Метод подбора экспертов на основе тематического анализа больших массивов научно-технических документов*

Д.В. Зубарев[I], И.В. Соченков [II, I], И.А. Тихомиров[I], О.Г. Григорьев[I]

[I] Федеральный исследовательский центр «Информатика и управление» РАН, г. Москва, Россия
[II] Сколковский институт науки и технологий, г. Москва, Россия

Аннотация. Статья посвящена задаче подбора экспертов для заданного объекта экспертизы. Представлен обзор зарубежной и отечественной литературы. Показаны основные проблемы, связанные с подходами, которые используются в настоящее время для решения поставленной задачи. Для преодоления этих проблем предлагается использовать накопленные массивы неструктурированной информации, имеющей прямое отношение к экспертам. Описывается метод, выполняющий поиск и ранжирование экспертов для заданного объекта экспертизы с использованием способа поиска тематически похожих документов. Дана методика оценки качества предложенного метода и результаты экспериментальных исследований, проведенных на коллекции заявок РФФИ.

Ключевые слова: научная экспертиза, подбор экспертов, анализ неструктурированных данных, анализ текстов, мера тематического сходства, поиск похожих документов.

DOI 10.14357/20718594190206

Введение

Рецензирование научных публикаций и экспертиза заявок на получение грантов являются важными процессами научной деятельности. Выбор независимых и компетентных экспертов – ключевой момент практически любой экспертизы, от которого зависит качество принимаемых решений. В настоящее время, как правило, назначение экспертов происходит на основании кодов составленного вручную классификатора. Эксперты и авторы самостоятельно присваивают коды из такого классификатора себе и объекту экспертизы (заявке, отчету, публикации и т.д.), а назначение экспертов происходит путем сопоставления присвоенных кодов. Классификаторы имеют свойство устаревать, редко обновляются, у них неравномерное покрытие предметной области (одному коду могут соответствовать тысячи объектов, а другому – десятки) и ряд других недостатков, свойственных составленным вручную таксономиям. Таким образом, поскольку эффективность назначения по кодам полностью зависит от используемого классификатора, вышеназванные причины не позволяют качественно решить задачу сопоставления «эксперт – объект экспертизы» только на основании сопоставления кодов. Кроме того, эксперты часто приписывают себе несколько кодов, притом, что покрытие ими компетенций экспертов неравномерно [1]. При наличии нескольких десятков экспертов, соответствующих одному и тому же коду (что бы-

* Работа выполнена при поддержке РФФИ (проект 18-29-03087).

✉ *Зубарев Денис Владимирович. E-mail: zubarev@isa.ru*

вает достаточно часто), дальнейший выбор крайне субъективен и непрозрачен (фактически осуществляется ручной выбор эксперта). Все это приводит к недостаточному соответствию компетенции выбранного эксперта и объекта экспертизы. Как следствие, из-за несовпадения реальных компетенций эксперта и предлагаемых ему объектов для рецензирования, происходят отказы от экспертизы, либо она проводится некомпетентными экспертами. Именно поэтому при назначении эксперта важно не только определить соответствие тематической направленности деятельности его и объекта экспертизы, но и иметь ряд дополнительных характеристик, позволяющих выполнить ранжирование экспертов.

В настоящее время информация о компетенциях экспертов в неявном виде накоплена в больших массивах данных и текстов (история ранее выполненных экспертиз, научные статьи, научно-технические отчеты, патенты и т.д.). Эта информация более полно характеризует область компетентности эксперта, нежели коды классификатора или ключевые слова, задаваемые вручную.

В статье описывается метод, выполняющий поиск и ранжирование экспертов для заданного объекта экспертизы с помощью способа поиска тематически похожих документов. Для работы метода необходимо наличие базы данных экспертов и большого массива текстов, связанных с экспертами. Предполагается, что метод станет основой для целого класса методов, использующих неструктурированную информацию при сопоставлении «эксперт – объект экспертизы».

1. Обзор литературы

Задача автоматизации поиска экспертов для проведения экспертиз давно является предметом научных исследований. Часто исследователи сужают область применения разрабатываемых методов, например, до назначения экспертов на рецензирование статей, подаваемых на конференции [2] или до их выбора для ответа на вопрос пользователей по корпоративной базе знаний [3]. Обычно методы назначения экспертов делят на две группы [4].

