УДК 001:004

DOI: 10.15827/0236-235X.122.275-283

Дата подачи статьи: 12.10.17

2018. Т. 31. № 2. С. 275–283

РЕКОМЕНДАТЕЛЬНАЯ СИСТЕМА НА ОСНОВЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА НАУКОМЕТРИЧЕСКОГО ПРОФИЛЯ ИССЛЕДОВАТЕЛЯ

Д.В. Валько ¹, магистрант, *ell.science@mail.ru*

¹ Южно-Уральский государственный университет

(национальный исследовательский университет), просп. Ленина, 76, г. Челябинск, 454080, Россия

Сегодня научные результаты каждого исследователя могут быть представлены в различных наукометрических базах и системах. Нередко они имеют определенную популярность не в силу своей актуальности, а лишь благодаря глобальной доступности. В то же время качественные научные результаты могут оказаться вне поля зрения научного сообщества просто потому, что не размещены в той или иной популярной наукометрической системе. С точки зрения научного менеджмента такая ситуация девальвирует исследователя безотносительно качества и актуальности достигнутых им научных результатов.

По мнению автора, формирование механизма выработки рекомендаций для отдельных исследователей, научных коллективов и их руководителей на всех уровнях менеджмента позволило бы обратить внимание на перспективные научные результаты и обоснованно аккумулировать необходимые ресурсы для включения таких результатов в популярные наукометрические системы.

Разработка инструментария, оперирующего большими объемами открытых данных современных наукометрических систем, не может обойтись без методов интеллектуального анализа. В настоящей работе показано, что на основе алгоритма интеллектуального анализа взаимосвязей (Argiori), адаптированного к наукометрическим данным в системе Scopus, могут быть сформулированы некоторые наборы ассоциативных правил, пригодные для прогнозирования вероятных будущих научных результатов, и разработаны автоматизированные рекомендации по улучшению публикационной активности. Разработаны и предложены методический инструментарий анализа публикационной активности с применением интеллектуального анализа наукометрических данных, а также модульная архитектура и прототип программной системы, позволяющей на основе интеллектуального анализа наукометрических данных формировать публикационный рейтинг и индивидуальные рекомендации по улучшению публикационной активности автора.

В статье показаны некоторые результаты экспериментов по оценке публикационного потенциала исследователей, аффилированных с Южно-Уральским государственным университетом.

Ключевые слова: наукометрия, рекомендательная система, интеллектуальный анализ данных, публикационная активность.

В мировой практике функционирования института науки в качестве основы оценки научной мысли и продуктивности научной деятельности используются два подхода – экспертный (качественный) и наукометрический (количественный) [1]. В последние годы в качестве инструмента оценки эффективности деятельности российских ученых, исследовательских организаций, отечественной науки в целом стали активно использоваться данные об уровне и числе публикаций, числе и качестве цитирований, представленных разнообразными наукометрическими показателями. Публикационные и цитатные наукометрические показатели рассматриваются как целевые индикаторы состояния науки в «Стратегии инновационного развития Российской Федерации на период до 2020 года» и в Указе Президента РФ от 7 мая 2012 года.

В области наукометрии и инфометрии в целом наработан обширный базис математических методов для исследования динамики научных исследований и публикационной активности. В международной практике известны аналитические методы и показатели, наиболее полно отвечающие задаче индивидуальной наукометрической оценки автора и т.п. Совершенствуется аналитический инструментарий (InCites, SciVal etc.) в программных системах международных реферативно-библиографических БД Scopus, Web of Science и других.

Вместе с тем недостаточное внимание как в теории, так и в практике уделено поддержке принятия управленческих решений в сфере менеджмента научной деятельности: недостаточно изучены вопросы формирования и управления публикационным потенциалом отдельных исследователей и исследовательских коллективов; не проработаны вопросы работы с большими наукометрическими данными с применением методов интеллектуального анализа; отсутствуют программные разработки в области поддержки принятия решения и в разработке индивидуальных рекомендаций по улучшению публикационной активности, наукометрической результативности и реализации исследовательского и публикационного потенциалов.

На современном этапе развития информационных технологий важную роль приобретают интеллектуальные рекомендательные системы. Рекомендательные системы – это программы, которые пытаются предсказать, какие объекты (фильмы, музыка, книги, публикации, веб-сайты) будут интересны пользователю, имея определенную информацию об их профиле [2]. Разумеется, и наукометрический профиль исследователя или исследовательской организации в целом в той или иной наукометрической БД вполне отвечает необходимым условиям разработки рекомендаций.

Таким образом, в настоящее время целесообразна разработка программной системы, призванной помочь руководителям исследовательских групп, а также индивидуальным исследователям принимать решения по улучшению публикационной активности, наукометрической результативности, определять слабые и сильные стороны, направления и конкретные решения в виде автоматизированных рекомендаций по повышению величины и репрезентативности наукометрических показателей. Рассматриваемая тема приобретает особую актуальность для российских университетов, включенных в государственную программу «Проект 5-100», направленную на повышение конкурентоспособности университетов среди ведущих мировых научно-образовательных центров.

