Инструменты оценки научно-технологического ландшафта страны

А.В. Рудик [1], Е.В. Антонов [1], А.А. Артамонов [1]

[1] *Национальный исследовательский ядерный университет «МИФИ», Каширское шоссе, д. 31, г. Москва, 115409, Россия*

Аннотация

Исследование научно-технологического потенциала является важной задачей в определении технологического лидерства стран. Для его оценки не существует универсального показателя, принято рассматривать совокупность косвенных показателей для осуществления подобных исследований. В работе рассматривается инструмент оценки научно-технологического ландшафта стран на примере Японии и Республики Корея, а также визуализация полученных результатов, которая позволяет представить большие объемы данных в простом для восприятия формате. В работе анализируются существующие методологии оценки научно-технологического ландшафта, включая методологию оценки потенциала модернизации промышленного комплекса, технологического прогнозирования и социальных изменений через сеть цитирования и анализ тем, патентный ландшафт. Рассматриваются сбор и обработка научных публикаций выбранных стран в количестве 1 803 000 научных публикаций с помощью программных инструментов, составлено хранилище данных. Представлен инструмент оценки научно-технологического ландшафта, включая его визуализацию в виде 3D графика в масштабе научной области и страны в целом.

Ключевые слова

Научно-технический потенциал, технологическое лидерство, технологическое прогнозирование, научная визуализация, инструменты оценки, программная обработка данных.

Tools for Assessing a Country's Science and Technology landscape

A.V. Rudik [1], E.V. Antonov [1], A.A. Artamonov [1]

[1] *National Research Nuclear University "MEPhI", Kashirskoye shosse, b. 31, Moscow, 115409, Russia*

Abstract

The study of scientific and technological potential is an important task in determining the technological leadership of countries. There is no universal indicator for its assessment, it is common to consider a set of indirect indicators for the implementation of such studies. The paper considers the tool for assessing the S&T landscape of countries on the example of Japan and the Republic of Korea, as well as the visualization of the results, which allows to present large amounts of data in an easy-to-understand format. The paper analyzes existing methodologies for assessing the S&T landscape, including the methodology for assessing the potential for modernization of the industrial complex, technological forecasting and social change through the citation network and topic analysis, patent landscape. The collection and processing of scientific publications of the selected countries of 1,803,000 by means of software tools are considered, and a data repository is compiled. A tool for assessing the S&T landscape is presented, including its visualization in the form of a 3D graph at the scale of the scientific field and the country as a whole.

Keywords

Scientific and technological potential, technological leadership, technological forecasting, scientific visualization, assessment tools, software data processing.

ГрафиКон 2023: 33-я Международная конференция по компьютерной графике и машинному зрению, 19-21 сентября 2023 г., Институт проблем управления им. В.А. Трапезникова Российской академии наук, г. Москва, Россия
EMAIL: avrudik@kaf65.ru (А. В. Рудик); eantonov@kaf65.ru (Е. В. Антонов); aartamonov@kaf65.ru (А. А. Артамонов)
ORCID: 0000-0002-1757-1681 (А. В. Рудик); 0000-0003-1498-9131 (Е. В. Антонов); 0000-0002-9140-5526 (А. А. Артамонов)

1. Введение

Научно-технологический ландшафт (НТЛ) — текущее состояние и развитие научно-технических областей в конкретном регионе или стране, коллективная экосистема научных знаний, технологических достижений, организаций, политических мер и отдельных лиц, которые способствуют развитию, распространению и применению знаний и технологий в определенной области, регионе или во всем мире. НТЛ является важным индикатором развития страны, позволяющим оценить ее научно-технический потенциал и определить направления дальнейшего развития в области науки, технологий и т.д. Для взаимодействия с научно-технологическим ландшафтом необходимы инструменты, позволяющие визуализировать, оценивать его текущее состояние и прогнозировать его развитие [1].Необходимость разработки инструментов оценки научно-технологического состояния страны присутствует всегда, однако в настоящее время дополнительными препятствиями для исследователей выступают ограничения доступа к международным реферативным базам данных на территории Российской Федерации, а также усложнение экспериментов в различных областях. В этих условиях важно понимать, что происходит в той или иной тематической области, а в частности, какие исследования привели ее к текущему состоянию. Именно поэтому главной целью работы является разработка инструмента, который поможет исследователю в поиске и анализе данных больших объемов.

