УДК 004.4 DOI: 10.17587/prin.8.556-562

А. С. Козицын, канд. физ.-мат. наук, вед. науч. сотр., e-mail: alexanderkz@mail.ru,
С. А. Афонин, канд. физ.-мат. наук, вед. науч. сотр., e-mail: serg@msu.ru,
МГУ имени М. В. Ломоносова, г. Москва

Разрешение неоднозначностей при определении авторов публикации с использованием графов соавторства в больших коллекциях библиографических данных

Рассмотрены некоторые подходы к решению задач обработки библиографических данных в системах наукометрии с помощью выделения статистических закономерностей в больших коллекциях таких данных. На примере представительной информационно-аналитической системы "ИСТИНА", которая эксплуатируется в МГУ им. М. В. Ломоносова, представлены результаты исследований по идентификации авторов по библиографическим данным статей. Предлагаемый алгоритм позволяет выделять авторов с точностью более 95 %.

Ключевые слова: *наукометрия, обнаружение закономерностей, библиографические данные, граф, цитирование, автор, тематический анализ*

Введение

В настоящее время во многих ведущих учебных заведениях и научных центрах активно внедряют системы автоматизации сбора данных о научной и педагогической деятельности работников. Это обусловлено тем обстоятельством, что эффективное управление большими организациями в сфере науки и образования невозможно без внедрения методов оценки эффективности деятельности отдельных работников таких организаций. Для решения подобных задач необходимо использовать современные средства и системы автоматизированного сбора, верификации, хранения и анализа больших объемов библиографической информации, которая описывает результаты научной деятельности сотрудников отдельных организаций и предоставляет данные для оценки эффективности их деятельности. Набор используемых для оценки показателей такой эффективности (индикаторов) должен быть достаточно большим, чтобы охватывать все основные сферы деятельности организации и учитывать особенности отдельных ее структурных подразделений. При этом необходимо принимать во внимание то обстоятельство, что любые количественные наукометрические показатели не могут являться абсолютным критерием оценки деятельности каждого работника, а представляют только начальную оценку, подлежащую дальнейшему анализу экспертами.

В качестве такой системы, на средствах анализа данных которой иллюстрируются результаты исследований, представленные в настоящей статье, рассмотрим информационно-аналитическую систему (ИАС) "ИСТИНА" [1, 2], активно используемую в МГУ им. М. В. Ломоносова для подготовки принятия управленческих решений (далее по тексту ИАС "ИСТИНА" или система). С учетом масштабов МГУ и относительно большого времени эксплуатации системы эти данные могут адекватно отражать результаты применения предлагаемых подходов в других организациях в сфере науки и образования.

Одним из важнейших показателей, которые необходимо собирать для оценки научной деятельности работников организации, является число публикаций, а также их цитируемость, распределение публикаций по журналам и темам. Кроме того, необходимо учитывать следующие показатели: участие работников в научных проектах и конференциях; получение ими свидетельств о защите интеллектуальной собственности; педагогическую деятельность; представление к премиям и наградам; участие в работе диссертационных советов и др. [3]. Сбор и верификация, агрегирование и анализ отмеченных данных позволяют адекватно готовить аналитические материалы для принятия управленческих решений, проводить автоматизированный расчет рейтинговых показателей отдельных работников, оценивать работу подразделений и оперативно, без больших затрат готовить различного рода отчеты. Однако при планировании и проведении работ по сбору и анализу показателей о научной и педагогической деятельности сотрудников необходимо принимать во внимание наличие обратных связей между системой анализа и измеряемыми показателями. Следует учитывать, что, как правило, проведение измерений характеристик достаточно сложного объекта начинает оказывать влияние на сам объект. В качестве аналогий можно привести факты, что при измерении термометром

