследовательского поиска применяются различные методы визуализации результатов [2, 11]. Наконец, с точки зрения Уайта и Рота [2] системы исследовательского поиска должны не просто возвращать набор релевантных документов, но и помогать понять и изучить их. Во внимание должны приниматься сложность каждого документа, их взаимосвязи (в каком порядке их надо изучать), уровень подготовки пользователя. К сожалению, практических примеров реализации этой функциональности нам обнаружить не удалось.

В следующей группе собраны инструменты, помогающие при проведении длительного поиска.

4.3. Инструменты поддержки длительного поиска

Исследовательский поиск в отличие от традиционного поиска итеративен и протяжен во времени. Пользователь будет время от времени прерывать поиск, продолжать с того же места или возвращаться к уже просмотренным ранее результатам и так далее. Системы исследовательского поиска должны предоставлять соответствующие средства поддержки.

Среди инструментов исследовательского поиска, выделяемых авторами [2], к инструментам поддержки длительного поиска можно отнести следующие три категории:

- групповой поиск (collaboration);
- поддержка историй, рабочих мест и отслеживания прогресса (histories, workspaces, and progress updates);
- поддержка управления задачами (task management).

Системы исследовательского поиска должны сохранять историю поиска и позволять быстро возвращаться к полученным ранее результатам. Пользователи должны иметь возможность сохранять заметки по ходу поиска или необходимую информацию более сложной структуры. Кроме того, система должна отслеживать, какие пути уже исследованы, а какие нет, отслеживать скорость продвижения и предоставлять соответствующие данные пользователю. Сложный поиск должно быть можно выполнять не в одиночку, а группой. Это может оказаться полезным, потому что разные пользователи обладают разным опытом и могут дополнять друг друга. При наличии подходящего поискового интерфейса пользователи могут перенимать друг у

друга как знания в изучаемой предметной области, так и знания о том, как искать.

Одна из классических универсальных систем для совместного поиска – SearchTogether [17]. SearchTogether предоставляла интегрированный интерфейс с функциями взаимного информирования о действиях всех членов группы, распределения работы и сохранения промежуточных результатов поиска.

Шесть лет спустя авторы [18] приходят к выводу, что большинство пользователей не использует никаких специальных средств для совместного поиска. Вместо этого они применяют универсальные инструменты: электронную почту, мессенджеры и социальные сети. Тем не менее, мы считаем, что для профессиональных пользователей поисковых систем необходимость изучения специализированного интерфейса не станет непреодолимым препятствием.

На этом мы заканчиваем краткий обзор инструментов исследовательского поиска. Мы описали три группы инструментов: инструменты помощи при формулировании запроса, инструменты помощи при анализе результатов и инструменты поддержки длительного поиска. В следующем разделе мы рассмотрим, какое применение они нашли в системах поиска научных публикаций.

5. Существующие подходы к решению задачи исследовательского поиска публикаций

Научные публикации представляют собой достаточно специфичный вид данных, что с одной стороны усложняет построение поисковых систем, а с другой стороны предоставляет дополнительные возможности повышения качества их работы по сравнению с универсальными поисковыми системами. Кратко рассмотрим основные специфические особенности научных публикаций.

Во-первых, специфичен текст статей: он написан в научном стиле, не окрашен эмоционально, включает большое количество узкоспециальных терминов, во многих областях науки – формул (например, химических, математических). Текст статей обладает достаточно устойчивой структурой: в большинстве статей есть введение, обзор существующих работ, заключение. Во многих областях науки фактическим стандартом является наличие отдельного раздела с указанием методов, использованных при проведении исследования.

Во-вторых, текст статей существует не сам по себе, но вместе с метаданными. К метаданным относится информация об авторах статьи: их имена и аффилиации, адреса электронной почты; место публикации статьи – название журнала или конференции; аннотация – она обычно свободно доступна, даже если доступ к тексту статьи ограничен; год публикации и другие.

В-третьих, статьи со связями цитирования образуют ориентированный направленный граф без циклов – граф цитирования. Кроме собственно графа

цитирования по статьям, также рассматриваются и производные от него: граф цитирования по авторам, графы социтирования по документам и по авторам [19].

Таким образом, хотя анализ непосредственно текстов научных статей и затруднен, за счет использования дополнительной информации – метаданных и графа цитирования, возможно повышение качества работы поисковых систем.

За прошедшие годы появилось немало исследовательских систем, реализующих те или иные средства исследовательского поиска для научных публикаций. Рассмотреть все их в рамках статьи не представляется возможным.

Авторы [20] в своем исследовании использования различных интерфейсов поиска рассмотрели две системы: ACM DL и Sowiport DL. С одной стороны, нам не известен исчерпывающий перечень таких систем, и даже известных систем слишком много, чтобы охватить их все в рамках одной статьи. С другой стороны, многие системы очень похожи между собой по отношению к предоставляемым функциям, и достаточно знать одну из них, чтобы понимать, чего ожидать от остальных.