Первая группа включает методы, требующие от экспертов или авторов объекта экспертизы некоторых дополнительных действий. Один из методов этой группы предполагает просмотр экспертом поданных аннотаций статей и самостоятельную оценку своей готовности отрецензировать какую-либо из рассматриваемых работ. Другой метод этой группы предполагает выбор экспертом ключевых слов, характеризующих его компетенции, из предоставленного организаторами конференции списка и сопоставление этих ключевых слов от эксперта с ключевыми словами, выбранными авторами статьи. Этот подход используется в системах управления конференциями EasyChair [5]. Он хорошо работает для конференций небольшого размера, но не масштабируется для мероприятий, в которых могут участвовать несколько десятков тысяч участников. Даже при относительно небольших конференциях, подход с ключевыми словами неприемлем, если набор тем мероприятия достаточно большой.

Вторая группа включает методы, которые автоматически строят модель компетенций эксперта по его статьям и/или другим данным и сопоставляют полученную модель со статьей для рецензирования, представленной в рамках такой же модели. Такой метод был исследован в работе [6]. В этой работе имя и фамилия эксперта посылались в качестве запроса в Google Scholar и CiteSeer. Для полных текстов найденных статей измерялось евклидово расстояние с поданной статьей. Такой метод не учитывает однофамильцев, динамику изменения интересов эксперта, возможный конфликт интересов и является вычислительно интенсивным. Другой метод [7] использует аннотации и заголовки для классификации статей по темам, заранее заданным организаторами конференции. Однако далеко не всегда возможно заранее задать конкретный набор тем.

В методе, представленном в работе [8], используются библиографические ссылки из списка литературы поданной статьи. Из библиографических ссылок извлекаются имена и фамилии авторов, для них, с помощью внешних ресурсов (DBLP), выявляются их соавторы и т.д. Таким образом, строится граф соавторства, на котором запускается вариация алгоритма ссылочного ранжирования, с целью выявления экспертов.

В работе [9] для определения близости между работами эксперта и поданной статьей используется специальная мера сходства, которая сравнивает библиографические списки. Сравнение происходит по заголовкам и авторам. Также учитывается случай, когда поданная ста-

тья цитирует работы экспертов. Сравнение библиографических списков – достаточно эффективная операция, однако представляется затруднительным оценить компетентность эксперта только по библиографическим связям, без использования полных текстов.

В работе [10] подход к решению задачи выбора экспертов с использованием семантической близости заключается в следующем: компетенции эксперта описываются с помощью онтологической модели предметной области. Модель строится на основе документов, автором которых он является (публикации, доклады, отчеты, проекты и др.). Документ, подлежащий экспертизе, также представляется своим онтологическим описанием. Онтологическое описание представляется вектором весов понятий (bag of concepts). Причем для каждого документа составляется набор моделей документа в различных областях знаний, который называется профилем документа. Аналогично профиль эксперта представляется как объединение всех профилей документов эксперта. Для вычисления близости профилей рецензируемого документа и эксперта используется мера близости, которая включает вычисление сходства между каждой парой понятий из сравниваемых профилей. Этот подход предполагает наличие готовых онтологий для различных областей знаний. Построение онтологий является очень ресурсоемкой задачей, в связи с необходимостью их постоянной актуализации. Это делает такой подход неприменимым для задач, где требуется проведение экспертиз по большому набору работ, относящихся к разным областям знаний.

Таким образом, существующие методы подбора экспертов не используют всю доступную информацию, релевантную этой задаче. Ряд методов основывается на анализе полных текстов статей для определения тематической принадлежности, но при этом не сопоставляются библиографические списки или дополнительные характеристики (сведения о соавторстве, аффилиациях). В других методах ограничиваются обработкой библиографических списков или аннотации с заголовками, игнорируя при этом полные тексты статей. Кроме того, следует отметить, что часть описанных методов являются вычислительно затратными, ввиду того, что они не используют эффективные средства индексации, и при подборе экспертов для каждого нового объекта экспертизы требуется повторять вычислительно сложные операции. В предлагаемом в этой статье подходе основная часть вычислительно интенсивных операций производится только один раз.

2. Метод подбора экспертов

На первом шаге для заданного объекта экспертизы (заявки) производится поиск тематически похожих документов по коллекциям научно-технических текстов [11]. Это могут быть научные статьи, патенты и прочие документы, связанные с экспертами. Предварительно эти коллекции индексируются. Перед индексацией текст подвергается полному лингвистическому анализу: проводятся морфологический, синтаксический и семантический анализ [12, 13]. В индексах сохраняются дополнительные характеристики для каждого слова (семантические роли, синтаксические связи и пр.) [14]. В процессе индексации создается несколько типов индексов, в том числе инвертированный индекс слов и словосочетаний, используемый при поиске тематически похожих документов. Индексация является инкрементальной, т.е. после первоначальной индексации возможно пополнение коллекций новыми текстами без переиндексации всего массива [11].