температуры газа происходит теплообмен газа с термометром, а при измерении скорости жидкости перед точкой, в которую помещен датчик, возникает зона торможения и происходит образование вихрей. Введение механизмов измерения показателей оценки научной деятельности также оказывает влияние на структуру измеряемых показателей. Проиллюстрируем это утверждение на примере числа соавторов публикуемых статей. В табл. 1 приведены данные за десять лет о среднем числе соавторов в статьях, зарегистрированных в ИАС "ИСТИНА", с разбивкой всех статей по годам, а также по статьям без учета коллабораций (больших устойчивых групп соавторов, которые публикуются вместе, например, ATLAS COLLABORATION, OPERA COLLABORATION). Из приведенных в табл. 1 данных можно сделать вывод, что с момента начала внедрения механизмов учета публикаций в 2011 г. наблюдается устойчивый рост среднего числа соавторов.

Подобный рост объясняется стремлением работников улучшить свои показатели качества работы и тем самым получить более высокий рейтинговый статус в организации. В данном случае увеличение числа соавторов используется для увеличения числа статей у сотрудников подразделения и для повышения показателя цитируемости. Подобные преднамеренные искажения объективных показателей в области наукометрии неизбежны при внедрении систем автоматизированной оценки деятельности работников. Вместе с тем при использовании подобных систем для эффективного управления крупной научной организацией или вузом возникает необходимость устранения непреднамеренных ошибок и искажений измеряемых показателей, которые являются результатом некорректного ввода данных в систему. Например, неправильное указание авторов при вводе данных о статьях, книгах и другой научной продукции значительно влияет на качество собираемых наукометрических данных. В связи с этим возникает необходимость автоматизации процесса определения авторов по фамилии и инициалам при вводе библиографических сведений о публикации.

Таблица 1

Среднее число соавторов статей

Год	Среднее число соавторов	Рост, %	Среднее число соавторов без учета коллабораций	Рост без учета коллабораций, %
До 2006	2,724	100	2,708	100
2006	2,791	102	2,780	103
2007	2,803	103	2,778	103
2008	2,739	101	2,714	100
2009	2,761	101	2,725	101
2010	2,760	101	2,726	101
2011	2,847	105	2,791	103
2012	3,017	111	2,942	109
2013	3,023	111	2,961	109
2014	3,013	111	2,954	109
2015	3,081	113	3,002	111
2016	3,194	117	3,125	115
2017	3,453	127	3,256	120

Распознавание авторов по библиографическим данным

Представляется очевидным тот факт, что только ученый или педагог может точно перечислить все результаты своей деятельности. Поэтому для достижения максимально полного состава таких данных необходимо предоставить конечным пользователям наукометрической системы возможность ввода сведений о результатах своей деятельности. При этом если какой-либо результат деятельности был получен в соавторстве, то сведения о нем должны быть введены в систему один раз, независимо от числа соавторов. Такой подход позволяет не только экономить человеческий ресурс, но и быстрее, особенно на начальном этапе становления системы, пополнять коллекцию данных. Он позволяет устранить ненужное дублирование и возникающие в связи с этим дополнительные ошибки при вводе данных.

Одной из задач, которые в связи с этим возникают в ходе разработки механизмов автоматизации процессов сбора библиографических данных о публикациях в больших наукометрических системах, является надежное определение (идентификация) автора по указанным в публикации фамилии и инициалам. Важным фактором в этом случае является то, насколько распространена фамилия автора публикации. Если для редко встречающихся фамилий поиск соответствия между указанной в статье фамилией и зарегистрированным в системе автором этот вопрос решается простым поиском совпадения строк, то для распространенных фамилий необходимо проводить более сложный анализ. Например, среди 90 тысяч авторов ИАС "ИСТИНА", являющихся сотрудниками МГУ им. М. В. Ломоносова, около тысячи имеют фамилии "Иванов", "Кузнецов", "Смирнов", "Петров", "Попов", а 50 тысяч авторов имеют более десяти однофамильцев. Кроме того, следует учитывать большое количество соавторов, которые не являются сотрудниками МГУ, но также должны правильно распознаваться системой при проведении автоматизированного разбора библиографических данных, добавляемых авторами статей. Отмеченный факт является важным, поскольку в современных научных исследованиях большое число проектов выполняется совместно работниками нескольких

организаций, а доля сотрудников МГУ среди полного списка авторов статей, загруженных в ИАС "ИСТИНА", составляет менее 30 %.

