Подбор экспертов на основе сопоставления графов цитирований и оценки тематической близости документов*

Д. В. Зубарев [I], А. А. Рыжова [I], Г. В. Овчинников [II], Д. А. Девяткин [I], И. В. Соченков [I]

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

Аннотация. Статья посвящена задаче отбора кандидатов, обладающих экспертными знаниями в определенной области. Предлагаются методы оценки сходства на основе графов цитирования и методы поиска похожих документов. Приводится методика оценки качества предлагаемых методов и результаты экспериментальных исследований, которые были проведены на коллекции заявок по грантам Российского Фонда Фундаментальных исследований. Выявлено, что классические методы сопоставления графов цитирований и методы, основанные на глубоком обучении, дают схожие результаты. Кроме того, методы, основанные на оценке тематического сходства текстов, имеют более высокую точность подбора экспертов, чем методы сравнения графов цитирований. Предложенные методы могут быть применены не только в ходе экспертизы заявок на гранты научных фондов, но и при подборе рецензентов для анализа любых объектов, представимых в виде текстов, содержащих цитирования.

Ключевые слова: подбор экспертов, анализ графов, SimRank, DeepWalk, поиск тематически-близких документов.

DOI 10.14357/20718632210405

Введение

Рецензирование научных публикаций и рассмотрение заявок на гранты являются важными этапами научной деятельности. Выбор независимых и компетентных рецензентов является ключевым практически в любой экспертизе, так как от этого выбора зависит качество принимаемых решений. В настоящее время, как правило, назначение экспертов основывается на классификаторах, составленных вручную (УДК, ГРНТИ). Эксперты и авторы самостоятельно присваивают коды из таких классификаторов себе и объекту экспертизы (заявке, отчету, публикации и т. д.), а эксперты назначаются путем сравнения присвоенных кодов**.** Классификаторы, как правило, устаревают, редко обновляются, неравномерно охватывают предметную область (тысячи объектов могут соответствовать одному коду, а десятки - другому) и имеют ряд других недостатков, присущих таксономиям, составленным вручную. Таким образом, поскольку результат присвоения кода полностью зависит от используемого классификатора, вышеперечисленные причины не позволяют качественно решить задачу сопоставления «эксперт – объект экспертизы» только на

* Работа выполнена при финансовой поддержке РФФИ, грант № 18-29-03087 мк

основе сравнения кодов. Кроме того, эксперты часто приписывают себе несколько кодов, несмотря на то что охват экспертных компетенций этими кодами может быть неоднородным. Если существует несколько десятков экспертов, соответствующих одному и тому же коду (что случается довольно часто), дальнейший выбор крайне субъективен и непрозрачен (по сути, эксперт выбирается вручную). Все это приводит к недостаточному соответствию между компетенцией выбранного эксперта и объектом экспертизы. В результате из-за несоответствия между реальными компетенциями эксперта и объектами, предложенными ему для рассмотрения, в проведении экспертизы может быть отказано, или экспертиза будет проводиться некомпетентными экспертами. Именно поэтому при назначении эксперта важно не только определить соответствие между тематической направленностью деятельности эксперта и объектом экспертизы, но также иметь ряд дополнительных характеристик, позволяющих оценивать экспертов. В настоящее время информация о компетентности экспертов неявно накапливается в больших объемах данных и текстов (история предыдущих экспертиз, научные статьи, научно-технические отчеты, патенты и т. д.). Эта информация более полно характеризует сферу деятельности эксперта, нежели коды классификатора или ключевые слова, составленные вручную.

В этой статье мы решаем задачу подбора экспертов, компетентных в данной теме исследования, с помощью инструментов анализа графов и метода поиска тематически похожих документов.

1. Обзор методов подбора экспертов

Методы подбора экспертов для рецензирования обычно делят на две группы [1]. К первой группе относятся методы, требующие дополнительных действий со стороны экспертов или авторов. Например, эксперта могут попросить составить список основных ключевых слов, выражающих область его компетенции, который потом сравнивается со списком ключевых слов статей для рецензирования, либо он должен ознакомиться с тезисами предоставленных статей и сам определить свою готовность рассмотреть предоставляемую работу. Такие подходы могли бы подойти для небольших конференций, неприменимы на практике при наличии большого количества тем, представленных на конференции.