При поиске тематически похожих документов заданный представляется в виде вектора, элементами которого являются tf-idf-веса ключевых слов и словосочетаний. Стоит отметить, что словосочетания выделяются на основе синтаксических связей между словами. Это позволяет выделять словосочетания, состоящие из слов, которые в предложении идут непоследовательно, но имеют синтаксическую связь между собой. Например, для фрагмента «поиск электронных изображений» будут выделены словосочетания «поиск изображений» и «электронные изображения». После построения векторного представления заданного текста из соответствующего индекса выбираются документы, которые пересекаются по словам и словосочетаниям с ключевыми словами и словосочетаниями заданного документа (сравнение идет по векторным представлениям). Между вектором исходного документа и векторами выбранных документов из индекса вычисляется степень близости. Для вычисления степени близости используется модифицированная мера Хэмминга. Основные параметры метода поиска тематически похожих документов представлены Табл. 1.

Табл. 1. Основные параметры метода подбора экспертов

№	Описание	Название
1	Доля слов и словосочетаний исходного документа, на основе которых определяется схожесть документов	TOP_PERCENT
2	Максимальное число ключевых слов и словосочетаний, которые используются для определения схожести документов	MAX_WORDS_COUNT
3	Минимальное число ключевых слов и словосочетаний, которые используются для определения схожести документов	MIN_WORDS_COUNT
4	Минимальный tf-idf-вес слова или словосочетания, входящего в топ ключевых слов документа	MIN_WEIGHT
5	Минимальное значение оценки близости	MIN_SIM
6	Максимальное количество похожих документов для исходного документа	MAX_DOCS_COUNT

На основе списка тематически похожих документов составляется список кандидатов в эксперты. Это тривиальная операция, потому что документы связаны с экспертом: эксперт является одним из авторов, он рецензировал эту статью/заявку и пр.

После этого происходит удаление экспертов из списка кандидатов по различным критериям при наличии необходимой метаинформации. Например, если для рецензируемого документа и эксперта доступна метаинформация о принадлежности к организациям, то появляется возможность отсеять часть экспертов по причине конфликта интересов. На данный момент этот шаг зависит от доступной метаинформации и связан с типом рецензируемого документа. В экспериментальной реализации метода использовано несколько фильтров, которые уместны для заявок на гранты:

1. Отсеиваются все эксперты, которые задействованы в качестве исполнителей или в качестве руководителя в заданной заявке.

2. Отсеиваются все эксперты, работающие в тех же организациях, что и руководитель заданной заявки. Причем учитываются все организации, в которых числились эксперты и руководитель заданной заявки, согласно предоставленным данным.

После этого рассчитывается релевантность каждого эксперта объекту экспертизы. При расчете учитывается оценка сходства документов, с которыми связан эксперт, с рецензируемым документом, а также ряд дополнительных критериев с разным весом. В случае если с экспертом связано несколько документов, то их оценка сходства усредняется. Набор дополнительных критериев зависит от типа рецензируемого документа. В реализованном методе использовался один дополнительный бинарный критерий: совпадение кода области знаний, присвоенного эксперту и рецензируемому документу. Общая оценка релевантности эксперта вычисляется по формуле: $W_{sim} \cdot S_{sim} + W_{sci} \cdot S_{sci}$, где S_{sim}, S_{sci} оценки сходства документов и совпадения научных направлений соответственно, а W_{sim}, W_{sci} – веса этих критериев с условием $W_{sim} + W_{sci} = 1$. В экспериментальной реализации значения этих параметров были равны 0,9 и 0,1 соответственно. Эти значения были выбраны эмпирически. Дополнительный критерий S_{sci} был полезен при ранжировании экспертов, являвшихся руководителями междисциплинарных проектов. Междисциплинарный проект может относиться к нескольким научным областям, однако руководитель является экспертом только в своей профильной области, поэтому он должен быть ранжирован ниже, чем эксперты, у которых совпадает область знаний. Оценка релевантности каждого эксперта может лежать в промежутке [0; 1]. После оценки эксперты ранжируются по убыванию релевантности.