За последнее десятилетие написано множество статей, касающихся данной темы. В рамках обзора существующих методов оценки НТЛ страны рассмотрены работы в различных русско- и англоязычных реферативных базах данных научных публикаций. По результату обзора литературы выявлено, что авторы в основном концентрируются на изучении тенденций определенной области (патентного ландшафта, модернизации промышленного комплекса, науки, технологического прогнозирования).

Например, в статье И. В. Макаровой и А. Д. Максимова [2] рассмотрена методология анализа и оценки потенциала промышленного комплекса. Подход, рассмотренный в статье, позволяет оценить весомость промышленного комплекса с точки зрения обеспечения долгосрочного потенциала модернизации. Представленные в статье данные и расчеты, подтверждающие указанные выводы не позволяют воспроизвести полученные результаты с той же точностью. Авторы не описаются на анализ научных публикаций, взаимосвязь развития области с развитием рубежа знаний в этой же области, отображенного в статьях ученых, вместо этого расчеты строятся на основе видов капитала и экономических ресурсов.

В своей работе Джованни Абрамо и Чириако Андреа Д'Анджело [3] приводят пример расчета, в котором учитывается, что в расходах на исследования факторы производства (входные данные) состоят из труда и капитала (все ресурсы, кроме труда, такие как научные приборы, базы данных, здания и т.д.). Предлагаемая методология определения научных сильных и слабых сторон страны позволяет избежать искажений, связанных с зависимостью от размера традиционных методик, т.е. методик, измеряющих долю статей, цитирований или высокоцитируемых статей страны по отношению к общему мировому объему. Отличительной особенностью этой статьи является рассмотрение доходов научных сотрудников, однако не рассматривается анализ содержания самих научных публикаций и взаимосвязей их с развитием науки в той или иной области.

В статье авторов Юя Кадзикава, Кристиана Меджия, Мэнцзя У и И Чжана [4] рассматривается академический ландшафт журнала “Technological Forecasting and Social Change” (TFSC), выделяются основные тенденции тематик журнала, используя библиометрические методы. Данная статья хорошо показывает, как много информации могут дать библиометрические показатели научных публикаций, особенно если масштабировать данную аналитическую методологию на большие массивы статей [5]. Единственное, что ограничивало авторов статьи – ее тематика, дающая возможность рассматривать только научные публикации выбранного журнала.

В работе Н.Г. Кураковой, Л.А. Цветковой, В.Г. Зинова [6] рассматривается анализ патентного ландшафта с выявлением драйверов технологического развития страны для выбора научно-технологических приоритетов страны.

В ходе обзора литературы выделено несколько методов, однако многими авторами не было представлено описания того, как в точности применять данные методы, к тому же во многих из

них не было представлено визуализации полученных результатов. Не было выявлено инструментов для оценки научно-технологического состояния страны в виде визуализаций на основе научных публикаций стран, сделан вывод о необходимости создания методики по собственным критериям оценки. Особенностью работы является рассмотрение научно-технологического ландшафта страны в нескольких масштабах (на уровне страны в целом, а также на уровне отдельной научной области). Данных подход позволяет при необходимости искусственно «углубляться» в тематическую область, рассматривать ландшафт областей, входящих в нее, а область задается не существующими рубрикаторами, а может задаваться заказчиком (пользователем).

2. Разработка инструмента

В рамках работы разработана методика построения инструмента для оценки НТЛ (рисунок 1).