Цель и задачи

Сегодня исследователю для решения задачи наиболее эффективного опубликования результатов исследований в числе прочего необходимо ответить на ряд вопросов:

- в каком виде и объеме опубликовать новые научные результаты;
- какие издания могут быть рекомендованы для опубликования;
- как выбрать авторитетный научный журнал;
- как на эти вопросы отвечают наиболее успешные коллеги.

Очевидно, что не всякий исследователь обладает свободным временем и компетенциями, чтобы ответить на эти вопросы, при этом от ответов на них во многом зависят его репутация и доступность результатов исследования для мирового научного сообщества. Данная проблема особенно обостряется на уровне руководства научными коллективами, лабораториями и исследовательскими организациями в целом, когда руководителю необходимо принимать управленческие решения по финансированию тех или иных исследований, обеспечению кадровыми и иными ресурсами. Руководитель нуждается в аналитической информации относительно перспективности тех или иных исследований, исследовательского и публикационного потенциалов тех или иных исследователей.

Целью данной работы является создание программной системы поддержки принятия решений в сфере менеджмента научных исследований, отличающейся применением методов интеллектуального анализа наукометрических данных и позволяющей формировать индивидуальные рекомендации по улучшению публикационной активности автора. Данная цель определила задачи:

- разработать методический инструментарий оценки публикационной активности с применением интеллектуального анализа наукометрических данных;
- спроектировать и реализовать программную систему, позволяющую работать с наукометрическими данными, формировать публикационный

рейтинг и индивидуальные рекомендации по улучшению публикационной активности автора.

Обзор программных решений в данной области

В настоящее время в рассматриваемой области преимущественно применяются встроенные аналитические приложения к реферативным БД. Каждое такое приложение является проприетарным (InCites, SciVal etc.) и работает только со своей наукометрической базой, в то время как обстоятельный анализ требует сопоставления показателей из РИНЦ, Scopus, Web of Science и внутриуниверситетской информационной системы, в том числе базы портфолио и т.п. Достижения в области информационных технологий ориентируют на открытость наукометрических аналитических приложений (например Google Scholar, Microsoft Academic Search), и данные сервисы наращивают потенциал [3]. Попытка интеграции нескольких источников данных на принципах открытости принята, например, в проекте «Карта российской науки» (<https://mapofscience.ru/>), но ее нельзя признать состоятельной, поскольку данный проект предполагает одновременную работу только с одним источником данных (в режиме переключения между источниками), а также не предоставляет серьезных аналитических возможностей.

Рассмотрим некоторые актуальные зарубежные прикладные разработки в данной области.

В работе [4] предложены алгоритм автоматизации процесса анализа цитирования соавторов и структура соответствующей БД. Исходные данные предварительно располагаются в специальной БД WebCitation. В основе разработанного программного инструмента лежат методы кластерного анализа, с помощью которых формируется карта кластеров связанных авторов по той или иной тематике.

В работе [5] представлен программный инструмент анализа библиометрических и наукометрических данных, который на основе расчета ряда стандартных показателей предоставляет нормированные результаты для исследователей, научно-исследовательских групп, исследовательских центров, факультетов или университета в целом. Инструмент разработан на Java и использует централизованную университетскую БД MySQL, в которую администратором подгружаются необходимые сведения об авторах, статьях и журналах.

Программный продукт Publish or Perish Bibliometric Data Tools, предложенный в [6], интегрирует набор утилит, позволяющих предобрабатывать исходные данные Google Scholar и Microsoft Academic Search, а затем на основе их анализа рассчитывать ряд показателей (www.harzing.com/resources/publish-or-perish) (см. http://www.swsys.ru/uploaded/image/2018_2/2018-2-dop/1.jpg):

- общее количество публикаций и общее количество цитирований;

- среднее число цитат на публикацию, на автора, число цитат в год и другое;
- базисный h -индекс и его основные модификации (в том числе среднегодовой прирост индекса).

Среди отечественных научных и прикладных разработок в данной области выделяется интеллектуальная измерительная система «Эйдос», адаптированная к анализу данных РИНЦ [7]. На основе согласования мнений экспертов и статистического анализа данных система формирует классификатор (см. http://www.swsys.ru/uploaded/image/2018_2/2018-2-dop/2.jpg) и определяет статус отдельного автора.

В работе [8] представлена модульная архитектура системы анализа публикационной активности сотрудников научно-образовательной организации, включающая модуль для управления пространством сотрудников научно-образовательной организации, модуль для управления областью публикационных объектов и модуль для вычислений и анализа. К сожалению, в данной и подобных работах, например в [9], представлено лишь описание концептуальных блоков и их общего взаимодействия в рамках системы, не ориентированное на какую-либо конкретную реализацию или варианты использования.