3. Описание экспериментальных исследований

3.1. Исходные данные

В результате сотрудничества с Российским Фондом Фундаментальных Исследований удалось провести ряд экспериментов на накопленных фондом заявках по различным конкурсам, проводимым с 2012 по 2014 годы. Фондом было предоставлено API для организации индексации полных текстов заявок. Текст заявки включал в себя:

- аннотацию проекта;

- описание фундаментальной научной проблемы, на решение которой направлен проект;
- цели и задачи исследования;
- предлагаемые методы;
- современное состояние исследований в данной области науки;
- ожидаемые научные результаты;
- описание научного задела;
- другие содержательные разделы, которые входят в форму, определенную конкурсом, в котором участвует заявка.

Для каждой заявки была предоставлена обезличенная метаинформация, содержавшая следующие поля:

- идентификатор документа;
- идентификатор руководителя;
- идентификатор организации, в которой работает руководитель;
- список идентификаторов исполнителей;
- год написания заявки;
- код области знаний, к которой принадлежит заявка (биология, химия и т.д.);
- основной код и дополнительные коды заявки;
- ключевые слова заявки.

Также была предоставлена обезличенная информация об экспертах, которые занимались рецензированием заявок:

- идентификатор эксперта;
- идентификатор организации, в которой работает эксперт;
- ключевые слова эксперта;
- код основной области знаний эксперта;
- заявки, в которых эксперт является руководителем (список идентификаторов);
- заявки, в которых эксперт является исполнителем (список идентификаторов);
- заявки, которые эксперт рецензировал (список идентификаторов);
- заявки, которые эксперт отказался рецензировать (список идентификаторов).

Размер коллекции заявок составлял около 65 тысяч документов. Также была получена информация о трех тысячах экспертов, при этом доля экспертов, являвшихся руководителем хотя бы одной заявки, составляла 78%. Сначала предполагалось использовать для поиска только заявки экспертов, в которых они являются руководителями. Но как оказалось, доля таких документов составляла около 9%. И как видно из Рис. 1 большая часть экспертов была связана только с одним проектом.

Чтобы увеличить количество документов, связанных с экспертами, мы стали также учитывать те заявки, в которых эксперт участвовал в качестве исполнителя. После этого доля экспертов с проектами возросла до 88%, а доля документов, связанных с экспертами, поднялась до 19%. Также почти вдвое уменьшилось количество экспертов, связанных только с одним документом, что видно из Рис. 2.

Рис. 1. Распределение количества проектов, связанных с экспертом

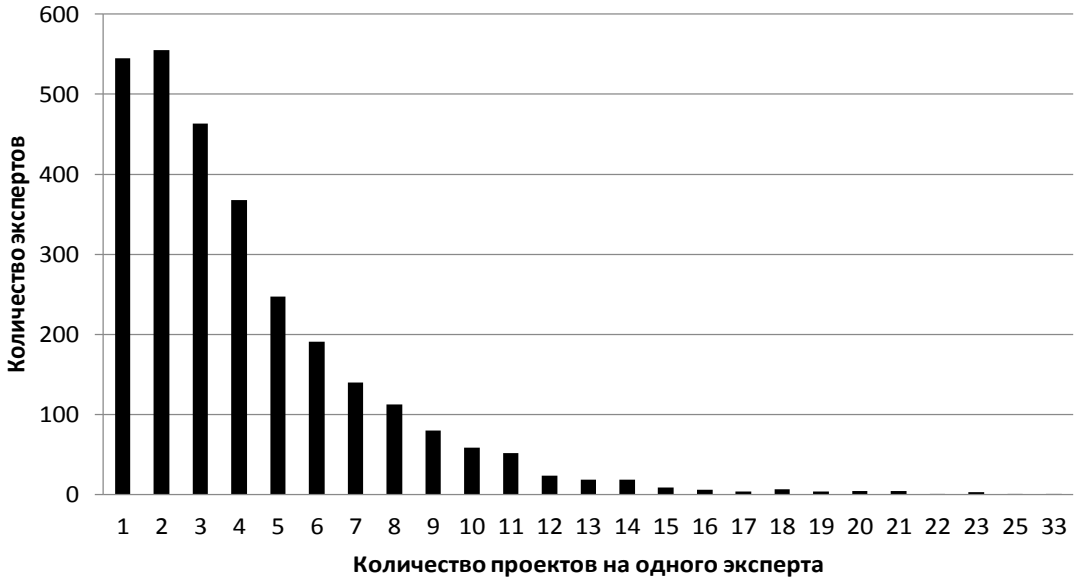

Рис. 2. Распределение количества проектов, связанных с экспертом с учетом экспертов-исполнителей

