Использование предметной онтологии единого цифрового пространства научных знаний в наукометрических задачах

В. А. Цветкова

ИНИОН РАН

Н. Е. Каленов

МСЦ РАН

Ю. В. Мохначева , И. А. Митрошин

БЕН РАН

Аннотация. Формирование Единого цифрового пространства научных знаний (ЕЦПНЗ) является одной из приоритетных задач современного периода и подразумевает использование унифицированных подходов к отбору информации, её представлению, навигации, терминологическому наполнению. При построении предметной онтологии научного направления – «Микробиология» использован метод статистического подсчёта частотного распределения ключевых слов в информационных массивах, сформированных в результате тематического поиска информации по микробиологии. Гипотеза исследования: частота встречаемости КС может являться индикатором интенсивности развития научной темы. Результаты исследований позволили предположить, что размер пула ключевых слов в рубриках может служить индикатором, показывающим широту спектра разнообразия исследований, их объектов и методов.

Ключевые слова: пространство знаний, классификаторы, цитатный анализ, частотное распределение, ключевое слово, предметная онтология.

Using the subject ontology of the unified digital space of scientific knowledge in scientometric tasks

V. A. Tsvetkova

ISIS RAS

Abstract. The Formation of a Common digital space of scientific knowledge (CDSSK) is one of the priorities of the modern period and involves the use of unified approaches to the selection of information, its presentation, navigation, and terminological content. When constructing an subject ontology of the scientific direction – «Microbiology» – we applied a method of statistical calculation of the frequency distribution of keywords (KW) in information arrays formed as a result of thematic search for information on Microbiology. We assume that the frequency of KW occurrence can be considered as indicator of the scientific topic development intensity. The results of the study show that the size of the pool of keywords in categories can serve as an indicator that shows the breadth of the spectrum of research diversity, methods and objects: the more keywords in a category, the greater the diversity of research directions.

Keywords: knowledge space, classifiers, citation analysis, frequency distribution, keywords, subject ontology.

В числе приоритетных задач современного периода рассматриваются подходы к формированию Единого цифрового пространства научных знаний (ЕЦПНЗ), что обусловлено активным переходом к формированию цифровой экономики, реализацией программ по осуществлению цифровизации страны в тех сферах, в которых фактор знания играет ключевую роль. В работе [1] детально рассмотрено Единое российское цифровое пространство научных знаний. Под онтологией понимается формальное описание некоторой области знания, которое может быть использовано для автоматической обработки компьютером [2].

Исследование направлено на создание единого пространства знаний и на оценку возможных направлений его развития на примере тематического направления «Микробиология», в основу которого положены информационно-лингвистические элементы. Исследования выполнены при поддержке грантов РФФИ № 18-00-00294комфи и № 18-00-00372комфи.

Формирование единого пространства научных знаний подразумевает использование унифицированных подходов к отбору информации, её представлению, навигации, терминологическому наполнению. В основу предложенной методологии положен анализ классификационных систем, описывающих данное научное направление, и установление связей между ними [3]; анализ публикаций, отраженных в базах данных цитирования по данному научному направлению, с выделением массивов ключевых терминов, установление связей между ними и с индексами классификационных систем, в первую очередь, таких как: Государственный рубрикатор научно-технической информации (ГРНТИ), Рубрикатор отраслей знаний ВИНИТИ РАН, Универсальная десятичная классификация (УДК), Международная классификация изобретений (МКИ) и др. Проблема заключается в том, что перечисленные классификаторы во многом не согласуются между собой и значительно отличаются друг от друга по множеству критериев [4].