Мы выделили несколько групп систем и выбрали наиболее известных, с нашей точки зрения, представителей в каждой из них:

- университетские разработки: CiteseerX [21], aMiner [22];
- общедоступные системы от крупных корпораций: Google Scholar и Microsoft Academic Search;
- SemanticScholar от института Аллена [23];
- социальные сети для ученых: Academia.edu, BibSonomy, Mendeley [24], ResearchGate;
- сайты электронных библиотек: ACM DL, arXiv.org, eLibrary, PubMed, ScienceDirect, Scopus, Springerlink, Web Of Science.

Далее в этом разделе мы покажем, какие средства поддержки исследовательского поиска используются в поисковых системах для научных публикаций и каким образом. Мы будем придерживаться разделения инструментов на три группы, введенного в предыдущем разделе. Сначала рассмотрим инструменты помощи при формулировании запроса, затем инструменты анализа результатов и, наконец, инструменты поддержки длительного поиска.

5.1. Инструменты помощи при формулировании запроса в современных системах поиска результатов научных исследований

Специфическая терминология, используемая в научных статьях, сильно затрудняет формулировку поискового запроса для пользователя, не знакомого с предметной областью. Как отмечали респонденты [10], при исследовании новой области им часто приходится сначала использовать универсальные источники данных, такие как Google и Wikipedia, и только потом, после

поверхностного ознакомления с предметной областью, они могут перейти к использованию специализированных поисковых систем. По этой причине особенно важно помогать пользователю на начальных этапах поиска, либо помогая ему сформулировать явный поисковый запрос, либо предполагая запрос неявно, за счет анализа истории поиска и уже отобранных пользователем релевантных статей. В этой группе инструментов мы рассмотрим средства быстрого уточнения запроса, использования контекста и рекомендаций. Сводные результаты можно увидеть в табл. 2.

С помощью быстрого уточнения запроса пользователь может выбирать из потенциально релевантных альтернатив, а не формулировать весь поисковый запрос целиком. Как ни странно, в большинстве рассмотренных систем никак не реализуется быстрое уточнение запроса, и в этом отношении они уступают универсальным поисковым движкам. В табл. 2 этот вариант обозначен словом *нет*. Среди оставшихся можно выделить две группы: в системах из первой группы реализуется классический вариант уточнения, основанный на агрегации запросов пользователей (обозначен в табл. 2 словом *классика*), системы из второй группы идут дальше и реализуют более доменоспецифичные подсказки (*адаптация к домену*). Например, MS Academic умеет распознавать в запросе названия журналов, а Academia.edu подсказывает так называемые исследовательские интересы (research interests).

Табл. 2. Поддержка инструментов помощи при формулировании запроса в современных системах поиска научных публикаций
Table 2. Support of tools to assist in the formulation of a query in modern systems of the search for scientific publications

Система	Быстрое уточнение запроса	Рекомендации
aMiner	адаптация к домену	для статьи
CiteseerX	нет	для статьи
Google Scholar	классика	для пользователя
MS Academic	адаптация к домену	для статьи
Semantic Scholar	классика, также есть возможность уточнить запрос со страницы с результатами	для статьи
Academia.edu	адаптация к домену	нет
BibSonomy	нет	нет
Mendeley	нет	для пользователя
ResearchGate	нет	для пользователя
ACM DL	нет	нет
arXiv.org	нет	нет
eLibrary	нет	нет
PubMed	классика	для статьи

ScienceDirect	нет	для пользователя*
Scopus	нет	нет
Springerlink	классика	нет
WOS	нет	нет

Современные универсальные поисковые движки автоматически отслеживают историю поисковых запросов пользователя и затем используют ее для ранжирования результатов (а также показа контекстной рекламы). Однако в специализированных поисковых системах для ученых аналогичную функциональности нам обнаружить не удалось. Мы делали по несколько запросов в режиме инкогнито, в обычном режиме и после авторизации на сайте (где это возможно), но не смогли обнаружить разницы в результатах, показанных на первых трех страницах поисковой выдачи. Таким образом, в этом аспекте все системы одинаковы, и мы не стали включать соответствующий столбец в сводную таблицу.

Если понимать под рекомендацией сценарий работы с системой, не предполагающий явного ввода запроса пользователем, то только в трех системах реализуется нечто подобное: Google Scholar, Mendeley и ResearchGate (значение *для пользователя* в сводной таблице). В ScienceDirect явным образом заявляется поддержка рекомендаций, но нам не удалось ими воспользоваться. Каждая из них так или иначе отслеживает активность пользователя и автоматически рекомендует статьи, релевантные его интересам. В некоторых системах реализована близкая функциональность – поиск связанных (related) статей (значение *для статьи* в сводной таблице). Остальные системы рекомендации не поддерживают. Следующий подраздел посвящен инструментам помощи при анализе результатов поиска.

5.2. Инструменты помощи при анализе результатов в современных системах поиска результатов научных исследований

В рассмотренных нами системах так или иначе реализуются четыре группы инструментов помощи при анализе результатов: фасеты, иерархическая классификация, фолксономия и визуализация результатов.