Вследствие описанных выше причин использования только фамилии и инициалов для определения автора статьи оказывается недостаточно. При вводе пользователем информации о статье после внесения библиографической информации необходимо правильно указать автора для каждой фамилии из списка соавторов статьи. Если система предлагает пользователю всех однофамильцев, то выбор проводится из слишком большого числа возможных вариантов. Многие пользователи затрудняются сделать правильный выбор, загружая в систему статьи с неправильными или незаполненными данными об авторах статей. Такие ошибки оказывают существенное влияние на корректность расчета наукометрических показателей (число статей в журналах из списка ВАК, RSCI, Web Of Science, Scopus, цитируемость и др.) как по отдельным сотрудникам, так и по организации в целом.

Точность решения задачи определения автора статьи по фамилии и инициалам, указанным в ее библиографическом описании, можно увеличивать на основе использования дополнительной информации об устойчивых группах соавторов. Один из таких методов, основанный на поиске максимально связанных подграфов в графе соавторства, описан в работах [4, 5]. Этот метод используется в ИАС "ИСТИНА" в настоящее время, однако он имеет ряд недостатков. Во-первых, этот метод имеет достаточно большую вычислительную сложность, во-вторых, при определении возможных авторов статьи не используется информация об авторизованном пользователе, который вводит информацию о статье. Последний аспект особенно важен, поскольку именно авторы статей наиболее заинтересованы в своевременном и правильном добавлении их работ в наукометрическую систему. Подтверждением тому является тот факт, что согласно проведенным статистическим оценкам по данным из ИАС "ИСТИНА" (данные приведены в табл. 2), более 93 % публикаций вносится одним из соавторов работ. При этом только менее 7 % работ вносят пользователи системы, не являющиеся авторами статей (секретари кафедр и лабораторий, а также другие лица, которым делегированы соответствующие права). В связи с этим информацию из учетной записи зарегистрированного пользователя необходимо использовать для формирования более точных механизмов распознавания авторов.

Таблица 2

Доля статей, вводимых одним из соавторов, в зависимости от числа соавторов в статье

Число соавторов статьи	Статей, тыс.	Статей, введенных авторами, тыс.	Статей, введенных авторами, %
1	246	228	93
2	113	107	94
3	84	79	94
4	58	54	94
5	37	35	94
6	24	23	94
7	14	14	94
8	9	8	94
9	6	5	94
Более 9	13	12	93

Для более точного решения задачи определения авторов по библиографическому описанию статьи требовалось построить алгоритм, который на основе библиографического описания и информации о пользователе, осуществляющем ввод данных о статье, позволяет автоматически сформировать набор идентификаторов авторов, которые адекватны библиографическому описанию. Разработанный авторами этой публикации алгоритм на первом этапе выделяет список возможных авторов для каждой фамилии, упоминающейся в библиографическом описании статьи.

Первый этап алгоритма состоит из следующих шагов.

Шаг 1. Поиск указанной в библиографических данных фамилии и инициалов среди зарегистрированных в системе фамилий авторов. При этом для каждого автора учитываются все варианты написания его фамилии и инициалов, встречавшиеся ранее. Такой анализ необходим для работы со статьями, изданными на других языках. Для фамилий требуется точное совпадение, а для инициалов — совпадение наиболее короткой строки с префиксом длинной строки. Таким образом, для записи "Иванов И И" будут добавлены в список возможных соавторов "Иванов Иван Иванович", "Иванов Иван Ив", но не будет добавлен вариант "Иванов Иван Петрович".

