УДК 004.048

Основные подходы к оценке эффективности публикационной деятельности профессорско-преподавательского состава кафедр

Малашков Валентин Борисович[a], Шульдова Светлана Георгиевна[b], Лапицкая Наталья Владимировна[c]

[a] *Белорусский университет информатики и радиоэлектроники, магистрант кафедры программного обеспечения информационных технологий, malashkovv@gmail.com*
[b] *Белорусский университет информатики и радиоэлектроники, кандидат технических наук, доцент, доцент кафедры программного обеспечения информационных технологий, shsg@bsuir.by*
[c] *Белорусский университет информатики и радиоэлектроники, кандидат технических наук, доцент, заведующий кафедрой программного обеспечения информационных технологий, lapan@bsuir.by*

Аннотация
В статье изложены подходы к оценке эффективности публикационной деятельности профессорско-преподавательского состава кафедр учреждений высшего образования, основанные на анализе наукометрических показателей и текста публикаций в рецензируемых научных журналах, а также методы и модели их реализующие.

Ключевые слова: публикация, цитируемость, автор, сеть, соавторство, наукометрический показатель, Google Scholar, граф.

Веб: http://library.miu.by/journals!/item.science-xxi/issue.9/article.4.html

Поступила в редакцию: 09.12.2020

Basic approaches of rating the publication activity of the teaching staff of the higher educational institution departments

Malashkov Valentin[a], Shuldova Svetlana[b], Lapitskaya Natalya[c]

[b] *Belarusian State University of Informatics and Radioelectronics, Master's degree student in the Department of Software of Information Technologies, malashkovv@gmail.com*
[b] *Belarusian State University of Informatics and Radioelectronics, PhD in Technical Sciences, Associate Professor, Associate Professor in the Department of Software of Information Technologies, shsg@bsuir.b*
[c] *Belarusian State University of Informatics and Radioelectronics, PhD in Technical Sciences, Associate Professor, Head of the Department of Software of Information Technologies, lapan@bsuir.by*

Abstract
The article describes approaches of rating the publication activity of the teaching staff of the departments of higher educational institutions, based on the analysis of scientometric indicators and the text of publications in peer-reviewed scientific journals, as well as methods and models that implement them.

Keywords: publication, citation, author, network, co-authorship, scientometric index, Google Scholar, graph.

Web: http://library.miu.by/journals!/item.science-xxi/issue.9/article.4.html

Received: 09.12.2020

Введение

Одним из обязательных условий научно-исследовательской деятельности, ее результатом и критерием являются научные публикации: тезисы, статьи, монографии, методические разработки.

Количество публикаций в расчете на одного педагогического работника из числа профессорско-преподавательского состава в рецензируемых научных журналах и позиция учреждения высшего образования в международных рейтингах являются критериями при оценке эффективности деятельности учреждений высшего образования в Республике Беларусь. Также наличие публикаций в рецензируемых научных изданиях Перечня Высшей аттестационной комиссии (ВАК) – критерий при присуждении ученых степеней и званий.

1. Количественные показатели публикационной деятельности

В последнее время в качестве инструментов оценки эффективности публикационной деятельности стали активно использоваться наукометрические показатели, к которым относятся [1]:

– количество публикаций (статьи, зачисленные в авторский профиль ученого или профиль научной организации; это показатель научной производительности, по которому можно оценивать автора (группу авторов), организацию, государство, журнал);

– суммарная цитируемость (показатель научной влиятельности или авторитетности, оценивает автора (группу авторов), организацию; цитируемость накапливается с годами);

– высокоцитируемые (имеющие наибольшее количество цитирований) публикации (показатель научной авторитетности, который отражает число действительно важных – в масштабах деятельности рассматриваемых автора или организации – публикаций);

– средняя цитируемость (отношение общего количество ссылок на статьи ученого к общему количеству статей);

– индекс Хирша (наукометрический показатель, предложенный в 2005 году Хорхе Хиршем из Калифорнийского университета в Сан-Диего; основан на учете числа публикаций исследователя и числа их цитирований; ученый имеет индекс h, если h из его N статей цитируются как минимум h раз каждая);

– импакт-фактор журнала (показатель авторитетности и влиятельности журнала).