3.2. Методика оценки

Для оценки полноты и точности использовались данные о предыдущих назначениях экспертов на заявки из конкурса А-2013 (всего 10 тысяч заявок, в среднем 3 эксперта на одну заявку). Для заявок из этого конкурса с помощью предложенного метода формировался ранжированный список экспертов. После этого список полученных экспертов сравнивался с экспертами, назначенными на заданную заявку. Полнота вычислялась по формуле:

$$Recall = \frac{F_{found}}{F_{total}},$$

где F_{found}– количество найденных экспертов из числа тех, которые были назначены на эту заявку; F_{total} – количество назначенных экспертов, которые могли бы быть найдены методом (т.е. учитываются только эксперты, с которыми связан хотя бы один документ).

Для усреднения по всем заявкам использовалось микроусреднение (т.е. для всех заявок суммировались F_{found} и F_{total}, и на основе этого вычислялась искомая метрика). Также полнота вычислялась отдельно для каждой области знаний.

Стандартный способ измерения точности в данной ситуации не подходит в связи с тем, что неизвестно, является ли релевантным найденный эксперт, который не был назначен на эту заявку. Он может оказаться релевантным для этой заявки, но не был назначен по причине занятости на других проектах или по другим причинам. Такая оценка в данной работе не проводилась. Поэтому для оценки точности использовалась информация об отказах от экспертизы данной заявки. Всего таких случаев было около 2 тысяч согласно предоставленным данным. Идея заключается в том, что в сформированном списке экспертов должны отсутствовать эксперты, отказавшиеся от рецензирования данной заявки. Точность вычислялась по формуле:

$$Precision = 1 - \frac{R_{found}}{R_{total}},$$

где R_{found}– количество найденных экспертов из тех, которые отказались от экспертизы этой заявки; R_{total} – количество отказавшихся экспертов, которые могли бы быть найдены методом (т.е. эксперты со связанными документами).

3.3. Оптимизация параметров

Оптимизация параметров алгоритма производилась на отдельной выборке объемом 5 тысяч заявок. Для оптимизации использовался перебор по сетке отдельного параметра алгоритма с фиксированным значением остальных параметров. Оптимизация производилась с целью максимизировать полноту. Полученные в результате этого значения параметров метода представлены в Табл. 2.

3.4. Результаты экспериментов

Было проведено два эксперимента: в первом для поиска экспертов использовались только

Табл. 2. Значения параметров метода после оптимизации

Название	Значение
TOP_PERCENT	0,6
MAX_WORDS_COUNT	350
MIN_WORDS_COUNT	15
MIN_WEIGHT	0,03
MIN_SIM	0,05
MAX_DOCS_COUNT	2000

заявки, в которых эксперт выступал в качестве руководителя (эксперты-руководители); во втором к экспертам также были привязаны документы, в которых они числились в качестве исполнителей (+эксперты-исполнители). В Табл. 3 приведены микроусредненные значения полноты и точности для этих двух экспериментов.

Видно, что добавление новых документов, связанных с экспертами, увеличило значение полноты, но в то же время почти на столькоже пунктов упала точность.

Так как результатом работы метода является ранжированный список экспертов, то есть возможность построить график зависимости полноты и точности от глубины выдачи (т.е. увеличения числа экспертов для данной заявки), изображенный на Рис. 3.

Табл. 3. Оценки качества

Значения	Эксперты-руководители	+Эксперты-исполнители
Полнота	0,59	0,71
Точность	0,62	0,49
Значения	Эксперты-руководители	+Эксперты-исполнители

Из графика видно, что достижение максимальной полноты происходит, если брать большой топ (80-100) выдачи релевантных экспертов, после этого полнота практически не изменяется. Стоит учитывать, что при промышленном применении подобранные эксперты должны быть распределены между несколькими десятками или сотнями заявок, и на каждую заявку обычно назначается несколько экспертов. Поэтому для каждой заявки требуется достаточно большой пул релевантных экспертов. Также были рассчитаны значения полноты для каждой области знаний отдельно. Полученные результаты представлены на Рис. 4, расшифровка кодов областей знаний – в Табл. 4.