Вопросы разработки инструментов автоматизации и интеграции научной информации в пространстве разнородных информационных систем, обсуждаемые, например, в [10], обусловили разработку способа автоматизированного сбора метаданных о публикациях в репозитории и некоторые шаблоны обработки данных [11].

Таким образом, среди актуальных отечественных и зарубежных разработок в данной области выделяется интеллектуальная измерительная система «Эйдос», адаптированная к анализу данных РИНЦ и использующая агрегированные экспертные оценки; а также программный продукт Publish or Perish Bibliometric Data Tools, который интегрирует набор утилит, позволяющих предобрабатывать исходные данные Google Scholar и Microsoft Academic Search, а затем на их основе рассчитывать ряд наукометрических показателей. Однако данные программные разработки не могут считаться рекомендательными системами.

В конечном счете, как показал анализ, сегодня отсутствуют программные разработки в области поддержки принятия решения и разработки индивидуальных рекомендаций по улучшению публикационной активности, наукометрической результативности и реализации исследовательского и публикационного потенциалов.

Оценка публикационного потенциала методами интеллектуального анализа данных

Результатом применения методики оценки публикационной активности является публикацион-

ный рейтинг, то есть некоторое упорядочение исследователей с присвоением рангов от 1 (лучший ранг) до N . Полученный рейтинг становится основой оценки публикационного потенциала, положение которой рассмотрим далее.

В общем смысле под *публикационным потенциалом* автора (исследователя) следует понимать его возможности по улучшению публикационной активности, в частности, по повышению наукометрических показателей, без снижения качества. В настоящей работе определение индивидуального публикационного потенциала на основе публикационного рейтинга опирается на квартильную меру. Квартили – значения, которые делят таблицу данных (или ее часть) на четыре группы, содержащие приблизительно равное количество кортежей (значений).

В данном случае принадлежать к первому квартилю (или верхнему, q_1) будут авторы, занимающие верхние 25 % мест в публикационном рейтинге; ко второму квартилю (q_2) – следующие 25 % мест и т.д. Иными словами, в первом квартиле будут находиться 25 % авторов, занимающих лучшие места в публикационном рейтинге, а в четвертом квартиле (q_4) – 25 % занимающих худшие места.

С учетом этого *индивидуальный публикационный потенциал* автора определяется как прогнозная оценка возможностей данного автора по переходу из своего квартиля по рейтингу публикационной активности к лучшему квартилю. Необходимые условия для подобного перехода в тривиальном случае представляют собой такой прирост наукометрических показателей автора, который обеспечил бы существенный подъем в публикационном рейтинге до места, занимаемого последним автором в лучшем квартиле. Поскольку при построении публикационного рейтинга предпочтение отдавалось более молодым авторам (по году первой публикации), определение публикационного потенциала на его основе сохраняет данную особенность. Это, впрочем, согласуется со спецификой жизненного цикла исследователя: предполагается, что более молодой автор располагает большим публикационным потенциалом.

В целях оценки публикационного потенциала автором предлагается адаптация алгоритма интеллектуального анализа взаимосвязей (поиска ассоциативных правил) Apriori [12], модифицированного в [13] к наукометрическим данным (на примере системы Scopus). С помощью данного алгоритма, как будет показано далее, могут быть сформулированы некоторые наборы ассоциативных правил, пригодные для прогнозирования вероятных будущих публикационных результатов автора.

Определение 1. Формальный контекст в пространстве наукометрических данных определяется как $C = (U, S)$, где S – набор атрибутов профилей авторов в системе Scopus; U – множество всех ав-

торов (как объектов). Набор S является синтетическим, то есть формируется на основе заданного соответствия (некоторого отображения, $\Leftrightarrow$) между набором квартилей Q , определенных на совокупности рангов (r^n) для n соответствующих показателей или фактических значений (x') последних по всем атрибутам авторов: $S = \{s \Leftrightarrow q\}$, $s \in S$, $q \in Q$.

Поскольку для любого существующего значения показателя $x' \in X$ или ранга атрибута профиля автора заведомо существует квартиль $q(x')$: $= \{q \in Q / q \Leftrightarrow x'\}$, а для любого автора $u \in U$ заведомо существует набор значений атрибутов профиля $x(u)$: $= \{x' \in X / x' \Leftrightarrow u\}$ (для рангов: $r(u)$: $= \{a_i \in A / a_i \Leftrightarrow u\}$), полагаются некоторые бинарные отношения для наборов $U \times X$ и $X \times S$. Отталкиваясь от определения S , имеем однозначное отношение $U \times S = \{s \in S / s \Leftrightarrow q(x(u))\}$.