Измерение наукометрических показателей осуществляется с помощью наукометрических баз данных и систем цитирования, которые с некоторой периодичностью получают из проиндексированных журналов библиографические описания статей, аннотации, ключевые слова и, главное, списки литературы – цитирования, или ссылки. Помимо журналов, индексируются книги и труды конференций. Затем система ищет в базе публикации, на которые ссылаются авторы добавленных работ, и прописывает соответствия, присоединяя ссылки к процитированным публикациям.

Наиболее известными научными поисковыми системами, которые объединяют электронные базы данных научного цитирования, являются Web of Science, Scopus, Microsoft Academic Search, Google Scholar. В отличие от Web of Science и Scopus Google Scholar имеет открытый доступ, индексирует большее число источников, содержит поддерживаемые авторами профили пользователей, однако не все индексируемые издания являются научными.

Представляется целесообразным разработать инструментарий, позволяющий выполнить анализ публикационной деятельности научно-педагогических работников на основе данных Google Scholar с учетом критериев ВАК. В этом случае к наукометрическим показателям необходимо добавить количество публикаций по категориям: монография, статья в зарубежном научном издании из Перечня ВАК Республики Беларусь, статья в зарубежном научном издании, статья в научном издании из Перечня ВАК Республики Беларусь, статья в научном издании Республики Беларусь, материалы конференций, сборники научных трудов, тезисы докладов.

2. Исходные данные для анализа

Источниками данных для анализа являются профиль ученого в Google Scholar, список зарубежных научных изданий, в которых могут быть опубликованы основные научные результаты диссертации на соискание ученой степени доктора и кандидата наук, и перечни научных изданий Республики Беларусь для опубликования результатов диссертационных исследований с 2005 года, размещенные на сайте ВАК. Также необходимо учитывать данные об авторах, месте их работы, остепененности, научных интересах.

Из открытого профиля сотрудника в Google Scholar по каждой статье могут быть получены следующие данные: цитируемость, автор (группа авторов), название, год, источник, издатель, URL-адрес статьи, URL-адрес цитирования, дата запроса, тип, DOI (Digital Object Identifier, цифровой идентификатор объекта, ЦИО), ISSN (International Standard Serial Number, международный стандартный номер для периодических изданий), выпуск, начальная страница, конечная страница, число цитирований в год, число цитирований на автора.

URL-адрес цитирования представляет собой ссылку на страницу со списком статей, цитирующих данную. Из открытой информации по отдельной статье также можно извлечь ее краткое описание.

Для извлечения данных используется программный интерфейс (Application Programming Interface, API) сервиса Google Scholar. На основе входных данных – фамилии и инициалов автора – можно получить информацию о его научных интересах, месте работы и список статей с URL-адресами цитирования.

При переходе по ссылкам статьи идентифицируются последовательно – одна за одной. В связи с тем, что данный API является внутренним и не предназначен для использования внешними пользователями, он защищен различными программными средствами от процесса веб-скрейпинга, что создает трудности при массовом извлечении данных.

Чтобы избежать блокировки IP-адресов, можно использовать сторонние прокси-сервисы. Также для оптимизации процесса веб-скрейпинга необходимы сопоставления со значениями справочных таблиц (перечни ВАК, издания и т. п.) и значительные преобразования уже на стадии извлечения информации.

Для извлечения данных написан скрипт на языке программирования Python с использованием библиотеки с открытым исходным кодом под названием Scholarly. Данная библиотека берет на себя роль готовой «обертки», а также имеет готовый функционал для переадресации запросов через предоставленный список прокси-серверов в случае блокировки очередного IP-адреса.

Процесс извлечения и загрузки данных в реляционные таблицы представлен в виде DFD-диаграммы на рисунке 1.