Лучшие результаты были получены в областях Науки о Земле (5) и Науки о человеке и обществе (6) – полнота около 90%. В остальных областях были получены хорошие результаты (полнота от 70 до 80%). Средние результаты были получены для областей Информационные

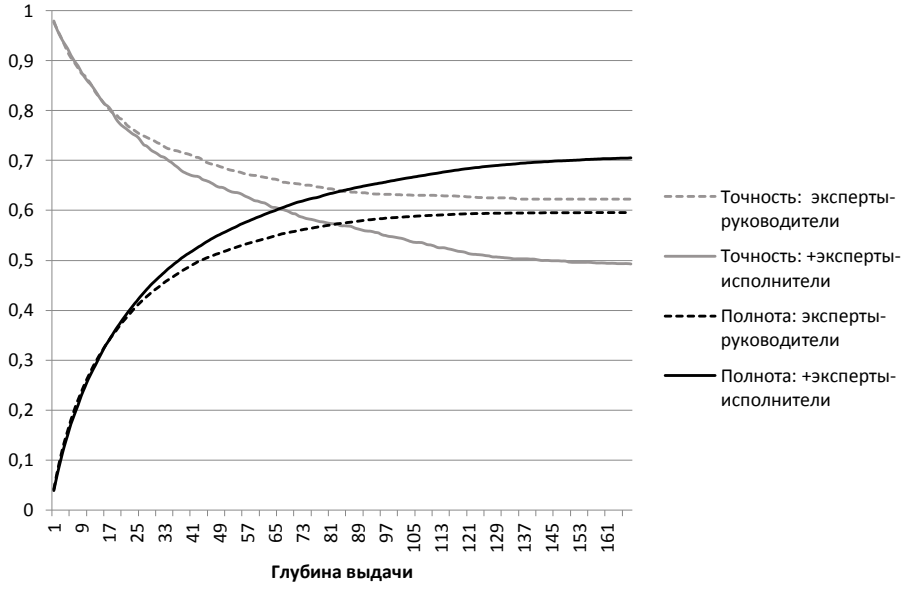

Рис. 3. Зависимость полноты и точности от глубины выдачи

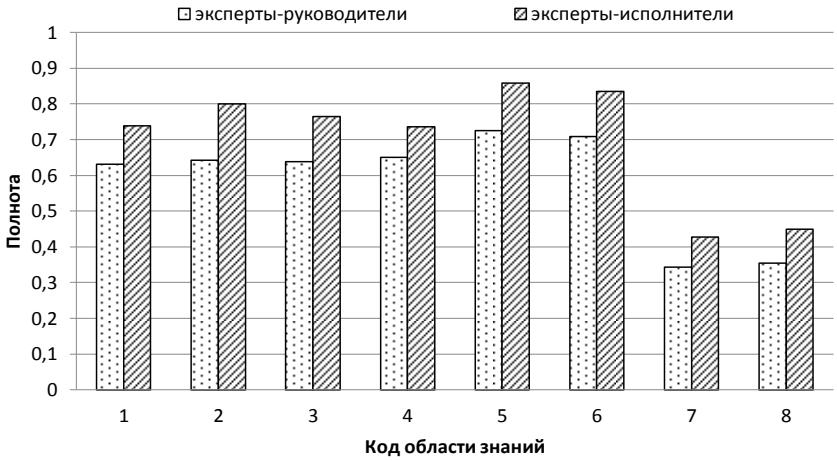

Рис. 4. Полнота для каждой области знаний

Табл. 4. Расшифровка кодов областей знаний

Код	Название
1	Математика информатика и механика
2	Физика и астрономия
3	Химия
4	Биология и медицинская наука
5	Науки о Земле
6	Науки о человеке и обществе
7	Информационные технологии и вычислительные системы
8	Фундаментальные основы инженерных наук

технологии (7) и Фундаментальные основы инженерных наук (8). Кроме того, рисунок показывает, что увеличение количества документов, связанных с экспертами, положительно влияет на полноту подбора экспертов во всех областях знаний.

Заключение

В статье описан метод подбора экспертов на основе анализа текстовой информации, представлены результаты экспериментов. Предложена новая методология оценки и проведены эксперименты на наборе данных научного фонда, что отличает данную работу от предыдущих. Метод показал свою состоятельность, однако требуется произвести доработку в первую очередь с целью увеличения полноты. Перспективным направлением является добавление дополнительных текстов, автором которых является эксперт: в первую очередь научных публикаций, а также научно-технических отчетов, патентов и пр. С учетом того, что научные публикации являются неструктурированной информацией и могут быть связаны с экспертами только посредством ФИО, необходимо также решить задачу разрешения неоднозначности имен авторов. Для решения этой задачи планируется использовать методы иерархической кластеризации. В качестве признаков могут быть использованы имена соавторов и заголовки публикаций, а также тематическая близость полных текстов. Также можно расширить список документов, связанных с экспертом, путем использования тех, к которым он имел иное отношение кроме авторства, например, рецензировал. Эти документы должны давать вклад в общую оценку релевантности с заниженным коэффициентом в связи с тем, что эксперт не имеет прямого отношения к тексту, однако если он регулярно рецензирует тексты определенной тематики, то это должно учитываться при назначении.