Вторая группа включает в себя методы, в которых модель компетентности автора строится автоматически и далее сравнивается с рецензируемой статьёй, которая представляется в том же виде. В работе [2] авторы используют аннотации и названия статей для их классификации по темам, темы определяют организаторы конференций. Недостатком данного подхода является то, что не всегда возможно определить список тем. В [3] используются специальная мера сходства, которая сравнивает списки литературы и используется для определения близости между публикациями экспертов и представлением статей. В работе [4] для представления объекта экспертизы и документов, связанных с экспертом, используется тематическое моделирование, причём распределение тем в работах эксперта корректируется с течением времени. Для измерения
сходства между распределением тем в работах эксперта и распределением тем рассматриваемого объекта используется косинусная мера близости. Также авторы применяют векторную пространственную модель с TF-IDF взвешиванием, а финальная оценка сходства получается с помощью взвешенной суммы оценок, полученной из двух подходов.

При подборе экспертов помимо информации о тематической близости документов могут быть использованы данные о цитированиях и совместном авторстве публикаций. Поиск экспертов на основе этих данных проводится путем анализа графов соавторства и цитирований. Для анализа этих графов применяют два подхода. Один из них основан на метриках близости, а второй использует векторные представления графов (graph embeddings). Для получения графа цитирования в заданной области могут использоваться разнообразные наукометрические базы, такие как Scopus, Web of Science, E-library (РИНЦ).

Меры близости на графах используются во многих прикладных задачах, таких как анализ отношений [5], поиск похожих документов ([6, 7]) и

многих других. Меры близости вершин графов используются также в анализе социальных сетей [8], электронной коммерции [9]. В статье [10] предлагается использовать меру близости под названием SimRank. В основе определения SimRank заложена идея о том, что «два объекта похожи, если на них ссылаются похожие объекты».

Векторные представления графов (*graph embeddings*) являются ещё одним популярным методом анализа графов. Они представляют собой отображение графа в пространство более низкой размерности, например, $\mathbb{R}^n$. В работе для построения таких представлений используется метод Deep-walk [11], в котором обучение представлений вершин графа осуществляется путем генерации коротких случайных блужданий по графу. Для построения этих представлений используется предобученная нейронная сеть. Полученные представления позволяют выявлять сходство вершин графа на основании их окрестностей и принадлежность к сообществам.

2. Метод SimRank

Пусть $G = (V, E)$ – граф, V – множество его вершин, а E – множество его рёбер. Порядком графа называется число его вершин $|V| = n$.

Пусть $s(a, b)$ – SimRank между вершиной $a \in V$ и вершиной $b \in V$. Предполагается, что величина $s(a, b)$ для вершин a и b принадлежит отрезку $[0,1]$, и каждая вершина максимально близка самой себе, то есть $s(a, a) = 1$ для любой вершины a. Величина SimRank для двух вершин a and b определяется следующим образом:

$$s(a,b) = \begin{cases} 1, \text{ если } a = b \\ 0, \text{если } I(a) = \emptyset \text{ или } I(b) = \emptyset \\ \dfrac{c}{I(a)I(b)} \displaystyle\sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s\left(I_i(a), I_j(b)\right) \end{cases} \quad (1)$$

где $I(v)$ – множество соседей вершины v, $c \in (0,1)$ – некоторая константа. Обозначим за S матрицу SimRank, в которой элемент (i, j) – SimRank между вершинами i и j:

$$S = [s_{ij}], \quad i, j = 1, \dots, n.$$

Чтобы найти матрицу SimRank для графа, необходимо выписать величины $s(a, b)$ для всех пар его вершин (a, b). После этого мы получаем систему из n^2 уравнений, которая имеет единственное решение [10]. В статье [12] предлагается следующее приближенное решение этой задачи.

Пусть A – матрица связности графа G, нормализованная по столбцам:

$$\sum_{i=1}^{n} A_{ij} = 1, \quad j = 1, \dots, n$$

и

$$W = \sqrt{c}A.$$

Для решения уравнения используется итерационный процесс с фиксированной точкой:

$$S = W^{\top} S W - \text{diag}(W^{\top} S W) + I. \quad (2)$$

Решение уравнения (2) существует, единственно и может быть записано в следующем виде:

$$S = \sum_{k=0}^{\infty} \mathcal{L}\left(\mathcal{D}^k(I)\right) \quad (3)$$

где $\mathcal{L}(X)$ дает решение дискретного уравнения Ляпунова L:

$$L = W^{\top} L W + X \quad (4)$$

и $\mathcal{D}(X)$ – оператор, определяемый формулой

$$\mathcal{D}(X) = -\text{diag}(W^{\top} \mathcal{L}(X) W) \quad (5)$$

Степень k оператора $\mathcal{D}(X)$ задаётся уравнением

$$\mathcal{D}^k(X) = \underbrace{\mathcal{D}\big(\dots \mathcal{D}(X)\big)}_{k \text{ times}},$$

где $\mathcal{D}\text{^}0\ (X) = X$. Таким образом, мы можем выписать ряд

$$\mathcal{S}(I) = \sum_{k=0}^{\infty} \mathcal{L}\left(\mathcal{D}^k(I)\right) \quad (6)$$