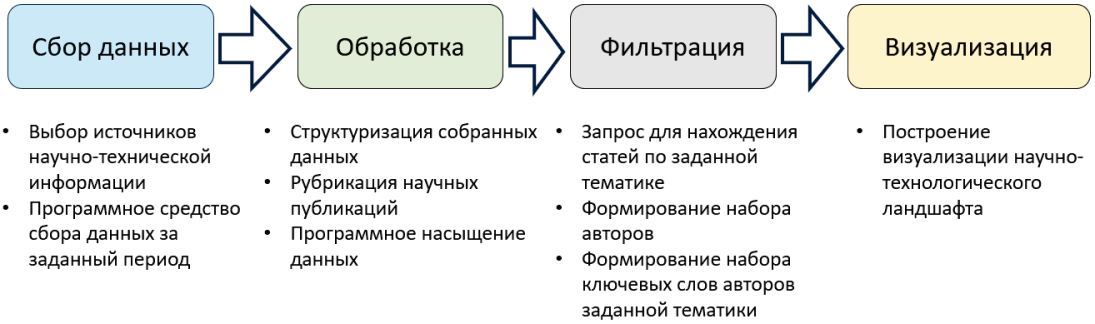

Рисунок 1 – Методика создания инструмента оценки НТЛ

На этапе сбора данных осуществлен выбор источников информации и разработано программное средство сбора данных за заданный период, используя автоматизированную технологию сбора данных [7, 8]. При этом от выбора источника данных напрямую зависят получаемые результаты (НТЛ, полученный из данных патентов, будет отличаться по содержанию от данных по научным публикациям [9]). Главная задача разработанного инструмента – избавить исследователя от рутинной работы по самостоятельному отбору групп статей для сбора, если в исходном ресурсе присутствуют ограничения на выгрузку публикаций за один раз, а также полная автоматизация процесса, не предполагающая дополнительного вмешательства в него пользователя после начала процесса выгрузки данных [10]. На этапе обработки произведена структуризация собранных данных, приведение их в используемый в дальнейшем для анализа форма. В рамках структуризации создано хранилище данных (озеро данных, см. рисунок 2), проведены рубрикация и программное насыщение данных.

Для насыщения данных ключевыми словами текста использовалась библиотека yake на языке программирования Python, она позволяет выделять ключевые слова из массива текста [11]. Первых двух этапов достаточно для проведения анализа НТЛ страны. На этапе визуализации на основе подготовленных данных осуществлена визуализация НТЛ с использованием библиотеки Plotly, которая дает возможность не только представлять данные в виде двумерных графиков, но и создавать интерактивные 3D визуализации [12]. Для отбора научных публикаций, относящихся к определенной тематике (для формирования НТЛ области), в существующую методику добавлен этап фильтрации.

Поэтапно процесс построения НТЛ области представлен на рисунке 2. Исходя из полученных данных публикаций страны и информации от экспертов (пользователей), выбирается научная область и формируются первичные ключевые слова для отбора первичной (родительской) статьи. Затем через несколько запросов в сформированное озеро данных выделяются необходимые публикации нужных авторов.