Шаг 2. Сравнение записей в полученном списке возможных авторов с данными о пользователе, который осуществляет ввод данных о статье.

Шаг 3. Если среди возможных авторов встречается авторизованный пользователь, осуществляющий ввод данных о статье, то он считается определенным, и остальные возможные авторы из списка для этой фамилии удаляются.

Результатом работы первого этапа алгоритма является список возможных уникальных ключей авторов, известных системе, для каждой фамилии автора из библиографического описания статьи. Программный компонент, реализующий алгоритм на первом этапе, выполнен в виде модуля в СУБД Oracle и может использоваться в SQL-запросах. Например, "select column_value IRID from table(get_man_id_by_fio_s('Иванов И И',12454))".

На втором этапе алгоритма осуществляется выделение наиболее правдоподобного уникального ключа автора для каждой фамилии. В качестве входных данных второй этап алгоритма принимает полученные на первом этапе списки возможных уникальных ключей авторов для каждого соавтора статьи. Кроме того, в качестве входных данных используется граф соавторства, вершинами которого являются известные системе авторы, а ребра имеют вес, равный числу общих публикаций двух авторов. Второй этап алгоритма состоит из следующих шагов.

Шаг 1. Сортировка всех соавторов статьи по числу вариантов возможных уникальных ключей. В начале списка располагаются соавторы, для фамилий которых на первом этапе подобрано небольшое число возможных вариантов. Соавтор, который определен на текущий момент как авторизованный пользователь, помещается на первое место списка.

Шаг 2. Если в начале списка оказались соавторы с единственным возможным уникальным ключом, которые считаются "распознанными", то для каждого из соавторов с несколькими возможными вариантами уникальных ключей выбираются ребра связи с "распознанными" первичными ключами. После этого выбирается уникальный ключ, имеющий ребро связи с наибольшим весом. Выбранный уникальный ключ считается "распознанным".

Шаг 3. Повторение шага 2, пока в списке соавторов остаются "нераспознанные" ключи или пока в результате выполнения шага 2 не перестают появляться новые "распознанные" ключи.

Шаг 4. Если не все соавторы распознаны, то среди этих соавторов определяется ребро с наибольшим весом для каждой пары вариантов ключей.

Тестирование алгоритма проводилось на графе соавторства, имевшем около 226 тыс. вершин (авторов) и 5 млн ребер. Для построения графа соавторства использовалась информация из статей, тезисов, книг и проектов. Вес ребра определяется как число общих работ у заданных двух авторов. Следует отметить, что большинство ребер в анализируемом графе имеет вес 1. Данные о весах ребер используемого графа приведены ниже.

Вес ребра	**Число ребер**
1	3 506 738
2	808 090
3	270 726
4	119 922
5	69 254
6	44 388
7	31 652
8	23 418
9	18 116
10	15 086
Более 10	104 926

В ходе тестирования было обработано 540 тыс. статей. Совпадение результатов расчета алгоритма с реальными данными составило 520 тыс. записей:

Результат разбора	**Число записей**
Автор распознан и совпадает с автором, указанным в системе пользователем	503 824
Автор не распознан и не указан в системе пользователем	18 807
Автор распознан, но не указан в системе пользователем	47 244
Автор распознан, но не совпадает с автором, указанным пользователем	19 337
Автор не распознан, но указан в системе пользователем	3256

Следует отметить причины того, что для значительной части статей автор распознан, но не указан пользователем. С одной стороны, это может быть результат плохого качества автоматического определения автора при вводе данных о статье в предыдущей версии системы. С другой стороны, это может быть связано с отсутствием заинтересованности пользователя в правильном выборе соавторов вручную. Как следствие отмеченных причин, значительная часть статей введена в систему с единственным автором из всего списка соавторов, а именно — пользователем, который вносил данные о статье в систему. Использование нового алгоритма позволит предлагать пользователю более адекватный (точный) распознанный список авторов для новых вносимых в систему статей, а также определить соавторов для старых статей, в которых при загрузке были указаны не все авторы.