На основе особенностей и логики существующих классификаторов был разработан локальный базовый рубрикатор (ЛБР) по микробиологии, основные принципы и технология создания которого были подробно изложены в публикациях [3–4]. При разработке классификатора были установлены тезаурусные связи его рубрик с индексами существующих классификационных систем (КС). Установленные связи между терминами ЛБР и элементами КС принимают одно из значений: «эквивалентны» (тождественны), «содержит...» (шире), «входит в...» (уже), «пересекается с...». Последний вид связи эквивалентен принятому в классификационных системах указанию «см. также...». Локальные рубрики, сформулированные на языке современных терминов микробиологии, использовались для формирования глоссария понятий, являющихся основой предметной онтологии. Глоссарий включает ключевые термины (КТ), основой для отбора которых являлась база данных Web of Science Core Collection (WoS CC).

Микробиология – очень широкая и одна из самых быстроразвивающихся областей знания. Непрерывающееся

эволюционирование этого научного направления приводит к тому, что из классификаторов какие-то коды исчезают вообще, а какие-то появляются впервые. Кроме того, анализируемая предметная область не всегда вписывается в таксономию, предлагаемую существующими классификаторами.

При построении онтологии научного направления «Микробиология» использовался метод статистического подсчёта частотного распределения ключевых слов (КС) в информационных массивах, сформированных в результате тематического поиска информации по микробиологии. Как уже отмечено, в качестве основной информационной базы была использована Web of Science Core Collection. При построении структуры рубрикатора и составлении перечня КС учитывалась

частотность повторов того или иного термина в документальных массивах. В дальнейшем эти КС легли в основу глоссария терминов (понятий) для разработки онтологии.

На следующем этапе были построены деревья классификации понятий (иерархии классов), которых в онтологии может быть несколько. Именно эти понятия и устанавливают основные связи между классами. Всё это даёт возможность построить алгоритмы для реализации метода автоматизированного построения онтологии конкретной предметной области. Затем эксперты предметной области, для которой разрабатывается онтология (в нашем случае — «Микробиология»), определяют правила логических выводов, предоставляющие возможность использовать полученные данные, представленные в онтологии, и извлекать из неё новые знания.

Для слежения за динамикой терминологии тематического направления очень хорошо показал себя подход, разработанный на основе библиометрических методов. При этом в качестве единицы анализа выступают КС, или некоторая совокупность слов, выражающих проблему (предмет, отрасль, направление и др.). По изменениям частоты встречаемости единиц анализа в текстах можно делать выводы

об изменениях в научных направлениях, выявлять наиболее релевантные изучаемому предмету документы.

В исследовании использовался пул публикаций, сформированный из наиболее цитируемых документов по микробиологии. Наиболее цитируемые работы рассматривались как наиболее востребованные научным сообществом. Из этих документов в отдельный массив отбирались авторские КС и определялась частотность их встречаемости. Все КС были переведены на русский язык. Таким образом, был заложен фундамент для формирования онтологии.

Массив публикаций с наибольшей цитируемостью формировался по следующей схеме. Сначала был сформирован полный пул сведений о документах по микробиологии за необходимый период, затем производилось нисходящее ранжирование публикаций по их цитируемости. Суммарная цитируемость всего массива документов была разделена на 3 части, в которых суммарная цитируемость документов была приблизительно равна между собой. Для последующего анализа была отобрана первая (верхняя) треть от массива публикаций, на чью долю приходилась треть от всех ссылок. Эти публикации мы принимали как наиболее цитируемые, и дальнейшая работа производилась на этом массиве. Специалистами-микробиологами были исключены «шумовые» термины, после чего все КТ были переведены на русский язык, данные были загружены в модифицированную систему терминологических словарей [5]. Таким образом, был заложен фундамент для формирования предметной онтологии.

По нашему мнению, частота встречаемости КС может являться индикатором интенсивности развития научной темы. Однако частота встречаемости ни в коей мере не может служить критерием оценки значимости научного направления [6].