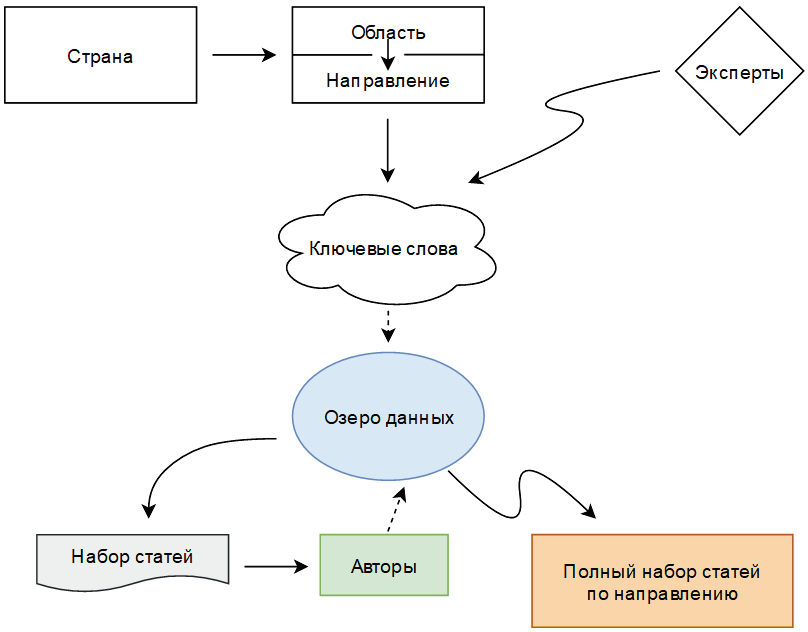

Рисунок 2 – Схема сбора и обработки данных

На первом шаге этапа фильтрации для формирования НТЛ области используются ключевые слова авторов (входящие в состав первичного набора сырых данных), а также ключевые слова из этих статей (собраны программно при помощи библиотеки yake на этапе структуризации данных). По собранным наборам ключевых слов осуществляется поиск релевантных для заданной тематики научных публикаций, выделяется набор связанных с тематикой авторов. Таким образом составляется первичный набор статей авторов, на данный момент занимающихся изучением заданной тематики. На следующем шаге производится повторный запрос в созданное хранилище для сбора всех публикаций выявленных авторов (схематично алгоритм представлен на рисунке 3). В результате формируется расширенный набор статей этих авторов за выбранных период времени как по данной, так и по смежным тематикам.

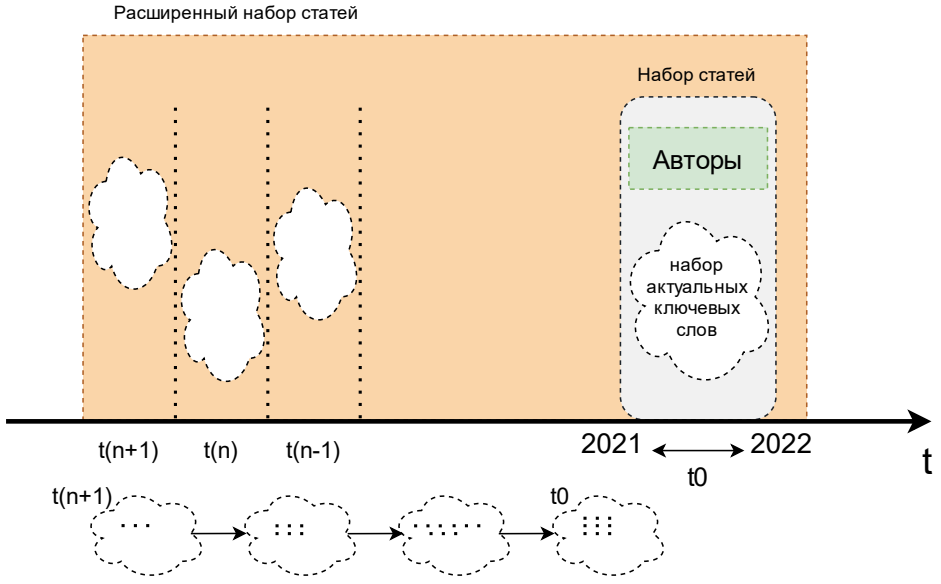

Рисунок 3 – Схема выбора научных публикаций определенной научной области на основе выделенных авторов, занимающиеся этой областью

3. Результаты и обсуждение

Для апробации разработанной методики выбраны две страны: Япония и Республика Корея. В ходе апробации создано хранилище данных. Хранилище данных по 2 странам за 2010-2021 годы состоит из JSON-файлов и занимает 40.6 ГБ памяти, из которых 24.3 ГБ (1 050 000 статей) – Япония, 16.3 ГБ (753 000 статей) – Республика Корея. Эти данные насыщались дополнительными сведениями, полученными в ходе обработки полученного текста (добавлены ключевые слова текста; произведена рубрикация, т.е. «присвоение» публикации рубрик, к которым она относится, с весами – показателями достоверности этого соотношения).

Инструмент оценки НТЛС реализован на языке программирования Python 3.8 с использованием библиотек openpyxl, selenium, bs4, plotly. Разработанный программный код может быть запущен в системах с операционной системой UNIX и Windows. Необходимым условием использования является наличие библиометрических данных научных публикаций за исследуемый период в формате, предоставляемый Scopus. Авторы планируют развивать данный инструмент в части разработки программного и графического интерфейсов и формирования Docker-образа с опубликованием на GitHub.