Определение 2. Если для данного автора u существует такое $S_u = u \times S \neq \emptyset$, где S_u – набор атрибутов профиля, относящихся к автору u , то данный автор может использоваться в поиске ассоциативных правил.

Пусть D – набор транзакций, где каждая транзакция T – это набор элементов из $I = U \times S$, $T \subseteq I$. Тогда транзакция T содержит N – некоторый набор элементов (*antecedent*) из I , если $N \subset T$. Ассоциативным правилом называется импликация $N \Rightarrow M$, где $N \subset I$, $M \subset I$ и $N \cap M = \emptyset$.

Полагаем, что ассоциативное правило $N \Rightarrow M$ имеет поддержку σ (*support*), если σ (%) транзакций из D содержат $N \cup M$: $\sigma(N \Rightarrow M) = \sigma(N \cup M)$. Правило $N \Rightarrow M$ справедливо с достоверностью δ (*confidence*), если δ (%) транзакций из D , содержащих N , также содержат M (*consequent*): $\delta(N \Rightarrow M) = \sigma(N \cup M) / \sigma(N)$.

В реализации К. Боргелята каждому ассоциативному правилу назначается также некоторое измерение λ (*lift*), такое, что $\lambda(N \Rightarrow M) = \delta(N \Rightarrow M) / \delta(D \Rightarrow M)$. Очевидно, что, чем выше λ , тем выше значимость данного правила, поэтому в дальнейшем будем исключать из рассмотрения правила, у которых $\lambda < 1$, и избирать правило с большим λ при одинаковой сопоставимой достоверности δ .

Таким образом, в соответствии с предпосылками применяемого метода интеллектуального анализа методика оценки публикационного потенциала и алгоритм выявления авторов, имеющих такой потенциал, будут состоять в следующем.

- Для заданного набора авторов строится публикационный рейтинг (на основе некоторой методики оценки публикационной активности).

- На основе полученного публикационного рейтинга набор авторов разбивается на четыре группы – квартили (q_1 , q_2 , q_3 , q_4).

- По каждому атрибуту наукометрического профиля (наукометрическому показателю), участвующему в построении рейтинга, производится

дискретизация, то есть формируется набор синтетических атрибутов, представляющих собой метки вида *attrib_QX*, где *attrib* – имя атрибута, X – признак квартиля, например *h_index_Q1*.

Кроме того, значения атрибута, эквивалентные минимуму (чаще всего для наукометрических показателей это ноль), заведомо приобретают метку *Q4*. Данное коррекционное правило приводит к искажению квартильной структуры по некоторым атрибутам, но обеспечивает верную интерпретацию наукометрического профиля автора.

- Для полученных наборов синтетических атрибутов по каждому автору, представляющих собой транзакцию, выполняется поиск ассоциативных правил. Затем производится выборка устойчивых правил, обнаруживающих переход авторов из собственных квартилей в лучшие по сочетанию атрибутов.

- На основе полученных правил выявляется набор авторов, им соответствующих, то есть имеющих потенциал перехода к лучшему квартилю рейтинга публикационной активности. При этом прогнозной мерой публикационного потенциала автора является мера достоверности подходящего устойчивого правила.

Разумеется, данная методика может быть легко реализована в рамках информационной системы научно-образовательной организации и использована для задач управления научной деятельностью. Полная автоматизация может быть достигнута интеграцией модуля автоматического сбора (выгрузки) наукометрических данных [14] из используемых реферативных баз.

В заключение рассмотрим принципы формирования индивидуальных рекомендаций. Фактически рекомендация для автора представляет собой тезисную интерпретацию результатов наукометрического анализа (в частности, оценки публикационного потенциала). Формирование рекомендации происходит на основе следующих альтернативных шаблонов:

...сфокусироваться на улучшении следующих показателей (в порядке понижения значимости): X , Y , Z ...

...по увеличению X : следует достигнуть уровня, не менее Y , то есть повысить свой показатель на Z ...

...необходимый прирост к следующему отчетному периоду составляет $+F$...

...по X следует сохранить положительную динамику (превышение на Y)...

...рекомендуются следующие журналы для публикации: X , Y , Z (см. рекомендованный список)...

...необходимо опубликовать дополнительно X статей в журналах с импакт-фактором, не менее Y (см. рекомендованный список)...

...рекомендовано рассмотреть возможность коллаборации с лучшими авторами в публикационной группе (см. рекомендованный список)...

Архитектура программной системы

Программная система предназначена для интеллектуального анализа наукометрических данных – некоторой совокупности значений наукометрических показателей публикационной активности для набора авторов. Пользователь программной системы представлен двумя ролями (актерами): автор (преподаватель, исследователь, имеющий публикации), использующий основной функционал программной системы для получения индивидуальных рекомендаций, и руководитель, использующий программную систему для анализа публикационной активности авторского коллектива, а также готовящий программную систему к работе с автором.