В дальнейших экспериментах также предполагается расширить набор критериев, влияющих на оценку релевантности эксперта для за-

данного объекта экспертизы, например, добавить ранг эксперта, вычисленный с помощью алгоритма ссылочного ранжирования на основе графа цитирований работ эксперта. Предполагается улучшить методику оценки качества назначения экспертов.

В методике, предложенной в статье, имеется ряд недостатков, а именно: зависимость полноты от первоначального назначения экспертов, которое могло быть субъективным; невозможность оценить подобранных рассматриваемым методом экспертов, которые не были задействованы при экспертизе данной заявки, что делает невозможным вычислить точность подбора. Измерение точности на основе случаев отказа от экспертизы также не является оптимальным, если учесть, что случаев отказа в 15 раз меньше, чем проведенных экспертиз, и отказ от экспертизы мог произойти по иным причинам, нежели несовпадение компетенций эксперта и тематики заявки. Однако вопрос, как оценивать работу метода подбора экспертов является в настоящий момент нерешенным. Привлечение внешних экспертов может значительно улучшить качество оценки, но для этого потребуется большое количество экспертов из разных областей знаний, которые, к тому же, должны быть хорошо знакомы с экспертным сообществом. Параметры, влияющие на оценку качества, могут быть улучшены при обработке существенно большего количества заявок с одновременным семантическим анализом причины отказа (эксперт должен объяснить отказ от экспертизы).

Предлагаемый метод можно использовать не только в процессах назначения экспертов для заявок на гранты научных фондов, но и при подборе рецензентов для любого объекта, представленного в виде текста: статей в научные журналы, тезисов конференций, заявок на получение патента и пр.

Литература

1. В Российском научном фонде прошло заседание экспертного совета по научным проектам: [Электронный ресурс]. URL: http://rscf.ru/ru/node/2367 (дата обращения: 09.12.2018)
2. Dumais S. T., Nielsen J. Automating the assignment of submitted manuscripts to reviewers //Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval. – ACM, 1992. – С. 233-244.
3. Balog K., Azzopardi L., De Rijke M. Formal models for expert finding in enterprise corpora //Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. – ACM, 2006. – С. 43-50.
4. Kalmukov Y., Rachev B. Comparative analysis of existing methods and algorithms for automatic assignment of reviewers to papers //arXiv preprint. – 2010. URL: https://arxiv.org/pdf/1012.2019.pdf (дата обращения: 09.12.2018).
5. EasyChair Conference Management System [Электронный ресурс]. URL: http://www.easychair.org / (дата обращения: 09.12.2018).
6. Pesenhofer A., Mayer R., Rauber A. Improving scientific conferences by enhancing conference management systems with information mining capabilities //Digital Information Management, 2006 1st International Conference on. – IEEE, 2006. – С. 359-366.
7. Ferilli S. et al. Automatic topics identification for reviewer assignment //International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. – Springer, Berlin, Heidelberg, 2006. – С. 721-730.
8. Rodriguez M. A., Bollen J. An algorithm to determine peer-reviewers //Proceedings of the 17th ACM conference on Information and knowledge management. – ACM, 2008. – С. 319-328.
9. Li X., Watanabe T. Automatic paper-to-reviewer assignment, based on the matching degree of the reviewers //Procedia Computer Science. – 2013. Т. 22. – С. 633-642.
10. Панкова Л. А., Пронина В. А., Крюков К. В. Онтологические модели поиска экспертов в системах управления знаниями научных организаций // Проблемы управления. – 2011. – №. 6. – С. 52-60.
11. Соченков И. В., Зубарев Д. В., Тихомиров И. А. Эксплоративный патентный поиск //Информатика и ее применения. – 2018. – Т. 12. – №. 1. – С. 89-94.
12. Osipov G. et al. Relational-situational method for intelligent search and analysis of scientific publications //Proceedings of the Integrating IR Technologies for Professional Search Workshop. – 2013. – С. 57-64.
13. Shelmanov A. O., Smirnov I. V. Methods for semantic role labeling of Russian texts //Computational Linguistics and Intellectual Technologies, Papers from the Annual International Conference “Dialogue. – 2014. – №. 13. – С. 607-620
14. Соченков И. В., Суворов Р. Е.. Сервисы полнотекстового поиска в информационно-аналитической системе (Часть 1) // Информационные технологии и вычислительные системы. №2.-2013.- С. 69-78