если $\lim_{k\to\infty} \| \mathcal{D}^k\ (X) \| = 0$

Чтобы построить приближение матрицы SimRank, мы суммируем линейные по W и $W^{\top}$ члены ряда (6) и добавляем остаточный член (малоранговую матрицу). Так как матрица SimRank S симметричная (2), остаточный член тоже представляется в симметричном виде. Финальная аппроксимация матрицы SimRank $\tilde{S}$:

$$\tilde{S} = I + W^{\top} W - \text{diag}(W^{\top} W) + U D U^{\top}, \quad (7)$$

где I – единичная $n \times n$ матрица, U и D – ортонормированная $n \times m$ матрица и диагональная $m \times m$ матрица соответственно. Малоранговая аппроксимация вычисляется с помощью рандомизированного подхода [13].

В ходе подбора экспертов с использованием наукометрических баз данных формируется граф цитирований работ экспертов и статей, приведенных в объектах экспертизы (заявках или препринтах публикаций), полученный граф обрабатывается алгоритмом SimRank и полученная матрица сходства используется для ранжирования экспертов. Строки матрицы сходства соответствуют объектам экспертизы, а столбцы – документам экспертов. Ранг каждого эксперта вычисляется как среднее сходство всех его документов с объектом экспертизы.

3. Метод DeepWalk

Случайное блуждание, берущее своё начало из вершины v_i обозначим за $\mathcal{W}_{v_i}$. Это стохастический процесс со случайными величинами $\mathcal{W}_{v_i}^1, \mathcal{W}_{v_i}^2, \ldots, \mathcal{W}_{v_i}^k$ такой, что $\mathcal{W}_{v_i}^{k+1}$ – это вершина, выбранная случайным образом из окрестности вершины v_k.

Оценим вероятность наблюдать вершину v_i при условии того, что предыдущие вершины уже были посещены в процессе случайного блуждания.

$$\Pr\left(v_i \mid (v_1, v_2, \cdots, v_{i-1}) \right)$$

Необходимо получить не только вероятностное распределение совместной встречаемости вершин графа, но и латентное представление этих вершин. Для этого мы вводим функцию отображения $\Phi: v \in V \mapsto \mathbb{R}^{|V| \times d}$. Это отображение Φ выражает скрытое социальное представление, связанное с каждой вершиной v в графе. (На практике, мы представляем матрицу Φ как $|V| \times d$ матрицу со свободными параметрами, которую в дальнейшем будем обозначать за X_E.) Таким образом, задача заключается в оценке правдоподобия:

$$\Pr\Big(v_i \mid \big(\Phi(v_1), \Phi(v_2), \cdots, \Phi(v_{i-1})\big) \Big) \qquad (8)$$

После преобразований необходимо решить задачу оптимизации:

$$\underset{\Phi}{\text{minimize}}\ P \text{, где}$$

$$P = -\log \Pr\big(\{v_{i-w}, \cdots, v_{i-1}, v_{i+1}, \cdots, v_{i+w}\} \mid \Phi(v_i) \big) \qquad (9)$$

Решение задачи оптимизации из уравнения (9) создаёт представления, которые отражают общие сходства в локальной структуре графа между его вершинами. Вершины, имеющие похожие окрестности, получат более близкие представления (кодирование сходства совместного цитирования), позволяя получать обобщения для задач машинного обучения.

Алгоритм состоит из двух фаз; сначала создаётся генератор случайного блуждания, а затем осуществляется процедура обновления (Алгоритм 1).

Алгоритм 1: DeepWalk

Параметры алгоритма	
Граф	$G(V, E)$
Размер окна	w
Размер эмбеддинга	d
Количество блужданий на вершину	γ
Длина блуждания	t
Матрица представлений вершин	$\Phi \in \mathbb{R}\text{^}(\lvert V\rvert \times d)$
Алгоритм	
Инициализация: семплировать Φ из $\mathcal{U}^{\lvert V\rvert \times d}$ Построить бинарное дерево T из V Начало цикла. От i = 0 до γ выполнить: $\mathcal{O} = \text{Shuffle}(V)$ Начало цикла. Для каждой вершины $v_i \in \mathcal{O}$ выполнить: $\mathcal{W}_{v_i} = \text{RandomWalk}(G, v_i, t)$ SkipGram(Φ, $\mathcal{W}_{v_i}$, w) Конец цикла Конец цикла	

В процессе подбора экспертов формируется граф цитирований работ экспертов и статей, упоминаемых в объектах экспертизы, далее полученный граф обрабатывается алгоритмом DeepWalk. Далее векторные представления вершин графа используются для оценки расстояния между рецензентами и объектами экспертизы.