Опишем кратко функциональные возможности (варианты использования) программной системы (рис. 1).

1. Формирование публикационного рейтинга. Прецедент, предоставляющий пользователю возможность просмотра таблицы рейтинга авторов, построенной по наукометрическим данным, ранее загруженным в систему (на основе оригинальной методики оценки публикационной активности).

2. Оценка публикационного потенциала. Позволяет пользователю ознакомиться с результатами прогноза возможностей авторов по переходу к лучшему квартилю в публикационном рейтинге (на основе поиска ассоциативных правил).

3. Поиск групп публикационной активности. В рамках данного прецедента выполняется формирование групп авторов, близких по своей публикационной активности, с выделением их типичных представителей (на основе кластеризации с выделением оптимального числа кластеров или с разбиением на заданное число кластеров).

4. Получение индивидуальных рекомендаций. Позволяет пользователю получить набор рекомен-

даций по продвижению в рейтинге публикационной активности и реализации публикационного потенциала (на основе анализа места автора в публикационном рейтинге относительно лучшего, типичного и худшего авторов, а также подбора журналов, в которых опубликованы работы авторов).

5. Подготовка наукометрических данных. Обобщенный прецедент, доступный только руководителю, реализующий программные возможности по подготовке и предобработке наукометрических данных авторов, их загрузке в программную систему и по необходимости обновлению.

6. Настройка аналитических моделей. Выполняется руководителем. Прецедент, реализующий программные возможности по настройке рабочих методик, а также параметров моделей интеллектуального анализа наукометрических данных, в дальнейшем используемых системой для работы.

Программная система состоит из трех основных компонентов, а также использует внешнюю интеллектуальную систему, обрабатывающую данные на платформе KNIME Analytics Platform (www.knime.org/knime-analytics-platform). Основные компоненты предоставляют и используют соответствующие интерфейсы (рис. 2):

- пользовательский интерфейс программной системы и/или интерфейс программирования (компонент Gateway);

- компонент хранения данных (компонент DataSystem);

- базовая вычислительная подсистема (компонент BaseSystem), реализующая взаимодействие с хранилищем данных, некоторые вычисления и обработку результатов работы внешней интеллектуальной системы KNIME Analytics Platform.

Автор в программной системе работает с именнованным набором наукометрических данных, предоставляемых хранилищем наукометрических

профилей (ProfileStore) и аналитических моделей (Model). Пользователь может создавать, удалять и изменять имеющиеся в системе наборы данных и моделей. Набор создается путем отбора данных и моделей из загруженных и подготовленных руководителем (модератором) ранее в зависимости от области исследований и других критериев. На основе имеющихся наборов данных, определенных атрибутами наукометрического профиля автора в системе Scopus, программная система строит публикационный рейтинг.

Аналитическая модель (Model) представляет собой созданный и настроенный модератором ин-

теллектуальный объект, который используется для оценки и прогнозирования принадлежности автора к той или иной публикационной группе и имплементации других методов интеллектуального анализа данных.

Базовая вычислительная подсистема (компонент BaseSystem) предусматривает модульную организацию таким образом, что каждый модуль независимо реализует один из вариантов использования. Структура основных модулей показана на рисунке 3.

Как видно из рисунка, ядром функционального модуля является вычислительный поток (KNIME

workflow), отдельный экземпляр которого обрабатывает данные на платформе KNIME. Из наименования модулей ясно, что *scopus_rating* реализует формирование публикационного рейтинга на основе данных системы Scopus, *scopus_rules* – оценку публикационного потенциала, *scopus_cluster* – поиск групп публикационной активности, а *scopus_preparing* – подготовку и предобработку наукометрических данных. Модульная структура, где каждый модуль автономен и реализует один из методов интеллектуального анализа, позволяет масштабировать систему в условиях распределенных вычислений и применения микросервисов. При этом компонент «шлюз» (gateway) позволяет каждому актору независимо получать результаты интеллектуального анализа и по-своему интерпретировать их для разработки рекомендаций. Полагаем, что такая архитектура позволит эффективно реализовать широкий инструментарий методов интеллектуального анализа данных, что выгодно отличает предложенную архитектуру от известных решений, например [8].

Выскажем некоторые соображения по организации пользовательского интерфейса (схема организации рабочего окна представлена на рисунке 4).

Исходя из привычного пользовательского дизайна распространенных операционных систем линейки Windows, рабочее окно должно содержать главное меню (1), предоставляющее доступ к основным параметрам программной системы, справке и помощи, а также область статуса (5), в которой будут отображены лог основных событий и прочая служебная информация.