Результаты исследований на основе данных за 2018–2019 гг. следующие. Было выявлено 5865 документов за 2018–2019 гг. Публикации учитывались однократно с удалением дублированных записей. По числу документов выделяется рубрика «Генетика дрожжей и микроскопических грибов»: она вобрала в себя наибольшее количество документов. Видимо, это вы-

звано развитием программ генетического картирования, целью которых служит получение бактериальных ферментов и лекарств, изучения антибиотической резистентности, развитие биотехнологии. Кроме того, в документах этой рубрики содержится значительная доля уникальных ключевых слов (УКС) – 83 %¹.

К числу активно развивающихся рубрик можно также отнести «Почвенную микробиологию»: 250 документов, в которых присутствует 92 % УКС. На третьем месте – «Геомикробиология»: 248 документов с 79 % УКС. На общем фоне также выделяются рубрики: «Взаимоотношение возбудителя и хозяина» – 235 документов с 80 % УКС и «Генетика бактерий» – 230 документов с 80 % УКС.

Можно предположить, что доля УКС в рубриках может служить индикатором, показывающим широту спектра разнообразия исследований, методов, организмов: чем больше доля таких слов, тем более разнообразными являются исследования по направлению. Проблема выявления наиболее активно развивающихся научных тем является очень актуальной. Слежение за динамикой терминологии научных направлений (в данном случае – Микробиологии) позволяет делать выводы об происходящих внутри направления изменениях. Показательным является анализ такой динамики у наиболее цитируемых публикаций, как наиболее востребованных научным сообществом. На примере разработанного нами рубрикатора «Микробиология» мы показали один из возможных подходов к определению наиболее активно развивающихся научных тем по основе метода частотного распределения ключевых слов.

Информация, сформированная на этом этапе, открывает широкие возможности для проведения наукометрических исследований данной области науки, базирующихся

¹ Уникальное ключевое слово – ключевое слово, учтённое только один раз, несмотря на возможную множественность упоминаний в разных документах. Доля УКС = кол-во УКС / количество всех КС в документах (включая повторы одних и тех же КС в разных статьях) × 100 %.

на сравнительном анализе терминологии, используемой в отдельных тематических направлениях. В качестве единицы анализа выступают КТ, отражающие разные аспекты исследований по данной проблеме (предмет, теоретические результаты, аспекты применения и др.). По изменениям частоты встречаемости единиц анализа в текстах можно делать выводы об изменениях в научных направлениях, выявлять наиболее релевантные изучаемому предмету документы. По нашему мнению, частота встречаемости КТ в репрезентативном массиве публикаций является одним из индикаторов интенсивности развития научной темы. Однако частота встречаемости ни в коей мере не может служить критерием оценки значимости научного направления [7].

Анализ на основе количественных оценок публикаций по различным разделам научной темы данного научного направления, динамика повторяемости и пересечения ключевых терминов, их изменение во времени (путем анализа документов, опубликованных за различные интервалы годов) позволяют делать выводы о наиболее и наименее динамично развивающихся научных темах, их взаимопроникновении, практическом применении результатов теоретических исследований [8].

Сравнительный анализ различных направлений исследований, относящихся к микробиологии, позволил выделить наиболее активно развивающиеся направления, которые характеризуются большим количеством уникальных (неповторяющихся) КТ и т. п. Работы в этом направлении предполагаются продолжить в направлении расширения анализируемой информационной базы за счет привлечения данных из Scopus и РИНЦ.

Исследования выполнены при поддержке грантов РФФИ № 18-00-00294комфи и № 18-00-00372комфи.

Литература

1. Антопольский А. Б., Каленов Н. Е., Серебряков В. А., Сотников А. Н. О едином цифровом пространстве научных знаний // Вестник Российской академии наук. 2019. Т. 89. № 7. С. 728–735.

2. Трусова Ю. О., Белоозеров В. Н. Представление классификационных систем в виде онтологий. Обзор // Науч.-техн. информ. Сер. 1. Орг. и методика информ. работы. 2015. № 11. С. 34–38.

3. Антопольский А. Б., Белоозеров В. Н., Каленов Н. Е., Маркарова Т. С. О развитии терминологической базы данных в виде комплекса отраслевых информационно-поисковых тезаурусов // Информационные ресурсы России, 2018. № 5 (165). С. 22–30.