4. Поиск тематически похожих документов

В этом методе на первом шаге для данного объекта экспертизы осуществляется поиск тематических похожих документов [14]. Для этого используются научные статьи, патенты и другие документы, которые имеют отношение к экспертам. Осуществляется полный лингвистический анализ текстов: морфологический, синтаксический и семантический [15, 16], а затем тексты индексируются. Индексы хранят дополнительные характеристики для каждого слова (семантические роли, синтаксические связи и т.д) [17]. Во время индексирования создаётся несколько типов индексов, в том числе инвертированный индекс слов и фраз, который используется для поиска тематически похожих документов. После первоначальной индексации новые тексты можно добавлять в коллекцию без повторной индексации всей коллекции [17].

Во время поиска тематически похожих документов данный документ представляется в виде вектора, элементами которого являются TF-IDF веса ключевых слов и фраз. Извлечение фраз осуществляется на основе синтаксических отношений между словами, что позволяет формировать фразы не только из слов, которые стоят рядом друг с другом. Затем между вектором данного документа и векторами документов проиндексированной коллекции вычисляется степень близости (мы использовали косинусную меру близости и расстояние Хэмминга). Основные параметры метода поиска похожих документов представлены в Табл. 1.

На основе списка тематически похожих документов формируется список кандидатов-экспертов. Далее при наличии мета информации некоторые кандидаты могут быть удалены из списка. В процессе экспериментов использовались следующие фильтры, которые подходят для заявок на гранты:

- Все эксперты, которые включены в список участников заявки на грант исключаются.
- Все эксперты, которые работают в той же организации, что и руководитель проекта по заявке, исключаются из списка.

После этого вычисляется соответствие каждого эксперта объекту экспертизы. При расчёте учитывается схожесть документов, а также равенство кодов области знаний, присвоенных эксперту и рассматриваемому документу. Общая оценка релевантности эксперта рассчитывается по следующей формуле (10):

$$W_{sim} \cdot S_{sim} + W_{sci} \cdot S_{sci}, \qquad (10)$$

где S_{sim} – значение близости рассматриваемого документа и документа эксперта (если у эксперта несколько документов, то значение мер близости усредняется), S_{sci} равняется 0, если код области знаний, присвоенный эксперту, не равен коду рассматриваемого документа, иначе 1. W_{sim} и W_{sci} – веса, с условием что

Табл. 1. Основные параметры метода поиска тематически похожих документов

Описание	Имя
Процент слов и фраз в исходном документе, которые определяют сходство документов	TOP_PERCENT
Максимальное число слов и фраз, которые используются для определения сходства документов	MAX_WORDS_COUNT
Минимальное число слов и фраз, которые используются для определения сходства документов	MIN_WORDS_COUNT
Минимальный TF-IDF вес слова или фразы для включения в топ ключевых слов документа	MIN_WEIGHT
Минимальное значение степени близости	MIN_SIM
Максимальное число похожих документов для исходного документа	MAX_DOCS_COUNT

$W_{sim} + W_{sci} = 1$. Критерий S_{sci} полезен при ранжировании экспертов, которые являются руководителями междисциплинарных проектов. Междисциплинарный проект может относиться к нескольким научным областям, но руководитель является экспертом только в одной области, поэтому его рейтинг должен быть ниже, чем у экспертов, обладающих той же областью знаний. Оценка релевантности каждого эксперта находится в интервале [0; 1]; после её вычисления эксперты ранжируются в порядке убывания значимости.

5. Данные и эксперименты

Для исследования методов подбора экспертов в качестве тестового набора данных мы собрали ретроспективные данные о назначении рецензентов по заявкам РФФИ на конкурс проектов А-2013 в области физики, математики и технических наук, а также список рецензентов, которые были назначены для рассмотрения этих заявок или отказались от их рассмотрения. Список включает в себя в общей сложности 500 пар (*эксперт, назначение)*. В рамках сотрудничества с РФФИ был предоставлен API для организации индексации полных текстов заявок. Текст заявки включал в себя:

- аннотацию проекта;
- фундаментальную научную проблему, которая должна быть решена в рамках проекта;
- цели и задачи исследования;
- предлагаемые методы решения проблемы;
- текущий статус исследования в данной области;
- ожидаемые научные результаты;
- описание научной базы;
- другие содержательные разделы, которые включены в форму для участников конкурса по данной заявке.

По каждой заявке также были предоставлены анонимизированные метаданные, содержащие следующие поля:

- идентификатор документа;
- ID руководителя проекта;
- идентификатор организации руководителя проекта;
- идентификаторы исполнителей;
- год написания заявки;
- код области знаний, к которой относится заявка (биология, химия и т. д.);
- основной код и дополнительные коды заявки;
- ключевые слова.