3.1. НТЛ страны

Для визуализации НТЛС (НТЛ страны) при помощи библиотеки plotly использовались все данные из хранилища о статьях из Японии и Республики Корея (за 2010–2021 гг.). Для построения ландшафта используются три оси:

1. по оси x – временной период (февраль 2010 г. – декабрь 2021 г.).
2. по оси y – тематика (всего тематик было выявлено 235). Принимая во внимание особенности отображения в Jupyter, все тематики одновременно на изображение вывести невозможно, поэтому автоматически была выведена лишь часть.
3. по оси z (вертикальная компонента) – количеству публикаций на каждую тему в данный период (месяц, год).

Анализ представленного графика по Японии (рисунок 4) показывает, что в период с 2010 по 2021 годы в Японии исследовались различные научные области. Из ландшафта видно, что Биомедицинские науки были самой популярной областью научного знания в Японии в период с 2010 по 2021 годы, с наибольшим количеством публикаций. Физические науки, химические науки и биологические науки также были популярными областями научного знания в Японии. Социальные науки, экономические науки и гуманитарные науки были наименее исследуемыми областями научного знания в Японии в период с 2010 по 2021 годы.

Рассмотрим НТЛ Республики Корея (рисунок 5). В последние годы Корея стала одним из ведущих мировых центров биотехнологических исследований, и многие корейские ученые внесли значительный вклад в области генетических исследований, что и отражено в полученной визуализации научно-технологического ландшафта.

Низкое количество публикаций в области гуманитарных наук может быть связано с тем, что эта область научного знания не является приоритетной для государственной политики Республики Корея. Однако описанные при анализе НТЛ Японии суждения о неоднозначности выводов про наименьший уровень популярности гуманитарных, социальных и экономических наук в стране также применимы и к Республике Корея.

При сравнении распределения публикаций в Корее и Японии, можно увидеть, что обе страны активно развивались в научных исследованиях в указанный период. В Корее наибольшее количество публикаций было в 2021 году (19 762), в Японии – тоже в 2021 году (22 980). В обеих странах этот пик приходится на область биомедицинских наук. Сравнивая обе страны, наблюдается одинаковая разреженность по некоторым областям, из чего сделан вывод о целесообразности дальнейшей фильтрации рубрик (изъятие рубрик с околонулевыми показателями на протяжении всего периода).

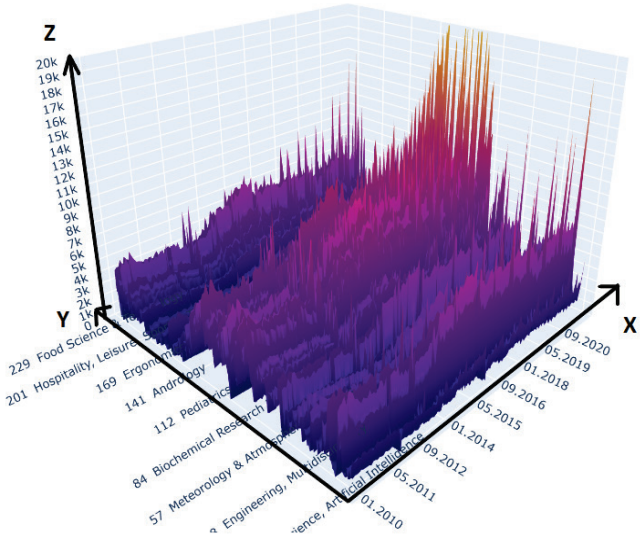

Рисунок 4 – НТЛ Японии на основе данных научных публикаций страны с февраля 2010 г. по декабрь 2021 г.