4. Цветкова В. А., Мохначева Ю. В., Харыбина Т. Н., Бескаравайная Е. В., Митрошин И. А. Пространство знаний: подходы к извлечению знаний из научных текстов // Информационные ресурсы России. 2019. № 2. С. 31–34.

5. Kalenov N., Senko A. Interactive System of Terminological Dictionaries as One of the Elements in the Ontology of Scientific Knowledge // Software Journal: Theory and Applications. 2019. № 4. <http://swsysweb.ru/en/interactive-system-of-terminological-dictionaries.html> (дата обращения 15.07.2020).

6. В. А. Цветкова, Т. Н. Харыбина, Ю. В. Мохначева, Е. В. Бескаравайная, И. Ю. Митрошина Особенности совмещения классификационных систем и формирования массива ключевых слов для определения пространства знаний по микробиологии // Науч. и техн. б-ки. 2019. № 11. С. 25–43.

7. Mengyang Wang, Lihe Chai Three new bibliometric indicators/approaches derived from keyword analysis // Scientometrics. 2018. V. 116. P. 721–750.

8. Цветкова В. А., Каленов Н. Е., Мохначева Ю. В., Митрошин И. А. Предметная онтология единого цифрового пространства научных знаний как источник наукометрических исследований // Информационные ресурсы России. 2020. № 5. С. 47–49.

References

1. Antopol'skij A. B., Kalenov N. E., Serebryakov V. A., Sotnikov A. N. O edinom cifrovom prostranstve nauchny`x znaniy // Vestnik Rossijskoj akademii nauk. 2019. T. 89. № 7. S. 728–735.

2. Trusova Yu. O., Beloozerov V. N. Predstavlenie klassifikacionny`x sistem v vide ontologij. Obzor // Nauch.-texn. inform. Ser. 1. Org. i metodika inform. raboty`. 2015. № 11. S. 34–38.

3. Antopol'skij A. B., Beloozerov V. N., Kalenov N. E., Markarova T. S. O razvitii terminologicheskoy bazy` danny`x v vide kompleksa otraslevy`x informacionno-poiskovy`x tezaurusov // Informacionny`e resursy` Rossii, 2018. № 5 (165). S. 22–30.

4. Tsvetkova V. A., Moxnacheva Yu. V., Xary`bina T. N., Beskaravajnaya E. V., Mitroshin I. A. Prostranstvo znaniy: podxody` k izvlecheniyu

znaniy iz nauchny`x tekstov // Informacionny`e resursy` Rossii. 2019. № 2. S. 31–34.

5. Kalenov N., Senko A. Interactive System of Terminological Dictionaries as One of the Elements in the Ontology of Scientific Knowledge // Software Journal: Theory and Applications. 2019. № 4. <http://swsys-web.ru/en/interactive-system-of-terminological-dictionaries.html> (data obrashheniya 15.07.2020).

6. Tsvetkova V. A., Xary`bina T. N., Moxnacheva Yu. V., Beskaravajnaya E. V., Mitroshina I. Yu. Osobennosti sovmeshheniya klassifikacionny`x sistem i formirovaniya massiva klyuchevy`x slov dlya opredeleniya prostranstva znaniy po mikrobiologii // Nauch. i texn. b-ki. 2019. № 11. S. 25–43.

7. Mengyang Wang, Lihe Chai Three new bibliometric indicators/approaches derived from keyword analysis // Scientometrics. 2018. V. 116. P. 721–750.

8. Tsvetkova V. A., Kalenov N. E., Moxnacheva Yu. V., Mitroshin I. A. Predmetnaya ontologiya edinogo cifrovogo prostranstva nauchny`x znaniy kak istochnik naukometricheskix issledovanij // Informacionny`e resursy` Rossii. 2020. № 5. S. 47–49.