Анонимная информация о рецензентах по заявкам включает в себя следующие поля:

- идентификатор эксперта;
- идентификатор организации, в которой работает эксперт;
- код основной области знаний эксперта;
- заявки, по которым эксперт является главным (список идентификаторов);
- заявки, в которых эксперт является исполнителем (список идентификаторов);
- заявки, рассмотренные экспертом (список идентификаторов);
- заявки, которые эксперт отказался рассматривать (список идентификаторов).

Для каждого из 500 вышеупомянутых экспертов были получены научные публикации из открытых русскоязычных источников (всего таких статей «нулевого» уровня насчитывается 50 тысяч). На 50 тысяч статей «нулевого» уровня было получено 95 тысяч текстов работ из 490 тысяч цитируемых публикаций первого уровня. Для 95 тысяч публикаций первого уровня было выделено 587 тысяч цитируемых публикаций. На втором уровне было получено 53 тысячи статей из 587 тысяч статей, которые ссылались на публикации первого уровня. Далее аналогичным образом были получены публикации на 3-м, 4-м и последующих уровнях. На основании вышеизложенного, стоит отметить, что были получены не все публикации, а только их часть. Из приведенных выше данных формировался граф цитирования, вершины которого являются документами и соединяются ребрами при наличии цитирования. Также был сформирован граф тематической близости документов, вершинами которого являются документы, а ребра взвешены методом из работы авторов статьи [14].

Для обучения моделей поиска тематически похожих документов использовались заявки, накопленные Российским фондом фундаментальных исследований (РФФИ) на различных конкурсах, проводимых с 2012 по 2014 год. РФФИ был предоставлен программный интерфейс (API) для индексации примерно 65 тысяч

заявок с информацией о трёх тысячах экспертов. Доля экспертов, являющимися руководителями одного и более проектов составила 78%. Сначала предполагалось оставить только те заявки, для которых существует информация о главном руководителе проекта, но оказалось, что доля таких документов составляет всего 9 процентов. Более того, большинство экспертов были связаны только с одним проектом по гранту. Для увеличения числа документов, связанных с экспертом, были также добавлены документы, где эксперт является соруководителем проекта. Также для поиска дополнительных научных публикаций рецензента использовался сборник научных работ, который в основном состоял из статей открытых полнотекстовых баз *mathnet.ru* и *cyberlinka.ru*. Сначала был проведён поиск документов, подтверждающих поддержку гранта с участием эксперта, в результате чего было получено дополнительно четыре тысячи документов. Кроме того, для каждой заявки был осуществлён поиск похожих документов. Для фильтрации похожих, но не связанных с экспертами документов, выполнено сравнение полных имен авторов статей с полными именами участников заявки по гранту. При совпадении хотя бы одного полного имени документ считался связанным с экспертом. В статьях чаще всего присутствуют только фамилии и инициалы авторов, в таких случаях искались два совпадения по авторам. Таким образом, при помощи поиска похожих документов было собрано ещё около 30 тысяч документов.

После выполнения всех этих процедур доля экспертов с документами возросла до 88 процентов. Кроме того, было значительно сокращено число экспертов, связанных только с одним документом.

Оптимизация параметров алгоритма поиска похожих документов была проведена на отдельной коллекции документов, состоящей из 700 заявок на гранты. Мы использовали алгоритм перебора по сетке, меняя один параметр модели и фиксируя остальные. Оптимизация основывалась на оптимизации полноты (recall). Эта оценка позволяет определить долю правильно найденных экспертов среди всех отобранных. Лучшие параметры алгоритма представлены в Табл. 2.

Были также рассмотрены различные комбинации функций близости и наборов признаков текстов (только слова, слова с именными фразами). Наилучшими параметрами по полноте (recall) оказались расстояние Хэмминга в качестве метрики близости и слова с добавлением фраз в качестве признаков (Табл. 3).

6. Результаты сопоставления методов подбора экспертов

NDCG (Normalized Discounted Cumulative Gain) была выбрана основной оценкой качества предложенных алгоритмов. По итогам экспериментального исследования алгоритмов выявлено, что SimRank и DeepWalk получают

Табл. 2. Значения параметров метода поиска похожих документов после оптимизации

Название параметра	Значение
TOP_PERCENT	0.4
MAX_WORDS_COUNT	200
MIN_WORDS_COUNT	15
MIN_WEIGHT	0.03
MIN_SIM	0.0
MAX_DOCS_COUNT	1000
W_{sci}	0.1
W_{sim}	0.9

Табл. 3. Оценки полноты (recall) поиска похожих документов при использовании различных признаков текстов и мер сходства векторов