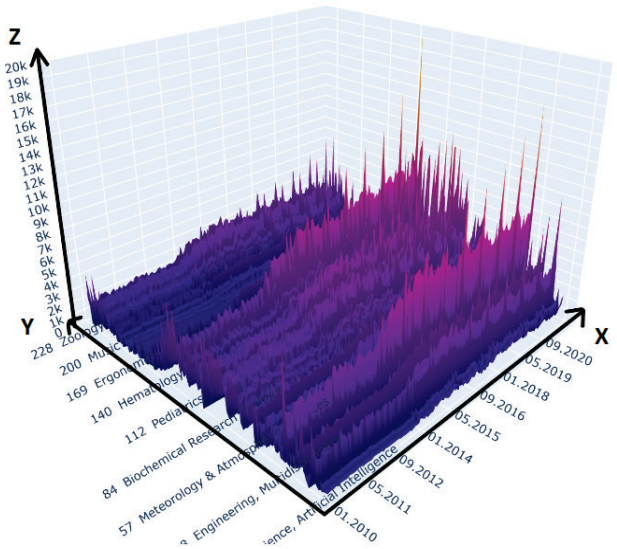

Рисунок 5 – НТЛ Республики Корея на основе данных научных публикаций страны с февраля 2010 г. по декабрь 2021 г.

Если говорить об отличиях, то главной отличительной особенностью является превосходящая почти в полтора раза интенсивность работ у Японии по сравнению с Республикой Корея. Также, несмотря на то что по большей части распределение количества публикаций по областям остается практически одинаковым в целом, существуют отдельные области научного знания, значительно различающиеся по популярности в двух странах. Так в Японии большое внимание уделяется радиологии и ядерной медицине, а также сельскохозяйственной инженерии, в то время как для Кореи эти направления являются значительно менее приоритетными. Плотное и относительно однородное (по сравнению с корейским) распределение по тематикам в Японии говорит о базовом характере проводимых исследований. В Японии дисциплины изучаются системно и методично, что нельзя сказать о Корее, где пиковый характер исследований повторяет рисунок графика Японии, однако при этом гряды значительно меньше по сравнению с японской стороной.

Если же смотреть в проекции времени, то по плотному и равномерно увеличивающемуся распределению научных публикаций как в Японии, так и в Корее можно судить о той же систематичности в исследовании областей в обеих странах, эта проекция обладает наибольшим количеством сходств.

3.2. НТЛ научной области

Разработанный инструмент апробирован на тематике «синхротронное и циклотронное излучение», а также «FLASH-терапия» [13]. Главная задача апробации – выяснить, дает ли инструмент возможность ответить на вопрос, какие области научного знания необходимо развивать, чтобы достичь подобных результатов в данной тематике. Чтобы найти все релевантные статьи по данной тематике произведена фильтрация статей из полученного из реферативных баз данных собранного хранилища данных по запросу: «FLASH-RT» OR «FLASH effect» OR «FLASH radiotherapy» OR «Radiation Therapy for Immuno-oncology» OR «External Beam Radiation and Radionuclide Therapies» OR «Radiobiology». Также набор был дополнен статьями авторов, занимающиеся этой областью (схематично выбор статей представлен на рис. 3). Для рубрикации научных публикаций использовался инструмент, разработанный в лаборатории кафедры №65 «Анализ конкурентных систем» НИЯУ МИФИ, его оценка не относится к данной работе [14].

После дополнительной обработки данных (приведение данных из структуры словаря к матричной форме для их распознания библиотекой plotly) выполнена программная визуализация данных (результат представлен на рисунке 6).

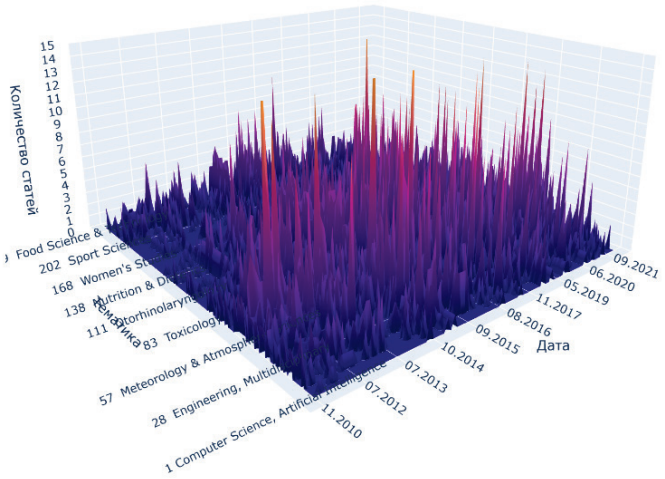

Рисунок 6 – НТЛ области «FLASH-RT» (Япония) по выбранным научным публикациям с февраля 2010 г. по декабрь 2021 г.