Табл. 3. Поддержка инструментов помощи при анализе результатов в современных системах поиска результатов научных исследований
Table 3. Support of assistance tools in analyzing the results in modern systems for search of research results

Система	Фасеты	Иерархическая классификация	Фолксономия	Визуализация
aMiner	расширенные	нет	есть	сложная
CiteseerX	нет	нет	нет	нет
Google Scholar	год	универсальный	нет*	простая*

MS Academic	расширенные	универсальный	нет	нет
Semantic Scholar	расширенные	нет	нет	простая
BibSonomy	нет	нет	есть	нет
Mendeley	нет	нет	нет	нет
ResearchGate	нет	нет	нет	нет
ACM DL	расширенные	доменный	нет	простая
arXiv.org	нет	универсальный	нет	нет
eLibrary	расширенные	универсальный	нет	нет
PubMed	расширенные	доменный	нет	нет
ScienceDirect	расширенные	универсальный	нет	нет
Scopus	расширенные	универсальный	нет	сложная
Springerlink	расширенные	универсальный	нет	сложная
WOS	расширенные	универсальный	нет	сложная

Academia.edu полноценный полнотекстовый поиск предоставляет только в платной версии, поэтому в этом подразделе мы ее не рассматривали.

Фасеты реализованы в большинстве систем, причем если базовые параметры, такие как год или место публикации, поддерживаются всеми, то более сложные параметры достаточно своеобразны. Например, Semantic Scholar умеет автоматически выявлять набор данных, использованный в статье для тестирования, MS Academic позволяет фильтровать результаты по аффилиациям авторов, PubMed поддерживает разделение по типу живых существ, исследованных в статье, и так далее. AMiner в своей подсистеме поиска экспертов предоставляет фасеты по индексу Хирша, полу и языку. eLibrary поддерживает фасеты для просмотра сохраненных коллекций, но не для результатов поиска.

Необходимо также отметить, что в большинстве систем поддерживается расширенный поиск, позволяющий сформулировать сложный запрос, учитывающий разные параметры результатов. В сводной таблице мы использовали только три значения: *нет*, если фасеты не поддерживаются, *год*, если доступен только фасет по году публикации, *расширенные*, если в системе доступен более широкий набор фасетов.

Иерархические классификаторы в системах научного поиска используются только для структурирования направлений исследований. Картина здесь достаточно пестрая, и сравнивать разные системы между собой трудно. Можно разделить системы на адаптированные к конкретному домену, такие как PubMed или ACM DL, и универсальные, такие как Google Scholar или eLibrary. Адаптированные к домену системы, как правило, предоставляют более богатые классификаторы, а универсальные – менее богатые. С другой стороны, в некоторых системах используются классификаторы, сделанные вручную (как например MeSH в PubMed), а некоторые строят классификатор автоматически,

как, например, MS Academic. Кроме того, польза от использования классификатора зависит от того, каким образом к нему привязываются статьи. К сожалению, ответы на эти вопросы как правило недокументированы, поэтому в этом случае мы ограничились классификацией на три класса: *нет* классификатора, *доменный* классификатор, *универсальный* классификатор.

Фолксономия существующими системами поддерживается слабо, поддержка есть только в aMiner и специальных системах, таких как Bibsonomy. В Google Scholar можно пользоваться приватными тегами (что, строго говоря, нельзя называть фолксономией). Как ни странно, поддержки тегов нет ни в одной из систем, которые можно считать социальными сетями. Ни в Academia.edu, ни в Mendeley, ни в ResearchGate поддержки фолксономии нам обнаружить не удалось.

В рассмотренных системах, как правило, применяются достаточно простые виды визуализаций или не применяется никаких. Наиболее частый вариант – столбчатая диаграмма по годам. Встречаются и более сложные варианты: столбчатые диаграммы по каждому значению фасета (Springerlink), диаграммы влияния авторов (Semantic Scholar). Microsoft Academic Search и CiteseerX полностью пренебрегают визуализацией (хотя Microsoft Academic Search на странице часто задаваемых вопросов обещает вернуть визуализацию графа цитирования из предыдущей версии системы). Semantic Scholar предоставляет гистограммы по годам для результатов поиска статей и для цитирований статей и авторов на соответствующих страницах. aMiner на странице автора предоставляет диаграмму изменения интересов автора по годам.

Куда более мощные средства визуализации больших объемов статей, которые могут быть получены в результате поиска по ключевым словам, предоставляют средства, разрабатываемые в области картографирования науки, такие как Citespace, Sci2Sci, SciMat. Эти системы берут на вход данные о статьях и их цитированиях и позволяют анализировать заданную таким образом предметную область. Один из основных сценариев работы с такими системами заключается в выполнении поиска, например, в Scopus, экспорта результатов и затем их анализа. К сожалению, автоматически получить существенное количество результатов поискового запроса в большинстве поисковых систем невозможно и далеко не все из них предоставляют информацию о цитированиях. Кроме того, при таком сценарии использования отсутствует возможность интерактивного взаимодействия с поисковой системой. После уточнения запроса необходимо снова вручную выполнять экспорт результатов поиска и заново начинать анализ. С другой стороны, средства анализа набора статей, реализованные в системах картографирования науки, существенно богаче всего, что реализовано в рассмотренных поисковых системах. Поддерживается и графовая кластеризация, и автоматическая генерация меток для них, и различные способы визуализации результатов анализа.