	Recall
Косинусное расстояние, только слова	0.73
Косинусное расстояние, слова и фразы	0.75
Расстояние Хэмминга, только слова	0.76
Расстояние Хэмминга, слова и фразы	0.77

Табл. 4. Результаты оценки методов подбора экспертов, NDCG@1000

SimRank	DeepWalk	Поиск тематически похожих документов
0.175	0.172	0.328

примерно одинаковые оценки NDCG. Предположительно, причиной такого результата является недостаточное содержание информации в графе цитирования – в нём отсутствует много важных взаимосвязей между статьями, особенно в популярных областях исследований. Таким образом, в условиях разреженного графа цитирований методы, основанные на тематической близости текстов, дают более точный результат по мере NDCG (Табл. 4).

Заключение

В работе предложены и исследованы методы подбора экспертов для рецензирования заданных объектов научно-технической деятельности с учетом тематической близости текстов и данных о цитированиях и совместном авторстве публикаций. Эти методы могут быть использованы не только для подбора экспертов в ходе оценки заявок на гранты научных фондов, но и при выборе рецензентов для анализа любых объектов, представимых в виде текстов, содержащих цитирования.

Результаты исследований показывают, что в условиях разреженного графа цитирований методы, основанные на тематической близости текстов, дают более точный результат. Однако, методы, основанные на тематической близости текстов, ограниченно применимы для анализа заявок и отчетов по мультидисциплинарным проектам, содержащим большое количество разнообразных направлений исследований; в то же время полные тексты статей потенциальных экспертов могут быть недоступны. Решением этой проблемы могла бы стать интеграция методов подбора экспертов на основе цитирований и тематической близости текстов.

Еще одним перспективным направлением развития представленных методов является применение кросс-языковых подходов к поиску тематически-близких документов. Внедрение этих подходов позволило бы учитывать при подборе экспертов наличие заделов на русском и английском языке одновременно.

Литература

1. Kalmukov Y., Rachev B. 2010. Comparative analysis of existing methods and algorithms for automatic assignment of reviewers to papers. ArXiv preprint arXiv:1012.2019.
2. Rodriguez, M.A., Bollen, J.. 2008. An algorithm to determine peer-reviewers. In: Proceedings of the 17th ACM conference on Information and knowledge management, 319–328. ACM.
3. Li, X., Watanabe, T. 2013. Automatic paper-to-reviewer assignment, based on the matching degree of the reviewers. Procedia Comput. Sci. 22, 633–642.
4. Peng, H., Hu, H., Wang, K., Wang, X. 2017. Time-aware and topic-based reviewer assignment. In: Bao, Z., Trajcevski, G., Chang, L., Hua, W. (eds.) DASFAA 2017. LNCS, vol. 10179, pp. 145–157. Springer, Cham.
5. Li, Lina, Cuiping Li, Hong Chen, and Xiaoyong Du. 2013. “Mapreduce-Based SimRank Computation and Its Application in Social Recommender System.” In Big Data (Bigdata Congress), 2013 IEEE International Congress on, 133–40. IEEE.
6. Li, Pei, Zhixu Li, Hongyan Liu, Jun He, and Xiaoyong Du. 2009. “Using Link-Based Content Analysis to Measure Document Similarity Effectively.” In Advances in Data and Web Management, 455–67. Springer.
7. Williams, Kyle, Jian Wu, and C Lee Giles. 2014. “SimSeerX: A Similar Document Search Engine.” In Proceedings of the 2014 Acm Symposium on Document Engineering, 143–46. ACM.
8. Akcora, Cuneyt Gurcan, Barbara Carminati, and Elena Ferrari. 2013. “User Similarities on Social Networks.” Social Network Analysis and Mining 3 (3). Springer: 475–95.
9. Putra, Aghny Arisya, Rahmad Mahendra, Indra Budi, and Qorib Munajat. 2017. “Two-Steps Graph-Based Collaborative Filtering Using User and Item Similarities: Case Study of E-Commerce Recommender Systems.” In 2017 International Conference on Data and Software Engineering (Icodse), 1–6. IEEE.
10. Jeh, Glen, and Jennifer Widom. 2002. “SimRank: A Measure of Structural-Context Similarity.” In Proceedings of the Eighth Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 538–43. KDD ’02. New York, NY, USA: ACM.
11. Perozzi B., Al-Rfou R., Skiena S. 2014. Deepwalk: Online learning of social representations. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 701–710.
12. Oseledets, Ivan, Ovchinnikov George, and Katrutsa Alexandr. 2017. “Fast, Memory-Efficient Low-Rank Ap-