Expert assignment method based on similar document retrieval from large text collections

D.V. Zubarev[I], I.V. Sochenkov[II, I], I.A. Tikhomirov[I], O.G. Grigoriev[I]

[I] Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia
[II] Skolkovo Institute of Science and Technology, Moscow, Russia

Abstract. The article is devoted to the task of expert assignment. The article provides an overview of methods, which are currently used to solve this task. We discuss the main problems of those methods and propose to leverage large collections of documents that are authored by the experts. The article describes a basic method for searching and ranking of experts for a given document, using similar document retrieval. For the evaluation of the proposed method we use private corpus of applications for a grant. Experimental studies show that the more documents are available that are authored by experts, the better recall becomes. In conclusion we discuss current limitations of proposed method and describe future work to use more features from texts, such as bibliography, co-authors information etc.
Keywords: scientific expertise, expert assignment, analysis of unstructured data, text analysis, similar document retrieval.

DOI 10.14357/20718594190206

References

1. V Rossijskom nauchnom fonde proshlo zasedanie ehkspertnogo soveta po nauchnym proektam [The Russian Scientific Foundation held a meeting of the expert council on scientific projects]. Available at: http://rscf.ru/ru/node/2367 (accessed: December 9, 2018).
2. Dumais, Susan T., and Jakob Nielsen. 1992. Automating the assignment of submitted manuscripts to reviewers. Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval. ACM. 233-244.
3. Balog, Krisztian, Leif Azzopardi, and Maarten De Rijke. 2006. Formal models for expert finding in enterprise corpora. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM. 43-50.
4. Kalmukov, Yordan, and Boris Rachev. 2010. Comparative analysis of existing methods and algorithms for automatic assignment of reviewers to papers. arXiv preprint Available at: https://arxiv.org/pdf/1012.2019.pdf (accessed: December 9, 2018).
5. EasyChair Conference Management System. Available at: http://www.easychair.org/ (accessed: December 9, 2018)
6. Pesenhofer, Andreas, Rudolf Mayer, and Andreas Rauber. 2006. Improving scientific conferences by enhancing conference management systems with information mining capabilities. Digital Information Management, 2006 1st International Conference on. IEEE. 359-366.
7. Ferilli, Stefano, et al. 2006. Automatic topics identification for reviewer assignment. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Springer, Berlin, Heidelberg. 721-730.
8. Rodriguez, Marko A., and Johan Bollen. 2008. An algorithm to determine peer-reviewers. Proceedings of the 17th ACM conference on Information and knowledge management. ACM. 319-328.
9. Li, Xinlian, and Toyohide Watanabe. 2013. Automatic paper-to-reviewer assignment, based on the matching degree of the reviewers. Procedia Computer Science 22: 633-642.
10. Pankova L. A., Pronina V. A., Kryukov K. V. 2011. Ontologicheskie modeli poiska ehkspertov v sistemah upravleniya znaniyami nauchnyh organizacij [Using ontology for expert finding in knowledge management systems of scientific organizations]. Problemy upravleniya [Control sciences]. 6:52-60.
11. Sochenkov I. V., Zubarev D. V., Tihomirov I. A. 2018. Eksplorativnyj patentnyj poisk [Exploratory patent search]. Informatika i ee primeneniya [Informatics and its Applications]. 12(1): 89-94.
12. Osipov, Gennady, et al. 2013. Relational-situational method for intelligent search and analysis of scientific publications. Proceedings of the Integrating IR Technologies for Professional Search Workshop. 57-64.
13. Shelmanov, A. O., and I. V. Smirnov. 2014. Methods for semantic role labeling of Russian texts. Computational Linguistics and Intellectual Technologies. Proceedings of International Conference Dialog. Vol. 13. No. 20. 607-620.
14. I. V. Sochenkov, R. E. Suvorov. 2013. Servisy polnotekstovogo poiska v informacionno-analiticheskoj sisteme (Chast' 1) [Full-text search in the information-analytical system (Part 1)]. Informacionnye tekhnologii i vychislitel'nye sistemy [Journal of Information Technologies and Computing Systems]. 2:69-78.