Подводя итог этому подразделу отметим, что основным механизмом описания результатов поиска являются фасеты, дополненные иерархическими

классификаторами областей исследования. Более мощные средства предоставляются оффлайновыми инструментами картографирования науки, которые поддерживают и кластеризацию (в том числе с автоматической генерацией меток), и графическую визуализацию всех результатов с учетом автоматически вычисляемой важности отдельный статей. Однако оффлайновые системы не позволяют переформулировать поисковый запрос и не имеют доступа к полной базе статей, вынуждая пользователей постоянно проходить цикл поиска, экспорта данных, автоматического анализа и снова поиска с уточненными параметрами, переключаясь между двумя системами.

5.3. Инструменты поддержки длительного поиска в современных системах поиска результатов научных исследований

Рассматривая инструменты поддержки длительного поиска, реализованные в исследуемых системах, мы выделили четыре группы: поддержка совместной работы, подписка на обновления, история поиска и поддержка рабочих мест.

Поддержка совместного проведения исследований в рассмотренных системах реализована слабо. Только в трех системах удалось обнаружить функции, предназначенные для работы в команде. BibSonomy позволяет создавать группы. Каждый член группы, взаимодействуя с системой – создавая новый пост, или делясь ссылкой, может ограничить видимость своего действия конкретной группой. ResearchGate позволяет создавать так называемые проекты, в которые можно добавить нескольких участников, оставить список релевантной литературы и список вопросов, на которые нужно получить ответы. Кроме того, в рамках проекта можно вести ленту обновлений. На проект можно подписаться и получать уведомления об изменениях. Mendeley тоже позволяет создавать группы, похожие по своей функциональности на проекты в ResearchGate. В табл. 4 мы использовали два значения: *общие проекты* для трех описанных выше систем и *нет* для всех остальных. Ни одна система не поддерживает собственно одновременный групповой поиск в духе SearchTogether.

Следующая группа инструментов для поддержки длительного поиска – это подписки. Основной вариант использования для подписок – поддержание осведомленности о состоянии предметной области. Рассмотренные системы поиска для ученых поддерживают следующие виды подписок в разных сочетаниях:

1. подписки на новые статьи выбранных авторов (*авторы*);
2. подписки на новые цитирования выбранных статей (*цитирования*);
3. подписки на новые результаты поисковых запросов (*запросы*).

Есть некоторые индивидуальные особенности: Semantic Scholar не позволяет подписываться на результаты поиска, но позволяет подписаться на обновления одной из поддерживаемых областей исследований, Google Scholar позволяет подписаться на цитирования любой статьи выбранного автора. Если достаточно

тщательно настроить все виды подписок одновременно, то можно быть почти уверенным, что никакие новые статьи по интересующим вас темам не будут упущены. Для известных авторов можно подписаться на все их новые статьи. Чтобы получать статьи не только известных авторов, можно подписаться на все цитирования важных в области статей. Также можно подписаться на все цитирования собственных статей, чтобы отслеживать обратную связь на свои исследования. Наконец, можно подписаться на обновления результатов поисковых запросов, чтобы иметь возможность заметить новые исследования в смежных областях, которые могут не процитировать важные статьи, на которые вы уже подписаны.

Следующий инструмент, который мы рассмотрим, более важен для исследования новой области. Это истории поисковых запросов.

Процесс изучения новой предметной области можно сравнить с поиском в пространстве состояний. Пользователь начинает с некоторых поисковых запросов, просматривает результаты поиска на некоторую глубину, делает уточняющие запросы, когда хочет исследовать некоторую подзадачу полнее, и возвращается назад, когда хочет вернуться к основной задаче. История поисковых запросов должна помогать в этом процессе, позволять быстро переключаться между анализом результатов разных поисковых запросов. Многие поисковые системы позволяют создавать сложные запросы с большим количеством ограничений по разным параметрам статей. Создание таких запросов может быть достаточно трудоемким и важно, чтобы, однажды составив их, пользователь мог легко к ним вернуться.