proximation of Simrank." Journal of Complex Networks 5 (1). Oxford University Press: 111–26.
13. Halko N., Martinsson P. G., Tropp J. A. 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM review. Vol. 53(2). 217-288.
14. Sochenkov, Ilya Vladimirovich, Denis Vladimirovich Zubarev, and Ilya Alexandrovich Tikhomirov. 2018. "Exploratory Patent Search." Computer Science and Its Applications 12 (1). Russian Academy of Sciences, Department of Nanotechnology; Information Technologies: 89–94.
15. Osipov, G., et al. 2013. Relational-situational method for intelligent search and analysis of scientific publications. In: Proceedings of the Integrating IR Technologies for Professional Search Workshop, 57–64.
16. Shelmanov, A.O., Smirnov, I.V. 2014. Methods for semantic role labeling of Russian texts. In: Proceedings of International Conference Dialog on Computational Linguistics and Intellectual Technologies, vol. 13, no. 20, pp. 607–620.
17. Shvets, A., Devyatkin, D., Sochenkov, I., Tikhomirov, I., Popov, K., Yarygin, K. 2015. Detection of current research directions based on full-text clustering. In: 2015 Science and Information Conference (SAI), pp. 483–488. IEEE.

Зубарев Денис Владимирович. Федеральный исследовательский центр «Информатика и управление» РАН, Москва, младший научный сотрудник. Количество печатных работ: 24. Область научных интересов: компьютерный анализ естественного языка, методы интеллектуального информационного поиска, анализ больших массивов данных, разработка распределенных программных систем. E-mail: zubarev@isa.ru

Рыжова Анастасия Александровна. Федеральный исследовательский центр «Информатика и управление» РАН, Москва, инженер-исследователь. Количество печатных работ: 7. Область научных интересов: компьютерный анализ естественного языка, машинное обучение, нейронные сети, классификация и кластеризация текстов, семантический анализ. E-mail: ryzhova@tesyan.ru

Овчинников Георгий Викторович. Сколковский институт науки и технологий. Старший научный сотрудник, к.ф-.м.н. Количество печатных работ: 29. Область научных интересов: быстрое вычисление мер подобия, индуцированных топологией, на графах и численное моделирование многомасштабных систем тензорными методами, E-mail: g.ovchinnikov@skoltech.ru

Девяткин Дмитрий Алексеевич. Федеральный исследовательский центр «Информатика и управление» РАН, Москва, научный сотрудник. Количество печатных работ: 70. Область научных интересов: компьютерный анализ естественного языка, методы извлечения информации из текстов, машинное обучение, анализ больших массивов данных, наукометрический анализ, методы оценки эффективности научной деятельности. E-mail: devyatkin@isa.ru

Соченков Илья Владимирович. Федеральный исследовательский центр «Информатика и управление» РАН, Москва, ведущий научный сотрудник, к.ф.-м.н. Количество печатных работ: 120. Область научных интересов: компьютерный анализ естественного языка, методы интеллектуального информационного поиска, динамическая контентная фильтрация, машинное обучение, анализ больших массивов данных, наукометрический анализ, методы оценки эффективности научной деятельности. E-mail: sochenkov@isa.ru

Graph and Topical Similarity-Based Methods for Assignment of Experts

D. V. Zubarev [I], A. A. Ryzhova [I], G. V. Ovchinnikov [II], D. A. Devyatkin [I], I. V. Sochenkov [I]

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

Abstract. The paper tackles the problem of selecting candidates with expert knowledge in a particular field. We propose methods for assessing similarity based on citation graphs and topical similarity of documents retrieval methods to select experts. The paper also provides a methodology for assessing the accuracy of the proposed methods and the results of experiments that were carried out on a dataset of grant applications from the Russian Foundation for Basic Research. The experimental results show that the classical methods of citation graph comparison and deep learning provide similar results. In addition, similar document retrieval methods have higher accuracy in selecting experts than citation-based methods. The proposed methods can be used not only for selecting experts for the evaluation of grant applications of scientific foundations but also for the assignment of reviewers for the analysis of any objects with text and citations.

Keywords: expert selection, graph analysis, SimRank, DeepWalk, topically similar document retrieval