Основная рабочая область (3) окна предоставляет пользователю различные способы отображения результатов наукометрического анализа, в частности, таблицы, текстовые сведения, графическую информацию. Информация, представленная в рабочей области, должна предоставляться в зависимости от выбранного варианта использования и примененных инструментов, при этом кнопки инструментария размещаются на инструментальной панели (4).

Рис. 4. Схема рабочего окна пользовательского интерфейса

Fig. 4. GUI scheme

Выбор доступных вариантов использования системы осуществляется путем переключения вкладок (2), каждая из которых отвечает за свой вариант использования и предоставляет свой инструментарий на панели (4).

Разумеется, данная схема интерфейса – только достаточный вариант представления функциональных возможностей системы. В силу модульной организации системы и в случае ее реализации в виде, например, микросервисов пользовательский интерфейс может быть выполнен на основе независимого web-представления.

Некоторые результаты использования программной системы

Прототип описанной системы был реализован в виде десктопного приложения со встроенным ADO-хранилищем данных, развернутого на операционной системе Windows 10 для непосредственного взаимодействия с развернутой здесь же системой KNIME Analytics Platform. В рамках апробации программной системы в части реализованных методик построения публикационного рейтинга и оценки публикационного потенциала рассмотрим результаты работы на примере Южно-Уральского государственного университета (см. табл. 1). Для этой цели в хранилище программной системы импортированы данные наукометрических профилей авторов, отнесенных Scopus к области исследования Computer Science и аффилированных с университетом на 1 июля 2017 г.

Таблица 1

Общая характеристика объекта исследования

Table 1

General characteristic of the research object

Международное наименование университета	Общий QS-рейтинг за 2016	QS-research in 2016	Число авторов в области Computer Science
South Ural State University (SUSU)	151–200	Medium	228

Подготовительные этапы интеллектуального анализа в соответствии с применяемой методикой формируют наборы ассоциативных правил, имеющих вид, представленный на рисунке 5.

Для поиска правил анализируются наборы с не менее чем 5 элементами, что, с одной стороны, гарантирует гибкость правил (поскольку соответствие 7 из 7 элементов обеспечивает сам публикационный рейтинг), а с другой – отсекает огромное множество заведомо недостоверных правил с малым числом элементов.

В целом программная система на имеющихся данных строит 4 163 правила при минимальной поддержке и достоверности в 5,0 %. Выборка с надежностью не менее 50,0 % обнаруживает 78 правил. Однако для перехода авторов в публикационные квантили рейтингов Q1 и Q2, а именно

Поддержка набора (support), %	Достоверность правила (confidence), %	Значимость правила (lift), %	Результирующая группа (consequent)		Исходный набор уровней показателей (antecedent)		
5,12	93,33	392,47	rank_Q1	←	h_index_Q3	snip_mean_Q1	...
5,61	88,50	371,99	rank_Q1	←	cit_count_Q1	snip_mean_Q1	...
...	...	...	...		...	...	...

Рис. 5. Схема набора полученных ассоциативных правил

Fig. 5. A scheme of a set of obtained associative rules

они в большей степени интересуют с точки зрения анализа публикационного потенциала, выделено 20 лучших правил, отвечающих заданным критериям достоверности.

В соответствии с выделенными правилами 28 авторов могут быть потенциально (с разной степенью достоверности) переквалифицированы в лучший квартиль. Из этических соображений ID авторов в системе Scopus не публикуются, поэтому в таблице 2 представлены групповые оценки.

Таблица 2

Публикационный потенциал авторов в области Computer Science

Table 2

The publication potential of authors in Computer Science

Исходный квартиль	Число авторов	Средняя поддержка правил по квартилю, %	Средняя достоверность правил по квартилю, %	Целевой квартиль (consequent)
Q2	13	6,0	90,2	Q1
Q3	48	9,0	65,8	Q2
Q4	4	7,6	60,2	Q2

Детализация оценок публикационного потенциала в разрезе отдельных авторов может стать основой для принятия управленческих решений в сфере менеджмента научных исследований, в частности, о целевом стимулировании авторов в целях прогнозируемого улучшения наукометрических показателей. На рисунках (см. http://www.swsys.ru/uploaded/image/2018_2/2018-2-dop/3.jpg, http://www.swsys.ru/uploaded/image/2018_2/2018-2-dop/4.jpg) показаны примеры индивидуальных рекомендаций, формируемых программной системой.

Заключение

В ходе исследования в целом показано, что в настоящее время целесообразна разработка программной системы, призванной помочь руководителям исследовательских групп, а также индивидуальным исследователям принимать решения по поводу улучшения публикационной активности, наукометрической результативности, определять слабые и сильные стороны, направления и конкретные решения в виде автоматизированных рекомендаций по повышению величины и репрезентативности наукометрических показателей.

В ходе проектирования сформулированы общие и технические требования к программной системе; с использованием среды моделирования Microsoft

Visio построены необходимые диаграммы в нотации UML 2.0, описывающие систему.