Табл. 4. Инструменты поддержки длительного поиска в современных системах поиска результатов научных исследований
Tab. 4. Tools to support long search in modern systems for search of research results

Система	Совместная работа	Подписки	Истории	Рабочие места
aMiner	нет	авторы	автоматически	теги
CiteseerX	нет	цитирования	нет	несколько папок
Google Scholar	нет	авторы, запросы, цитирования	вручную	теги
MS Academic	нет	авторы, цитирования	нет	одна папка
Semantic Scholar	нет	авторы, цитирования	нет	одна папка
Academia.edu	нет	авторы	автоматически	одна папка
BibSonomy	общие проекты	нет	нет	теги
Mendeley	общие проекты	авторы	нет	несколько папок
ResearchGate	общие проекты	авторы, цитирования	нет	несколько папок
ACM DL	нет	нет	вручную	несколько папок

arXiv.org	нет	нет	нет	нет
eLibrary	нет	нет	автоматически	несколько папок
PubMed	нет	авторы, запросы	автоматически	несколько папок
ScienceDirect	нет	нет	нет	нет
Scopus	нет	авторы, запросы, цитирования	вручную	несколько папок
Springerlink	нет	запросы	нет	нет
WOS	нет	авторы, запросы, цитирования	вручную	несколько папок

Частично сохранение истории поисковых запросов может быть реализовано средствами браузера, возможно поэтому около половины рассмотренных систем никак не реализуют эту функциональность. Тем не менее, возможностей браузера может быть недостаточно (как минимум, в истории браузера не сохранятся POST-запросы) и многие системы позволяют сохранять историю. Мы выделили два варианта поддержки истории:

1. ручной режим, когда пользователь должен самостоятельно сообщить системе, что конкретный поисковый запрос должен быть сохранен;
2. автоматический режим, когда система сама запоминает все запросы пользователя и позволяет ему просматривать эту историю и возвращаться к ранее сделанным запросам.

К сожалению, ни в одной из систем нам не удалось обнаружить возможности вернуться к конкретной странице результатов поиска, чтобы продолжить с того места, где работа была прервана. Таким образом, в этой части остается очевидное пространство для дальнейших улучшений.

Наконец, последний вид инструментов поддержки длительного поиска – это поддержка рабочих мест.

Ядром рабочего места пользователя в поисковых системах для научных публикаций является отобранный им список литературы, т.е. промежуточный результат выполняемого им длительного поиска. Поддержка рабочего места реализована не везде, а где есть, то обеспечивается в достаточно примитивном виде: это либо единый список статей, либо с разбивкой по папкам, либо с разбивкой по тегам. Единый список статей может быть полезен, только если ученый работает ровно в одной, достаточно узкой области или, например, использует систему для написания одной статьи. Когда количество отобранных статей начинает приближаться к сотне, единый список не может оставаться удобным и требуется какая-то возможность его рубрикации.

Очевидный способ – разделение списка на части аналогично разделению файлов по папкам. Этот способ организации наиболее популярен и поддерживается в половине рассмотренных систем. Несколько папок для отобранных статей удобнее, чем один плоский список, но остается

существенное ограничение: каждая статья может быть сохранена только в одной папке, что часто неудобно.

Наиболее гибкий способ организации хранения отобранных статей – это тэги. Каждую статью можно пометить произвольным набором тегов и затем легко отбирать подмножества статей с заданными тэгами, в некотором смысле динамически формируя виртуальные папки. Google Scholar поддерживает поиск по статьям из библиотеки, что может быть очень удобно для поиска с целью обучения, когда надо найти уже встречавшуюся ранее статью.

Подытоживая этот подраздел, мы вынуждены констатировать, что хотя многие рассмотренные системы достаточно хорошо поддерживают поиск с целью поддержки осведомленности (в первую очередь за счет подписок), исследование новой предметной области поддерживается значительно хуже. Только в социальных сетях для ученых есть некоторые достаточно ограниченные средства организации командной работы, остальные поисковые системы не поддерживают ее вовсе. Поддержка рабочих мест также оставляет желать лучшего, следует отметить разве что Google Scholar с его возможностями организации статей с помощью тегов и локальному поиску среди отобранных статей.

6. Выводы и дискуссия

Мы предприняли попытку объединить результаты исследований поискового поведения ученых с наработками, полученными при изучении исследовательского поиска. Исторически исследовательский поиск определялся, отталкиваясь от традиционного поиска и в противовес ему. Определение исследовательского поиска достаточно размыто, и многие инструменты, относимые к исследовательскому поиску, сегодня внедрены в традиционные поисковые системы.

Мы придерживались определения исследовательского поиска, данного в работе Уайта и Рота [2]: исследовательский поиск – это поиск, контекст проблемы которого является открытым, постоянным и многоаспектным, а процесс поиска является оппортунистическим, итеративным и мультитактическим. Опираясь на это определение, мы проанализировали основные задачи поиска, возникающие перед учеными, и обнаружили, что большая их часть может быть уверено отнесена к исследовательскому поиску.

Поисковое поведение ученых исследуется не первое десятилетие. Мы кратко проследили эти исследования, начиная от работ Эллиса и др. [4-5] и заканчивая достаточно современной работой Атукоралы и др. [10], в которой исследовалось поисковое поведение ученых в области компьютерных наук и то, как они используют современные электронные поисковые системы.

Обнаружив, что поисковое поведение ученых в большинстве случаев относится к исследовательскому поиску, мы выяснили, какие существуют инструменты,

предназначенные для исследовательского поиска, и выделили среди них три группы:

1. инструменты помощи при формулировании запроса;
2. инструменты помощи при анализе результатов;
3. инструменты поддержки длительного поиска.