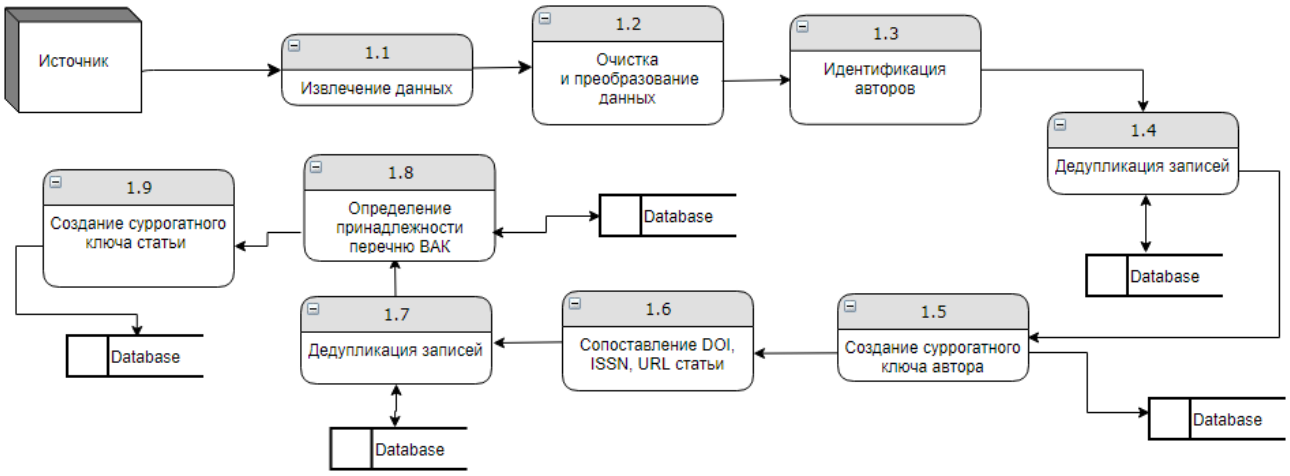

Рисунок 1 – Процесс извлечения, преобразования и загрузки данных

Данные в исходном виде выгружаются в два табличных файла формата CSV. Первый файл содержит в себе информацию об авторах, второй – о статьях.

Далее осуществляются очистка и дедупликация данных, после чего они загружаются в хранилище. Информация об авторах содержит в себе внутренний идентификатор (ID) сервиса Google Scholar, что позволяет легко отбросить дупликаты и создать на его основе суррогатный первичный ключ.

К сожалению, статьи не имеют внутреннего ID, поэтому для дедупликации и создания первичного ключа в хранилище данных можно использовать DOI, ISSN, а также URL статьи.

Схема базы данных для хранения и подготовки их к анализу показана на рисунке 2. Для организации хранения данных использована технология хранилищ данных, поскольку она ориентирована на поддержку процессов анализа.

Таблица «Сотрудники» может быть заполнена данными из базы данных информационной системы УВО. Эта таблица относится к типу 2 медленно меняющихся измерений, так как ученая степень и ученое звание могут со временем измениться. Поэтому таблица содержит суррогатный первичный ключ и дополнительные поля с датами. Поле «Дата изменения» не заполняется для текущих значений полей.

В качестве фактов в хранилище используется факт публикации – таблица «Публикации авторов», содержащая два внешних ключа (ключ, указывающий на автора, и ключ, указывающий на публикацию), а также факт цитирования – таблица «Цитирования» (см. рисунок 2).

3. Задачи оценки значимости публикационной деятельности и способы их решения

На первом этапе был собран небольшой набор данных и на его основе произведен анализ распределения количества цитирований на статью. На рисунке 3 представлена гистограмма частотного распределения количества статей в выборке в зависимости от количества цитирований.

Для анализа использованы статьи, имеющие количество цитирований более 10. В дальнейшем планируется для каждого автора определять относительное минимальное значение количества цитирований. Основной задачей анализа на основе собранных данных будет определение направления научных исследований, а также ассоциированных с ними научных сообществ, авторов, журналов и т. д. Для этой задачи основными источниками будут текст, цитирования, а также информация о соавторстве.