Скорость обработки тестируемой реализации представленного алгоритма составила более 10 000 тыс. статей за 40 с, или 0,004 с на одну статью, что позволяет давать рекомендации пользователю по выбору авторов при вводе статьи без задержек. Программный компонент, реализующий алгоритм на втором этапе, также выполнен в виде модуля в СУБД Oracle и может использоваться в SQL-запросах. Например, "SELECT * FROM TABLE(get_man_by_name(ARRAY_VARCHAR2('КОЗИЦЫН А С', 'АФОНИН С А', 'ЗАНЧУРИН М А', 'КОРШУНОВ А А'),0)) ".

Дальнейшее улучшение полученных на настоящее время результатов возможно за счет использования двудольного графа (пользователь — автор), позволяющего учесть факт регулярного ввода информации о публикации нескольких авторов одним пользователем. Например, в случае ввода информации обо всех сотрудниках кафедры ученым секретарем кафедры (лаборатории) или другим лицом, которому эти полномочия делегированы. Перспективным также является использование весовой функции при определении возможных авторов на первом шаге алгоритма. Однако некоторые ошибки являются неустранимыми, например, в случае смены фамилии автором или неправильного ее указания.

Недостатком предлагаемого алгоритма является отсутствие возможности его использования для статей, написанных одним автором. Однако этот недостаток частично компенсируется тем обстоятельством, что данные о таких статьях авторы обычно

вводят самостоятельно и авторство однозначно определяется по авторизованному пользователю.

Разработанный алгоритм позволяет при вводе данных в систему подсказывать пользователю правильный вариант выбора автора для каждой фамилии из библиографических данных статьи. Он повышает точность вносимых данных, а также позволяет анализировать внесенные данные на наличие потенциальных ошибок.

Выделение тематических направлений исследований авторов

Автоматическое определение областей интересов пользователей позволяет решить две практически важные задачи: поиск "похожих" по тематике авторов и поиск авторов по описанию темы, например, по ключевым словам. Кроме того, такой подход позволяет рекомендовать пользователям конференции и журналы на основе статистики публикаций авторов схожей тематики.

При проведении классификации необходимо учитывать то обстоятельство, что авторы могут иметь публикации в нескольких тематических направлениях. Интересы людей могут со временем меняться, автор может одновременно с успехом работать в нескольких научных областях. Следует также отметить наличие отдельного класса авторов, которых можно определить как "руководители". Для таких авторов, как правило, характерно наличие соавторства в статьях, проектах и докладах сразу по многим тематическим направлениям.

Наиболее простым способом классификации по областям научной деятельности является классификация авторов по тематической направленности журналов, в которых авторы печатают свои работы. В настоящее время в ИАС "ИСТИНА", например, используется несколько различных тематических классификаторов журналов: ГРНТИ (1800 рубрик), Scopus (310 рубрик), Web of Science (230 рубрик) и медицинский классификатор Medline MeSH (124 рубрики) [6]. Недостатком такого метода является отсутствие возможности точного определения тематики статьи и тематических направлений ее авторов, поскольку тематика журнала, как правило, формулируется слишком широко. Таким же недостатком обладает тематическая классификация авторов на основе описания своих достижений с указанием этих рубрик.

В экспериментах, которые проводили с целью анализа и устранения отмеченных недостатков, использовали векторную модель описания тематических интересов авторов, а тематическую близость авторов оценивали как косинус угла между их векторами. В качестве координат векторов использовали четыре варианта характеристик:

- число статей автора, которые напечатаны в журнале с заданной рубрикой;
- число журналов с данной рубрикой, в которых напечатаны статьи автора;
- число статей автора, которые напечатаны в журнале с заданной рубрикой, деленное на общее число статей в журналах с данной рубрикой;
- число журналов с данной рубрикой, в которых напечатаны статьи автора, деленное на общее число журналов с данной рубрикой.