Разработан методический инструментальный анализ публикационной активности с применением интеллектуального анализа наукометрических данных.

Разработаны модульная архитектура, алгоритмы и реализована программная система, позволяющая на основе интеллектуального анализа наукометрических данных формировать публикационный рейтинг и индивидуальные рекомендации по улучшению публикационной активности автора.

Проведены эксперименты по оценке публикационного потенциала исследователей, аффилированных с Южно-Уральским государственным университетом.

Работа выполнена при финансовой поддержке Правительства РФ (Постановление № 211 от 16.03.2013 г.), соглашение № 02.A03.21.0011.

Литература

1. Питерс Д., Марш Р. Rate my research dot com: измеряем то, что ценим, ценим, что измеряем // Научная периодика: проблемы и решения. 2011. № 1. С. 40–45.
2. Koenigstein N., Dror G., Yehuda Koren Y. Yahoo! music recommendations: modeling music ratings with temporal dynamics and item taxonomy. Proc. 5th ACM Conf. on Recommender systems. ACM, 2011, pp. 165–172.
3. Butler D. Computing giants launch free science metrics. Nature, 2011, vol. 476, p. 18.
4. He Y., Hui S.C. Mining a Web Citation Database for author co-citation analysis. Information Processing and Management, 2002, vol. 38, pp. 491–508.
5. Padrós-Cuxart R., Riera-Quintero C., March-Mir F. Bibliometrics: a Publication Analysis Tool. Proc. 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR 2016), 2016, pp. 44–53.
6. Harzing AW., Alakangas S. Microsoft Academic: is the Phoenix getting wings? 2016, vol. 106. DOI: 10.1007/s11192-015-1798-9.
7. Луценко Е.В., Орлов А.И., Глухов В.А. Наукометрическая интеллектуальная измерительная система по данным РИНЦ на основе АСК-анализа и системы «Эйдос» // Науч. журн. КубГАУ. 2016. № 122. С. 1–56.
8. Сеницын А.А., Никифоров О.Ю., Андреев М.А. Концепция и структура информационно-аналитической системы анализа публикационной активности сотрудников научно-образовательной организации // Фундаментальные исследования. 2014. № 11. С. 1276–1280.
9. Сеницын А.А., Никифоров О.Ю., Андреев М.А. Особенности применения информационно-аналитической системы для оценки направления поддержки по созданию результатов интеллектуальной деятельности научно-образовательной организации // Фундаментальные исследования. 2014. № 11-6. С. 1271–1275.
10. Мбого И.А., Прокудин Д.Е. Подходы к развитию инструментов автоматизации и интеграции ресурсов информационного пространства поддержки междисциплинарного научного направления // Интернет и современное общество IMS-2015: сб. науч. стат. XVIII объединен. конф. СПб, 2015. С. 290–302.

11. Мбого И.А., Прокудин Д.Е., Чугунов А.В. Разработка инструментов интеграции научной информации в пространстве разнородных информационных систем // Научный сервис в сети Интернет: тр. XVIII Всерос. науч. конф. М.: Изд-во ИПМ им. М.В. Келдыша, 2016. С. 249–258. DOI: 10.20948/abrau-2016-44.

12. Agrawal R., Srikant R. Fast algorithms for mining association rules. Proc. 20th Int. Conf. Very Large Databases, Santiago, Chile, 1994, pp. 487–499.

13. Borgelt C. Efficient Implementations of Apriori and Eclat. Proc. Workshop of Frequent Item Set Mining Implementations (FIMI 2003, Melbourne, FL, USA). URL: http://www.borgelt.net/papers/fimi_03.pdf (дата обращения: 02.10.2017).

14. Валько Д.В., Колташев А.С. Модуль агрегации наукометрических данных открытых сервисов в сети Интернет: пат. 2016619028. Рос. Федерация. № 2016616489; заявл. 21.06.16; опубл. 11.08.16, Бюл. № 9. 1 с.

Software & Systems

DOI: 10.15827/0236-235X.122.275-283

Received 12.10.17

2018, vol. 31, no. 2, pp. 275–283

A RECOMMENDATION SYSTEM BASED ON DATA MINING OF A SCIENTOMETRIC RESEARCH PROFILE

D.V. Valko¹, Graduate Student, ell.science@mail.ru

¹ South Ural State University (National Research University), Lenin Ave. 76, Chelyabinsk, 454080, Russian Federation

Abstract. Nowadays scientific results can be represented in various scientometric bases and systems. They are often popular not because of their relevance, but due to global availability. In fact, scientific results may be out of the scope of a scientific community simply because they are not placed in a popular scientometric system. From scientific point of view, such situation devalues a researcher regardless of quality and relevance of his scientific results.