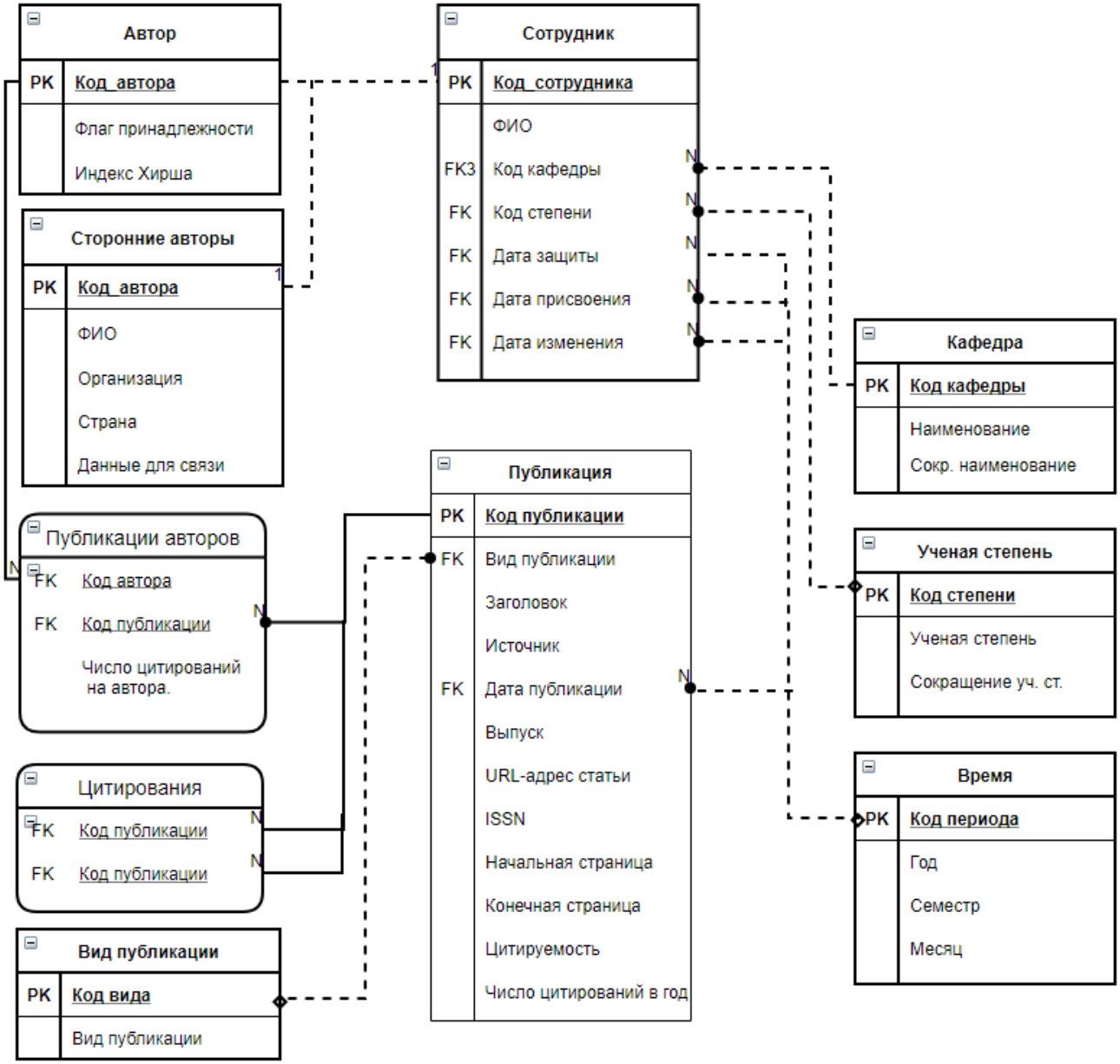

Рисунок 2 – Реляционная схема хранилища данных

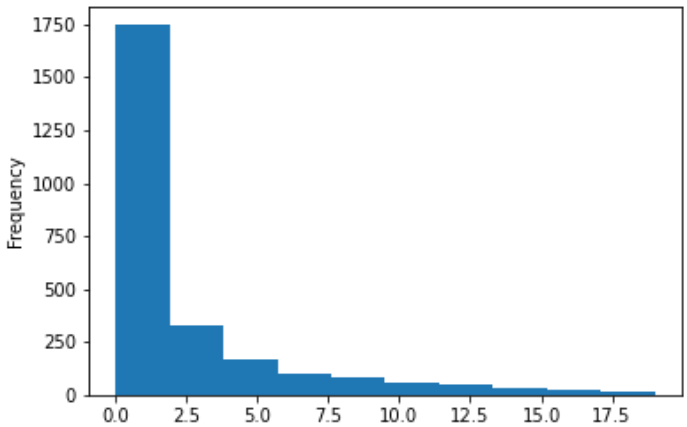

Рисунок 3 – Гистограмма частотного распределения количества статей от количества цитирований на статью

В качестве текста для анализа доступны название и краткое описание статьи. Чтобы выделить отдельные тематики в тексте научных исследований, используются алгоритмы машинного обучения для задач тематического моделирования. Относительно популярным и универсальным алгоритмом в данной области является латентное размещение Дирихле (LDA, latent Dirichlet allocation) [2], суть которого заключается в следующем.

Для начала данные необходимо разделить на отдельные слова или токены. Затем токены очистить от различных стоп-слов (это могут быть союзы и местоимения). Чаще всего уже существуют готовые решения, содержащие списки часто встречаемых стоп-слов для различных языков. Особенностью собранного набора данных является то, что он может содержать описание статей как на русском языке, так и на английском, что необходимо учитывать при поиске нужного инструмента. Вместо отдельных слов можно рассмотреть отдельные выражения. Статья может содержать уже устоявшиеся научные термины, такие как «преобразование Фурье» или «машинное обучение». Для выделения отдельных ключевых фраз существуют алгоритмы машинного обучения. Они могут быть как простыми, где можно задать максимальную длину искомой фразы, так и более сложными, где присутствует возможность задавать «шаблон фразы» с помощью частеречной разметки [3].