При анализе сформированного ландшафта можно сделать вывод о том, что в Японии при исследовании рассматриваемой тематики распределение по времени сохранялось приблизительно на одном уровне, что нельзя сказать о распределении по областям. В то время как гуманитарные науки практически не рассматривались, внимание уделялось областям, занимающимся изучением природных ресурсов, однако наибольший вклад в развитие данного направления внесены исследования в области физических технологий, медицины, инженерии и точных наук. При этом стоит отметить, что даже при значительном росте популярности технологий машинного обучения и использования ИИ в исследованиях, данная область в изучении тематики FLASH-терапии не претерпела изменений по временной оси и не являлась приоритетной областью исследований за выбранный период.

Также одним из использованных инструментов является представление ключевых слов в виде облака. Облака ключевых слов – инструмент, который используется для визуального представления и анализа ключевых слов, используемых в научных публикациях [15]. Облака ключевых слов позволяют быстро определить наиболее часто используемые ключевые слова в тексте и оценить, какие темы и области научного знания обсуждаются в публикации. Используя облака ключевых слов, можно определить, какие ключевые слова наиболее часто встречаются в тексте, и как они связаны между собой. Рассмотрим облака ключевых слов авторов Японии по тематике «FLASH-RT», представленные на рисунке 7.

Рисунок 7 – Облака ключевых слов авторов Японии по годам по тематике "FLASH-RT" по выбранным научным публикациям с февраля 2010 г. по декабрь 2021 г.

В 2010 году набор ключевых слов разнородный, приоритетность различных направлений примерно одинаковая, однако большинство данных ключевых слов находятся в области исследования патологий, генетических мутаций и, как следствие, методов их лечения, например – ПЭТ (позитронно-эмиссионная томография). В 2012 году, судя по набору популярных

ключевых слов, акцент исследований смещен на изучение видов излучений, «излучающих систем» и их влияния на организм, «гибель клеток». В 2013 году популярность набрали такие ключевые слова как «атомная бомба», «выжившие в ядерных бомбардировках», начали вестись исследования в области «брахитерапии». В 2016 году начали вестись исследования в области «углеродно-ионной терапии», а также «повреждения» и «репарации» ДНК. В 2018 году наблюдается рост по следующим ключевым словам: «рак шейки матки», «углеродно-ионной терапии» и «брахитерапия».

Таким образом облака тэгов являются важным инструментом в оценке научно-технологического потенциала страны в области, они позволяют взглянуть на общую картинку под другим углом, посмотреть на содержание изучаемых в стране областей через наиболее упоминаемые слова/выражения, что является важным дополнением к анализу НТЛ.

4. Заключение

В работе представлен инструмент оценки НТЛ. Для этого были разработаны методы построения и представления данных, необходимые для анализа НТЛ как в масштабе определенной научной области, так и страны. Произведена апробация обоих методов построения и визуализации НТЛ (Японии и Республики Корея, а также области «FLASH-RT» для Японии). Отличием данной работы от рассмотренных в процессе обзора литературы публикаций является наличие и способ представления данных в виде визуализации (графика-ландшафта), позволяющего рассмотреть большие данные по НТЛ в удобном для анализа формате.

Разработанный инструмент может быть интересен исследователям в различных областях, ученым, аналитикам. Еще одним достоинством является его независимость от фиксированных рубрик, (например, таких, какие представлены в Scopus, Web of Science). Однако на данный момент его недостатком является использование только научных публикаций в качестве данных, ограниченных еще и базой данных исходного источника данных, а также отсутствие графического интерфейса взаимодействия с пользователем.