Во всех случаях точность определения "похожих" авторов была очень низкой и не превышала 20 %. Это объясняется двумя факторами. Во-первых, многие журналы используют очень широкий набор классификаторов. Например, журнал "Информационное общество" имеет 19 общих рубрик, в том числе "Социология", "Философия", "Государство и право. Юридические науки", "Информатика", "Кибернетика". Публикация статьи в таком журнале создает много тематических неправильных связей между авторами. Во-вторых, даже в одном журнале публикуются авторы разных тематических направлений.

Повышение качества определения "похожих" авторов возможно за счет предварительной оценки количества рубрик разных тематических направлений. Журналы с узкой специализацией имеют, как правило, не более четырех основных тематических направлений (не связанные между собой позиции тематического рубрикатора ГРНТИ, Scopus или Web of Science). Ниже показано распределение числа журналов по количеству основных тематических направлений, зарегистрированных в системе "ИСТИНА".

Число тематических направлений	Число журналов
1	7158
2	5359
3	4047
4	3296
5	2180
6	1540
Более 6	2114

Как видно из представленных данных, большая часть журналов (77 %) имеет только четыре основных направления. Журналы с большим числом тематических направлений исключались из расчета тематической близости авторов. Кроме того, при проведении расчетов вес тематических направлений каждого журнала делили на число общих тематических направлений в этом журнале. Таким образом, узкоспециализированные журналы оказывали большее влияние на определение тематических интересов автора.

С учетом изложенных выше фактов авторами данной статьи был разработан алгоритм автоматического определения близости авторов статей по тематическим направлениям. Разработанный алгоритм определения тематической близости авторов состоит из следующих шагов.

Шаг 1. Выделение списка журналов с числом рубрик не более четырех.

Шаг 2. Расчет веса тематических направлений каждого журнала.

Шаг 3. Расчет вектора тематических направлений для каждого автора как число журналов с данной рубрикой, в которых напечатаны статьи автора, деленное на общее число журналов с данной рубрикой.

Шаг 4. Исключение из расчета пар авторов, векторы которых ортогональны (не имеют общих тематических направлений).

Шаг 5. Сортировка пар авторов по значению косинуса угла между соответствующими им векторами.

Шаг 6. Отбор для каждого автора N наилучших пар (при оценке результатов работы использовалось $N = 30$).

Исключение из расчета журналов с широким набором классификаторов и учет веса тематических рубрик позволили существенно повысить точность получаемых результатов. Точность определения пар авторов, имеющих одинаковые общие тематические направления, составила 90 %. Правильность результатов определения на тестовой выборке проверяли вручную по списку статей каждого автора.

Основным преимуществом предлагаемого подхода является небольшая вычислительная сложность по сравнению с алгоритмами тематического анализа полнотекстовых документов, а также высокая скорость работы на больших объемах библиографических данных. В первую очередь это объясняется тем фактом, что описанный алгоритм использует только связи между объектами (статьями, авторами, журналами), доступ к которым осуществляется с использованием индекса. Алгоритм не требует использования методов морфологической и синтаксической обработки полнотекстовых документов. Вместе с тем выделить точное тематическое направление он не способен в силу описанных выше причин. Как следствие, задача поиска похожих авторов по тематике не может быть решена в рамках подобного подхода. Предлагаемый алгоритм может быть использован только для выделения общих тематических направлений и подбора журналов и конференций, соответствующих тематическим интересам каждого пользователя системы.