Выделив набор токенов, можно применить алгоритм LDA и в качестве гиперпараметров определить количество токенов на одну тематику, а также количество тематик. На основе полученных токенов и тематик, ассоциированных с каждой из статей, можно оценить, насколько тематики научных исследований пересекаются друг с другом, а также какие авторы, институты и журналы являются ключевыми в данных научных направлениях.

Располагая информацией о датах публикаций, возможно проследить динамику развития обозначенной тематики научных исследований.

Данные о цитировании являются основой для построения сети цитирования публикаций автора $N = (P, R)$, в которой узлами P являются научные статьи, а направленными (ориентированными) связями R – ссылки на статьи внутри другой статьи, то есть [4]

$$p_i \, R \, p_j \equiv p_j \text{ цитирует } p_i.$$

Такая сеть может быть представлена в виде ориентированного графа $G = (V, E)$, в котором вершины $V = \{V_1, V_2, .., V_n\}$ соответствуют статьям, а дуги – отношениям цитирования $E = V \times V$, $e = \langle v_i, v_j \rangle \in E$, $i \neq j$, так как самоцитирование исключено. Также необходимо учесть, что p_j может цитировать p_i только один раз. Если p_j цитирует p_i, то p_i не может цитировать p_j, то есть петель, кратных ребер и циклов в графе нет, а вес ребра, соединяющего цитируемую и цитирующие статьи, $f(e) = 1$. В таблице «Цитирования» (см. рисунок 2) каждая строка – ребро направленного графа.

Для определения научных сообществ основным источником данных является информация о соавторстве. На основе этих данных можно построить неориентированный взвешенный граф $G = (V, E)$, в котором вершины соответствуют авторам, а дуги – отношениям соавторства. Вес каждой дуги определяется количеством общих статей двух авторов.