DOI 10.14357/20718632210405

References

1. Kalmukov Y., Rachev B. 2010. Comparative analysis of existing methods and algorithms for automatic assignment of reviewers to papers. ArXiv preprint arXiv:1012.2019.
2. Rodriguez, M.A., Bollen, J.. 2008. An algorithm to determine peer-reviewers. In: Proceedings of the 17th ACM conference on Information and knowledge management, 319–328. ACM.
3. Li, X., Watanabe, T. 2013. Automatic paper-to-reviewer assignment, based on the matching degree of the reviewers. Procedia Comput. Sci. 22, 633–642.
4. Peng, H., Hu, H., Wang, K., Wang, X. 2017. Time-aware and topic-based reviewer assignment. In: Bao, Z., Trajcevski, G., Chang, L., Hua, W. (eds.) DASFAA 2017. LNCS, vol. 10179, pp. 145–157. Springer, Cham.
5. Li, Lina, Cuiping Li, Hong Chen, and Xiaoyong Du. 2013. "Mapreduce-Based SimRank Computation and Its Application in Social Recommender System." In Big Data (Bigdata Congress), 2013 IEEE International Congress on, 133–40. IEEE.
6. Li, Pei, Zhixu Li, Hongyan Liu, Jun He, and Xiaoyong Du. 2009. "Using Link-Based Content Analysis to Measure Document Similarity Effectively." In Advances in Data and Web Management, 455–67. Springer.
7. Williams, Kyle, Jian Wu, and C Lee Giles. 2014. "SimSeerX: A Similar Document Search Engine." In Proceedings of the 2014 Acm Symposium on Document Engineering, 143–46. ACM.
8. Akcora, Cuneyt Gurcan, Barbara Carminati, and Elena Ferrari. 2013. "User Similarities on Social Networks." Social Network Analysis and Mining 3 (3). Springer: 475–95.
9. Putra, Aghny Arisya, Rahmad Mahendra, Indra Budi, and Qorib Munajat. 2017. "Two-Steps Graph-Based Collaborative Filtering Using User and Item Similarities: Case Study of E-Commerce Recommender Systems." In 2017 International Conference on Data and Software Engineering (Icodse), 1–6. IEEE.
10. Jeh, Glen, and Jennifer Widom. 2002. "SimRank: A Measure of Structural-Context Similarity." In Proceedings of the Eighth Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 538–43. KDD '02. New York, NY, USA: ACM.
11. Perozzi B., Al-Rfou R., Skiena S. 2014. Deepwalk: Online learning of social representations. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 701–710.
12. Oseledets, Ivan, Ovchinnikov George, and Katrutsa Alexandr. 2017. "Fast, Memory-Efficient Low-Rank Approximation of Simrank." Journal of Complex Networks 5 (1). Oxford University Press: 111–26.
13. Halko N., Martinsson P. G., Tropp J. A. 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM review. Vol. 53(2). 217-288.
14. Sochenkov, Ilya Vladimirovich, Denis Vladimirovich Zubarev, and Ilya Alexandrovich Tikhomirov. 2018. "Exploratory Patent Search." Computer Science and Its Applications 12 (1). Russian Academy of Sciences, Department of Nanotechnology; Information Technologies: 89–94.
15. Osipov, G., et al. 2013. Relational-situational method for intelligent search and analysis of scientific publications. In: Proceedings of the Integrating IR Technologies for Professional Search Workshop, 57–64.
16. Shelmanov, A.O., Smirnov, I.V. 2014. Methods for semantic role labeling of Russian texts. In: Proceedings of International Conference Dialog on Computational Linguistics and Intellectual Technologies, vol. 13, no. 20, pp. 607–620.
17. Shvets, A., Devyatkin, D., Sochenkov, I., Tikhomirov, I., Popov, K., Yarygin, K. 2015. Detection of current research directions based on full-text clustering. In: 2015 Science and Information Conference (SAI), pp. 483–488. IEEE.

Zubarev D. V. Federal Research Center "Computer Science and Control" RAS, Moscow, junior researcher. Number of publications: 24. Research interests: natural language processing, intelligent information retrieval, big data, development of distributed software systems. E-mail: zubarev@isa.ru

Ryzhova A. A. Federal Research Center "Computer Science and Control" RAS, Moscow, research engineer. Number of publications: 7. Research interests: natural language processing, machine learning, neural networks, classification and clustering of texts, semantic analysis. E-mail: ryzhova@tesyan.ru

Ovchinnikov G. V. Skolkovo Institute of Science and Technology, senior researcher, Ph.D. Number of publications: 29. Research interests: fast computation of similarity measures induced by topology on graphs and numerical modeling of multiscale systems by tensor methods, E-mail: g.ovchinnikov@skoltech.ru

Devyatkin D. A. Federal Research Center "Computer Science and Control" RAS, Moscow, researcher. Number of publications:70. Research interests: natural language processing, information extraction, machine learning, big data, scientometric analysis, methods for assessing the effectiveness of scientific activities. E-mail: devyatkin@isa.ru

Sochenkov I. V. Federal Research Center "Computer Science and Control" RAS, Moscow, leading researcher, Ph.D. Number of publications: 120. Research interests: natural language processing, intelligent information retrieval, dynamic content filtering, machine learning, big data, scientometric analysis, methods for assessing the effectiveness of scientific activities. E-mail: sochenkov@isa.ru