Одним из альтернативных подходов к классификации тематических интересов авторов является использование методов анализа текстовой информации из названий, аннотаций и ключевых слов, которые авторы указывают в описании статей, проектов, достижений, докладов на семинарах и выступлениях в СМИ. Использование полных текстов статей затруднено, поскольку издательства многих журналов не размещают полные версии статей в сети Интернет. Из перечисленных выше методов анализа тематической близости наиболее эффективным является использование набора ключевых слов [7], поскольку авторы самостоятельно отбирают слова, наиболее точно описывающие тематику статьи. Однако именно необходимость ручной работы является существенным препятствием для широкого использования такого подхода. От авторов требуется аккуратное заполнение ключевых слов на этапе ввода данных о статье в систему. Например, из 600 000 статей, зарегистрированных в системе "ИСТИНА", ключевые слова указаны менее чем для 2000. Аннотации заполнены для 100 000 статей. Таким образом, дальнейшее уточнение тематических интересов авторов возможно с использованием методов анализа текстов для аннотаций и названий статей.

Заключение

Решение описанных в статье задач идентификации авторов и определения близости тематических интересов пользователей с использованием методов анализа больших объемов библиографических данных позволяет создавать удобные инструментальные средства для сбора, верификации и обработки информации о научной продукции крупных организаций, а также предоставляет расширенные возможности по поиску, агрегации информации и составлению аналитических отчетов при принятии управленческих решений. Использование нового алгоритма определения авторов по библиографическим данным статьи позволяет повысить качество вносимых пользователем данных и увеличить точность последующего расчета наукометрических показателей. Автоматическое определение тем научных исследований позволяет оперативно готовить аналитические материалы для принятия управленческих решений в разрезе отдельных тематических направлений, а также может использоваться для предоставления пользователям наукометрической системы удобного интерфейса по поиску журналов и конференций.

Список литературы

1. **Садовничий В. А., Васенин В. А., Афонин С. А.** и др. Информационная система "ИСТИНА" как big data — инструментарий в области управления на основе анализа наукометрических данных // Материалы Всероссийской конференции с международным участием "Знания-Онтологии-Теории" (ЗОНТ-2015), 6—8 октября 2015 г. Т. 1. Новосибирск: Институт математики им. С. Л. Соболева СО РАН, 2015. С. 115—123.

2. **Афонин С. А., Бахтин А. В., Бухонов В. Ю.** и др. Интеллектуальная система тематического исследования научно-технической информации (ИСТИНА) / Под ред. В. А. Садовничего. М.: Изд-во МГУ, 2014. 262 с.

3. **Васенин В. А., Афонин С. А., Козицын А. С.** Автоматизированная система тематического анализа информации// Информационные технологии. Приложение. 2009. № 4. 32 с.

4. **Афонин С. А., Гаспарянц А. Э.** Автоматическое построение функции оценки качества в задаче разрешения неоднозначности имен авторов научных публикаций // Программная инженерия. 2015. № 10. С. 31—37.

5. **Афонин С. А., Гаспарянц А. Э.** Разрешение неоднозначности авторства публикаций при автоматической обработке библиографических данных // Программная инженерия. 2014. № 1. С. 25—28.

6. **Медицинские** предметные рубрики. URL: https://ru.wikipedia.org/wiki/Medical_Subject_Headings

7. **Афонин С. А., Лунев К. В.** Выявление тематических направлений в коллекции наборов ключевых слов // Программная инженерия. 2015. № 2. С. 29—39.

The Resolution of Ambiguities in the Identification of Authors of the Publication with the Use of Co-Authors' Graphs in Large Collections of Bibliographic Data

A. S. Kozitsin, alexanderkz@mail.ru, **S. A. Afonin,** serg@msu.ru, Lomonosov Moscow State University, Moscow, 117223, Russian Federation

Corresponding author:

Kozitsin Alexander S., Researcher, Lomonosov Moscow State University, Moscow, 117223, Russian Federation
E-mail: alexanderkz@mail.ru