К данным сетей цитирования и соавторства можно применить один из оптимизационных алгоритмов для разбиения сети на отдельно взятые модули. В качестве начального примера такого алгоритма можно взять лувенский метод [5]. Это позволит выделить отдельные кластеры научных сообществ авторов, а также дать оценку модулярности внутри данного сообщества.

При использовании ранее отобранных тематик на основе текста появляется возможность оценить, какие именно тематики и в какой степени ассоциируются с каким из научных сообществ. Для этого можно выделить все научные статьи, опубликованные в рамках сообщества авторов, а затем выделить топ N тематик на основе ранее проведенного анализа LDA текстов научных публикаций.

Для определения ключевых авторов можно воспользоваться как количественными данными, такими как количество цитирований, количество публикаций, индекс Хирша, так и ориентированным графом цитирований, описанным ранее, на стадии сбора данных. Для данной задачи подходит классический оптимизационный алгоритм PageRank для анализа соединений ориентированного графа [6]. Он позволяет дать количественную оценку релевантности той или иной вершины в зависимости от того, какое количество вершин на нее ссылается, а также на основе их оценки релевантности. Самый очевидный подход – это анализ на уровне отдельных публикаций, а затем суммирование всех оценок релевантности отдельных публикаций в рамках одного автора.

Заключение

На основе собранных данных из открытых профилей сотрудников учреждений высшего образования из числа профессорско-преподавательского состава можно определить количественные показатели публикационной деятельности, а также решить ряд задач, обуславливающих научную значимость публикаций:

– определить перспективные научные направления;

– идентифицировать научные сообщества и ключевых авторов.

Изложенные подходы к оценке эффективности публикационной деятельности основаны на использовании технологий многомерного и интеллектуального анализа данных. Спроектирована реляционная схема данных для обеспечения многомерного анализа. Дальнейшее расширение схемы предполагает

добавление данных о периодических изданиях, что позволит оценить влияние импакт-фактора издания на показатели научной значимости публикации. Для анализа в этом направлении особое значение также имеет принадлежность изданий к перечням ВАК.

ЛИТЕРАТУРА / REFERENCES

1. Индикаторы науки: 2017 [Электронный ресурс]: стат. сб. / Ю.Л. Войнилов [и др.]. – М.: НИУ ВШЭ, 2017. – 304 с. – Режим доступа: https://www. hse.ru/primarydata/in2017. – Дата доступа: 25.04.2020.

Indikatory nauki: 2017 [Electronic resource]: stat. sb. / Yu.L. Voynilov [et al.]. – M.: NIU VSHE, 2017. – 304 p. – Mode of access: https://www. hse.ru/primarydata/in2017. – Date of access: 25.04.2020.

2. Blei, D.M. Latent Dirichlet allocation / D.M. Blei, A.Y. Ng, M.I. Jordan // Journal of Machine Learning Research. – 2003. – № 3. – P. 993–1022.

3. Ванюшкин, А.С. Методы и алгоритмы извлечения ключевых слов / А.С. Ванюшкин, Л.А. Гращенко // Новые информационные технологии в автоматизированных системах. – 2016. – № 19. – С. 85–93.

Vanyushkin, A.S. Metody i algoritmy izvlecheniya klyuchevykh slov / A.S. Vanyushkin, L.A. Grashchenko // Novyye informatsionnyye tekhnologii v avtomatizirovannykh sistemakh. – 2016. – № 19. – P. 85–93.

4. Параметры «центральности» узлов сети цитирования научных статей / С.В. Бредихин [и др.] // Проблемы информатики. – 2016. – № 1. – С. 39–57.

Parametry «tsentral'nosti» uzlov seti tsitirovaniya nauchnykh statey / S.V. Bredikhin [et al.] // Problemy informatiki. – 2016. – № 1. – P. 39–57.

5. Fast unfolding of community hierarchies in large networks [Electronic resource] / V.D. Blondel [et al.]. – Mode of access: https://arxiv.org/pdf/0803.0476.pdf. – Date of access: 14.10.2020.

6. PageRank Citation Ranking: Bringing Order to the Web [Electronic resource]. – Mode of access: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf. – Date of access: 14.10.2020.