Для каждой из групп мы кратко описали относящиеся к ней инструменты.

Наконец, мы рассмотрели ряд специализированных поисковых систем для ученых, чтобы выяснить, насколько широко в них применяются инструменты исследовательского поиска и какие именно. Мы проанализировали адаптации универсальных поисковых систем к домену научных публикаций: Google Scholar и Microsoft Academic Search; экспериментальные академические разработки: aMiner, CiteseerX, SemanticScholar; поисковые системы, интегрированные в электронные библиотеки: ACMDL, arXiv.org, eLibrary, ScienceDirect, Scopus, SpringerLink, WOS; социальные сети для ученых: Academia.edu, BibSonomy, Mendeley, ResearchGate. Основная часть анализа поисковых систем была проведена в конце 2017 – начале 2018 года и затем актуализирована в конце 2018 года.

Из проведенного анализа можно сделать несколько выводов. И наиболее частые, и наиболее трудоемкие поисковые потребности ученых связаны с исследовательским поиском. При этом специализированные поисковые системы для ученых поддерживают проведение исследовательского поиска лишь частично.

Инструменты помощи при составлении запроса почти нигде не продвинулись дальше автодополнения, а во многих системах нет даже и этого. Таким образом, порог входа остается высоким, и неспециалисту придется потратить существенное время на то, чтобы подобрать запрос, соответствующий его потребностям.

Среди инструментов обзора результатов поиска наиболее развиты фасеты, с помощью которых опытный ученый сможет эффективно отобрать интересующие его статьи. Однако инструменты визуализации сравнительно примитивны и сильно уступают реализованным в инструментах картографирования науки. Насколько можно судить, не реализованы ни методы кластеризации, ни анализ графа цитирования, ни построение текстовых описаний.

Методы поддержки длительного поиска в лучшем случае сводятся к возможности откладывать релевантные статьи в именованные коллекции, аналогичные папкам на файловой системе. Кроме того, можно подписываться на новые статьи избранных авторов, цитирования избранных статей и обновления результатов заранее подобранных поисковых запросов. Этих инструментов в целом достаточно для реализации сценария поиска с целью поддержки осведомленности, но для групповой работы над одной задачей придется использовать общий аккаунт и внешние средства коммуникации.

Таким образом, наиболее перспективными направлениями для дальнейшей работы выглядят исследования по следующим направлениям:

1. понижение порога входа для начинающих исследователей: реализация специфических видов запросов, прежде всего рекомендаций;
2. интеграция методов картографирования науки для обзора результатов поиска;
3. интеграция методов поддержки совместного поиска.

Список литературы

[1]. G. Marchionini. Exploratory search: from finding to understanding. Communications of the ACM, vol. 49, no. 4, 2006, pp. 41–46.

[2]. R.W. White and R.A. Roth. Exploratory search: Beyond the query-response paradigm. Synthesis lectures on information concepts, retrieval, and services, Morgan and Claypool Publishers, 2009, 98 p.

[3]. G. Marchionini. Information seeking in electronic environments. Cambridge University Press, 1997, 240 p.

[4]. D. Ellis, D. Cox, and K. Hall. A comparison of the information seeking patterns of researchers in the physical and social sciences. Journal of documentation, vol. 49, no. 4, 1993, pp. 356–369.

[5]. D. Ellis. A behavioural approach to information retrieval system design. Journal of documentation, vol. 45, no. 3, 1989, pp. 171–212.

[6]. X. Niu, B. M. Hemminger, C. Lown, S. Adams, C. Brown, A. Level, M. McLure, A. Powers, M. R. Tennant, and T. Cataldo. National study of information seeking behavior of academic researchers in the United States. Journal of the Association for Information Science and Technology, vol. 61, no. 5, 2010, pp. 869–890.

[7]. X. Niu and B. M. Hemminger. A study of factors that affect the information-seeking behavior of academics cientists. Journal of the American Society for Information Science and Technology, vol. 63, no. 2, 2012, pp. 336–353.

[8]. L.I. Meho and H.R. Tibbo. Modeling the information-seeking behavior of social scientists: Ellis's study revisited. Journal of the Association for Information Science and Technology, vol. 54, no. 6, 2003, pp. 570–587.

[9]. E. Palagi, F. Gandon, A. Giboin, and R. Troncy. A survey of definitions and models of exploratory search. In Proc. of the 2017 ACM Workshop on Exploratory Search and Interactive Data Analytics, 2017, pp. 3–8.

[10]. K. Athukorala, E. Hoggan, A. Lehtio, T. Ruotsalo, and G. Jacucci. Information-seeking behaviors of computer scientists: Challenges for electronic literature search tools. In Proc. of the Proceedings of the 76th ASIS&T Annual Meeting of Association for Information Science and Technology, vol. 50, 2013, pp. 1–11.