Received on July 18, 2017
Accepted on August 02, 2017

This article addresses problems related to automated processing of bibliographic data in scientometric systems by means of statistical analysis of large collections of such data. Experimental results on authors identification in bibliographic data are based on the data set presented in ISTINA, a scientometric information system developed and deployed at Moscow State University. The new algorithm for authors identification, presented in this paper, shows 95 % accuracy on the considered data set. Utilization of this algorithm improves the quality of data entered into the system, thus leading to a more reliable scientometric characteristics of individual researchers and administrative units. The paper also discusses possible approaches to some related practically important problems, such as thematic search and classification of publications using coauthoring graph and thematic classification of scientific journals. Automatic discovery of researchers topics of interest allows for on-demand generation of various documents suitable for decision making in specific research areas and could be useful for supplying system users with information on relevant journals or upcoming scientific events.

Keywords: *succometrics, detection of regularities, bibliographic data, graph, citation, author, thematic analysis*

For citation:

Kozitsin A. S., Afonin S. A. The Resolution of Ambiguities in the Identification of Authors of the Publication with the Use of Co-Authors' Graphs in Large Collections of Bibliographic Data*, Programmnaya Ingeneria*, 2017, vol. 8, no. 12, pp. 556—562.

DOI: 10.17587/prin.8.556-562

References

1. **Sadovnichij V. A., Vasenin V. A., Afonin S. A., Kozitsin A. S., Golomazov D. D.** Informacionnaja sistema "ISTINA" kak big data — instrumentarij v oblasti upravlenija na osnove analiza naukometricheskih dannyh. (The information system "ISTINA" as big data is a tool in the field of management based on the analysis of scientometric data), *Materialy Vserossijskoj konferencii s mezhdunarodnym uchastiem "Znanija-Ontologii-Teorii" (ZONT-2015)*, 6—8 October 2015, vol. 1, Novosibirsk, 2015, pp. 115—123 (in Russian).

2. **Afonin S. A., Bahtin A. V., Buhonov V. Yu., Vasenin V. A., Gankin G. M., Gaspariants A. E., Golomazov D. D., Itkes A. A., Kozitsin A. S., Tumajkin I. N., Shapchenko K. A.** *Intellektual'naja sistema tematicheskogo issledovanija nauchno-tehnicheskoj informacii (ISTINA)* (Intellectual system of case research of scientific and technical information) / Eds. V. A. Sadovnichiy, Moscow, Moscow State University, 2014, 262 p. (in Russian).

3. **Vasenin V. A., Afonin S. A., Kozitsin A. S.** Avtomatizirovannaja sistema tematicheskogo analiza informacii (Automated system of thematic information analysis), *Informacionnye tehnologii, Supplement,* 2009, no. 4, 32 p. (in Russian).

4. **Afonin S. A., Gaspariants A. E.** Avtomaticheskoe postroenie funkcii ocenki kachestva v zadache razreshenija neodnoznachnosti imen avtorov nauchnyh publikacij (Automatic construction of the quality assessment function in the problem of resolving the ambiguity of names of authors of scientific publications), *Programmnaya Ingeneria,* 2015, no. 10, pp. 31—37 (in Russian).

5. **Afonin S. A., Gaspariants A. E.** Razreshenie neodnoznachnosti avtorstva publikacij pri avtomaticheskoj obrabotke bibliograficheskih dannyh. (Resolution of the ambiguity of the authorship of publications in the automatic processing of bibliographic data), *Programmnaya Ingeneria,* 2014, no. 1, pp. 25—29 (in Russian).

6. **Medical**_Subject_Headings, available at: https://ru.wikipedia.org/wiki/Medical_Subject_Headings

7. **Afonin S. A., Lunev K. V.** Vyjavlenie tematicheskih napravlenij v kollekcii naborov kljuchevyh slov (Identify the thematic areas in the collection of keyword sets. Software engineering), *Programmnaya Ingeneria,* 2015, no. 2, pp. 29—39 (in Russian).