According to the author, the development of recommendations for individual researchers, research teams and their managers at all levels of management would make it possible to pay attention to promising scientific results and reasonably accumulate necessary resources to include such results in popular scientometric systems.

Development of tools that operate the big scientometric open data cannot be without data mining methods. The paper shows that based on the algorithm of intellectual analysis of interrelations (like apriori), which is adapted to scientometric data in the Scopus, it is possible to formulate certain sets of associative rules suitable for forecasting probable future scientific results. It is also possible to develop automated recommendations for improving publication activity. The paper proposes the developed methodical tools for analyzing publication and scientometric data using data mining methods. In addition, it describes a modular architecture and a prototype of a software system that allows forming a publication rating and individual recommendations for improving author's publication activity based on scientometric data mining.

The paper shows some experimental results on assessing publication potential of researchers affiliated with the South Ural State University.

The authors built necessary diagrams in the UML 2.0 notation describing that software system using Microsoft Visio modeling environment.

Keywords: scientometrics, recommendation system, data mining, publication activity.

Acknowledgements. The work has been financially supported by the Russian government (Resolution no. 211 dated March 16, 2013), Agreement no. 02.A03.21.0011.

References

1. Pipers D., Marsh R. Rate my research dot com: we measure what we value, we value what we measure. *Nauchnaya periodika: problemy i resheniya* [Scholarly Communication Review]. 2011, no. 1, pp. 40–45 (in Russ.).
2. Koenigstein N., Dror G., Yehuda Koren Y. Yahoo! music recommendations: modeling music ratings with temporal dynamics and item taxonomy. *Proc. 5th ACM Conf. on Recommender Systems*. ACM Publ., 2011, pp. 165–172.
3. Butler D. Computing giants launch free science metrics. *Nature*. 2011, vol. 476, p. 18.
4. He Y., Hui S.C. Mining a Web Citation Database for author co-citation analysis. *Information Processing and Management*. 2002, vol. 38, pp. 491–508.
5. Padrós-Cuxart R., Riera-Quintero C., March-Mir F. Bibliometrics: a Publication Analysis Tool. *Proc. 3rd Workshop on Bibliometric-enhanced Information Retrieval (BIR 2016)*. 2016, pp. 44–53.
6. Harzing A.W., Alakangas S. Microsoft Academic: is the Phoenix getting wings? *Harzing.com*. 2016, vol. 106, DOI: 10.1007/s11192-015-1798-9.
7. Lutsenko E.V., Orlov A.I., Glukhov V.A. A scientometric intelligent measuring system of RSCI data based upon the ask analysis and Eidos system. *Nauchny zhurnal KubGAU* [Scientific J. of KubSAU]. 2016, no. 122 (08), pp. 1–56 (in Russ.).
8. Sinitsyn A.A., Nikiforov O.Yu., Andreev M.A. Concept and structure of the information-analytical system for to analyze scientific activity efficiency of the research and educational institution. *Fundamentalnye issledovaniya* [Fundamental research]. 2014, no. 11, pp. 1276–1280 (in Russ.).
9. Sinitsyn A.A., Nikiforov O.Yu., Andreev M.A. Features of the information-analytical system application for estimation the support areas for creation of the results of the intellectual activity of the research and educational institutions. *Fundamentalnye issledovaniya* [Fundamental research]. 2014, no. 11-6, pp. 1271–1275 (in Russ.).
10. Mbogo I.A., Prokudin D.E. Approaches to development of tools for automation and integration of information space resources to support interdisciplinary research. *Sb. nauch. st. XVIII Obyedinen. konf. "Internet i sovremennoe obshchestvo" IMS-2015* [Proc. 18th Joint Conf. "Internet and Contemporary Society" IMS-2015]. St. Petersburg, 2015, pp. 290–302 (in Russ.).
11. Mbogo I.A., Prokudin D.E., Chugunov A.V. Development of tools for integration of scientific information among heterogeneous information systems. *Nauchny servis v seti Internet: tr. XVIII Vseros. nauch. konf.* [Scientific Service on the Internet: Proc. 18th All-Russ. Sci. Conf.]. Moscow, IPM im. M.V. Keldysha Publ., 2016, pp. 249–258 (in Russ.).
12. Agrawal R., Srikant R. Fast algorithms for mining association rules. *Proc. 20th Int. Conf. on Very Large Databases*. Santiago, Chile, 1994, pp. 487–499.
13. Borgelt C. Efficient Implementations of Apriori and Eclat. *Proc. Workshop of Frequent Item Set Mining Implementations (FIMI 2003)*. Available at: http://www.borgelt.net/papers/fimi_03.pdf (accessed October 2, 2017).
14. Valko D.V., Koltashev A.S. Modul agregatsii naukometricheskikh dannykh otkrytykh servisov v seti Internet [Module for Aggregating Scientific Data of Open Services on the Internet]. Patent RF 2016619028, no. 2016616489.