[11]. T. Jiang. Exploratory search: a critical analysis of the theoretical foundations, system features, and research trends. In Library and Information Sciences, Springer, 2014, pp. 79–103.

[12]. А.О. Янина, К.В. Воронцов. Мультимодальные тематические модели для разведочного поиска в коллективном блоге. Машинное обучение и анализ данных, том 2, вып. 2, 2016, стр. 173-186.

[13]. J. Beel, B. Gipp, S. Langer, and C. Breitinger. Research-paper recommender systems: a literature survey. International Journal on Digital Libraries, vol. 17, no. 4, 2016, pp. 305–338.

[14]. П.А. Пархоменко, А.А. Григорьев, Н.А. Астраханцев. Обзор и экспериментальное сравнение методов кластеризации текстов. Труды ИСП РАН, том. 29, вып. 2, 2017, стр. 161–200. DOI: 10.15514/ISPRAS-2017-29(2)-6.
[15]. H. Tian and H. H. Zhuo. Paper2vec: Citation-context based document distributed representation for scholar recommendation. arXiv preprint arXiv:1703.06587, 2017.
[16]. X. Kong, M. Mao, W. Wang, J. Liu, and B. Xu. Voprec: Vector representation learning of papers with text information and structural identity for recommendation. IEEE Transactions on Emerging Topics in Computing, 2018.
[17]. M.R. Morris and E. Horvitz. Searchtogether: an interface for collaborative web search. In Proc. of the 20th Annual ACM symposium on User interface software and technology, 2007, pp. 3–12.
[18]. M.R. Morris. Collaborative search revisited. In Proc. of the 2013 Conference on Computer supported cooperative work. ACM, 2013, pp. 1181–1192.
[19]. C. Chen, F. Ibekwe-SanJuan, and J. Hou. The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. Journal of the American Society for information Science and Technology, vol. 61, no. 7, 2010, pp. 1386–1409.
[20]. L. McCay-Peet, A. Quan-Haase, and D. Kern. Exploratory search in digital libraries: a preliminary examination of the use and role of interface features. In Proc. of the 78th Annual Meeting of Association for Information Science and Technology, vol. 52, 2015, pp. 1–4.
[21]. H. Li, I. Councill, W.-C. Lee, and C.L. Giles. Citeseerx: an architecture and web service design for an academic document search engine. In Proc. of the 15th International Conference on World Wide Web, 2006, pp. 883–884.
[22]. J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In Proc. of the 14th ACM SIGKDD International conference on Knowledge discovery and data mining, 2008, pp. 990–998.
[23]. W. Ammar, D. Groeneveld, C. Bhagavatula et al. Construction of the literature graph in semantic scholar. arXiv preprint arXiv:1805.02262, 2018.
[24]. H. Zaugg, R.E. West, I. Tateishi, and D.L. Randall. Mendeley: Creating communities of scholarly inquiry through research collaboration. TechTrends, vol. 55, no. 1, 2011, pp. 32–36.

Exploratory search for scientific articles

[1]*Y.R. Nedumov <yaroslav.nedumov@ispras.ru>*
[1,2,3,4]*S.D. Kuznetsov <kuzloc@ispras.ru>*
[1] *Ivannikov Institute for System Programming of the Russian Academy of Sciences, 25, Alexander Solzhenitsyn st., Moscow, 109004, Russia*
[2] *Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russia*
[3] *Moscow Institute of Physics and Technology (State University) 9 Institutskiy per., Dolgoprudny, Moscow Region, 141700, Russia*
[4] *National Research University Higher School of Economics (HSE) 11 Myasnitskaya Ulitsa, Moscow, 101000, Russia*

Abstract. It is intuitively clear that the search for scientific publications often has many characteristics of a research search. The purpose of this paper is to formalize this intuitive understanding, explore which research tasks of scientists can be attributed to research search,

what approaches exist to solve a research search problem in general, and how they are implemented in specialized search engines for scientists. We researched existing works regarding information seeking behavior of scientists and the special variant of a search called exploratory search. There are several types of search typical for scientists, and we showed that most of them are exploratory. Exploratory search is different to information retrieval and demands special support from search systems. We analyzed seventeen actual search systems for academicians (from Google Scholar, Scopus and Web of Science to ResearchGate) from the exploratory search support aspect. We found that most of them didn't go far from simple information retrieval and there is a room for further improvements especially in the collaborative search support.