В дальнейшем развитии инструмента планируется добавление новых входных данных (например, объекты интеллектуальной собственности), разработка пользовательского интерфейса для выбора данных, выбора параметров фильтрации и отображения результатов.

5. Список источников

[1] Рычков Д.А., Орлова Е.Е., Сурьева О.А. [и др.]. Программная реализация сбора, хранения и обработки информации о режущих инструментах // Механики XXI веку, No. 14, 2015. pp. 122-126.

[2] Макарова И.В., Максимов А.Д.. Методология оценки потенциала модернизации промышленного комплекса // Журнал экономической теории, Vol. №4, 2011.

[3] Giovanni Abramo, Ciriaco Andrea D'Angelo. novel methodology to assess the scientific standing of nations at field level // Journal of Informetrics, Vol. 14, No. 100986, 2020.

[4] Yuya Kajikawa, Cristian Mejia, Mengjia Wu, Yi Zhang. Academic landscape of Technological Forecasting and Social Change through citation network and topic analyses // Technological Forecasting & Social Change, Vol. 182, No. 121877, 2022.

[5] Бастрикина В.В.. Проектирование веб-скрапера для получения данных с сайтов книжных издательств // Актуальные проблемы авиации и космонавтики., Vol. №14, 2018.

[6] Куракова Н.Г., Цветкова Л.А., Зинов В.Г., ПАТЕНТНЫЙ ЛАНДШАФТ РФ, СОЗДАННЫЙ РЕЗИДЕНТАМИ СТРАНЫ: АНАЛИЗ ВЫЯВЛЕННЫХ ПРОБЛЕМ // ЭКОНОМИКА НАУКИ, Vol. 2, No. 1, 2016.

[7] Antonov, E., Lopatina, E., Ionkina, K. Tretyakov, E. Agent data merging //Procedia Computer Science. – 2020. – Т. 169. – С. 473-478.

[8] Antonov, E. V., Artamonov, A. A., Rudik, A. V., Malugin, M. I. Trend Visualization of Academic Field: Proposed Method and Big Data Review // Scientific Visualization, 2022, volume 14, number 2, p. 62 – 76.

[9] Зубрилина Т.В., Юрьев В.Н., Базы данных. Проектирование реляционных баз и хранилищ данных с использованием CASE-технологий: учебное пособие. Санкт-Петербург: Изд-во Политехнического ун-та, 2007. 43 pp.
[10] Антонов Е.В., Артамонов А.А., Орлов А.В., Николаев В.С., Захаров В.П., Хохлова М.В., Концевая Ю.М., Бонарцев А.П., Воинова В.В.Обработка научно-технической информации в междисциплинарных исследованиях методами математико-лингвистического направления поиска на примере области изучения биоматериалов для тканевой инженерии // International Journal of Open Information Technologies, No. №11, 2022.
[11] R. Campos, A. Jorge, C. Nunes [et al.], YAKE! Keyword extraction from single documents using multiple local features// Information Sciences. – 2020. – Vol. 509. – P. 257-289. – DOI 10.1016/j.ins.2019.09.013. – EDN RLBHLY.
[12] Филина Е.В., Моделирование и визуализация данных на языке программирования Python с помощью библиотеки Plotly в различных областях знаний / Филина Е.В. // Вестник Саратовского областного института развития образования. – 2020. – № 1(21). – С. 81-88. – EDN VVSIQN
[13] Карташев Артем Владимирович, Бочкарева Татьяна Николаевна, Анохина Анастасия Сергеевна. FLASH-терапия: перспективное направление в борьбе с опухолью // BPP, No. №4, 2021.
[14] Onykiy, B., Antonov, E., Artamonov, A., Tretyakov, E., 2020. Information Analysis Support for Decision-Making in Scientific and Technological Development. International Journal of Technology. Volume 11(6), pp. 1125-1135
[15] Тимофеев М.В. Применение нормирования данных для последующей визуализации методом «облако тегов» / М.В. Тимофеев, А.В. Мазалькова // Научный аспект. – 2023. – Т. 2, № 4. – С. 147-154. – EDN UHUUGP.