Keywords: exploratory search; academic search engines; information seeking behaviour

DOI: -10.15514/ISPRAS-2018-30(6)-10

For citation: Nedumov Y.R., Kuznetsov S.D. Exploratory search for scientific articles. Trudy ISP RAN/Proc. ISP RAS, vol. 30, issue 6, 2018, pp. 171-198 (in Russian). DOI: 10.15514/ISPRAS-2018-30(6)-10

References

[1]. G. Marchionini. Exploratory search: from finding to understanding. Communications of the ACM, vol. 49, no. 4, 2006, pp. 41–46.
[2]. R.W. White and R.A. Roth. Exploratory search: Beyond the query-response paradigm. Synthesis lectures on information concepts, retrieval, and services, Morgan and Claypool Publishers, 2009, 98 p.
[3]. G. Marchionini. Information seeking in electronic environments. Cambridge University Press, 1997, 240 p.
[4]. D. Ellis, D. Cox, and K. Hall. A comparison of the information seeking patterns of researchers in the physical and social sciences. Journal of documentation, vol. 49, no. 4, 1993, pp. 356–369.
[5]. D. Ellis. A behavioural approach to information retrieval system design. Journal of documentation, vol. 45, no. 3, 1989, pp. 171–212.
[6]. X. Niu, B. M. Hemminger, C. Lown, S. Adams, C. Brown, A. Level, M. McLure, A. Powers, M. R. Tennant, and T. Cataldo. National study of information seeking behavior of academic researchers in the United States. Journal of the Association for Information Science and Technology, vol. 61, no. 5, 2010, pp. 869–890.
[7]. X. Niu and B. M. Hemminger. A study of factors that affect the information-seeking behavior of academics cientists. Journal of the American Society for Information Science and Technology, vol. 63, no. 2, 2012, pp. 336–353.
[8]. L.I. Meho and H.R. Tibbo. Modeling the information-seeking behavior of social scientists: Ellis's study revisited. Journal of the Association for Information Science and Technology, vol. 54, no. 6, 2003, pp. 570–587.
[9]. E. Palagi, F. Gandon, A. Giboin, and R. Troncy. A survey of definitions and models of exploratory search. In Proc. of the 2017 ACM Workshop on Exploratory Search and Interactive Data Analytics, 2017, pp. 3–8.
[10]. K. Athukorala, E. Hoggan, A. Lehtio, T. Ruotsalo, and G. Jacucci. Information-seeking behaviors of computer scientists: Challenges for electronic literature search tools. In Proc. of the Proceedings of the 76th ASIS&T Annual Meeting of Association for Information Science and Technology, vol. 50, 2013, pp. 1–11.

[11]. T. Jiang. Exploratory search: a critical analysis of the theoretical foundations, system features, and research trends. In Library and Information Sciences, Springer, 2014, pp. 79–103.
[12]. A. Yanina, K. Vorontsov. Multimodal topic modeling for exploratory search in collective blog. In Journal of Machine Learning and Data Analysis, vol. 2, no. 2, 2016, pp. 173-186 (in Russian).
[13]. J. Beel, B. Gipp, S. Langer, and C. Breitinger. Research-paper recommender systems: a literature survey. International Journal on Digital Libraries, vol. 17, no. 4, 2016, pp. 305–338.
[14]. P.A. Parhomenko, A.A. Grigorev, N.A. Astrakhantsev A survey and an experimental comparison of methods for text clustering: application to scientific articles. Trudy ISP RAN/Proc. ISP RAS, 2017, vol. 29, issue 2, pp. 161-200 (in Russian). DOI: 10.15514/ISPRAS-2017-29(2)-6
[15]. H. Tian and H. H. Zhuo. Paper2vec: Citation-context based document distributed representation for scholar recommendation. arXiv preprint arXiv:1703.06587, 2017.
[16]. X. Kong, M. Mao, W. Wang, J. Liu, and B. Xu. Voprec: Vector representation learning of papers with text information and structural identity for recommendation. IEEE Transactions on Emerging Topics in Computing, 2018.
[17]. M.R. Morris and E. Horvitz. Searchtogether: an interface for collaborative web search. In Proc. of the 20th Annual ACM symposium on User interface software and technology, 2007, pp. 3–12.
[18]. M.R. Morris. Collaborative search revisited. In Proc. of the 2013 Conference on Computer supported cooperative work. ACM, 2013, pp. 1181–1192.
[19]. C. Chen, F. Ibekwe-SanJuan, and J. Hou. The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. Journal of the American Society for information Science and Technology, vol. 61, no. 7, 2010, pp. 1386–1409.
[20]. L. McCay-Peet, A. Quan-Haase, and D. Kern. Exploratory search in digital libraries: a preliminary examination of the use and role of interface features. In Proc. of the 78th Annual Meeting of Association for Information Science and Technology, vol. 52, 2015, pp. 1–4.
[21]. H. Li, I. Councill, W.-C. Lee, and C.L. Giles. Citeseerx: an architecture and web service design for an academic document search engine. In Proc. of the 15th International Conference on World Wide Web, 2006, pp. 883–884.
[22]. J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In Proc. of the 14th ACM SIGKDD International conference on Knowledge discovery and data mining, 2008, pp. 990–998.
[23]. W. Ammar, D. Groeneveld, C. Bhagavatula et al. Construction of the literature graph in semantic scholar. arXiv preprint arXiv:1805.02262, 2018.
[24]. H. Zaugg, R.E. West, I. Tateishi, and D.L. Randall. Mendeley: Creating communities of scholarly inquiry through research collaboration. TechTrends, vol. 55, no. 1, 2011, pp. 32–36.