УДК 025 DOI: 10.31799/2077-5687-2021-4-57-67

ИССЛЕДОВАНИЕ НАУКОМЕТРИЧЕСКИХ ДАННЫХ В ОБЛАСТИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

М. С. Прокофьева
Санкт-Петербургский государственный университет аэрокосмического приборостроения

Исходя из современных тенденций, можно заметить острую необходимость в более глубоком освоении сфер, относящихся к искусственному интеллекту. Главным образом для сведения к минимуму влияния человеческого фактора. К сожалению, представить масштаб интереса к данной тематике крайне сложно, в связи с почти полным отсутствием визуальных представлений. Этот факт в совокупности с огромным количеством информации может вызвать затруднение не только в изучении, но и в поиске материалов. В связи с этим существует необходимость в анализе данных при помощи программы VOSVIEWER, которая способна создавать библиографические сети, представляющие визуализацию, ранее полученных, библиографических списков, и пакета BIBLIOMETRIX, позволяющего произвести количественный и статистический анализ статьей с целью определения цитируемости и публикационной активности. Материалы, подвергнутые исследованию, будет выгружены из крупнейшей библиографической и реферативной базы данных Scopus.

Ключевые слова: Глубокое обучение, машинное обучение, нечеткая логика, метод оптимизации, поверхностное обучение, нейросети, искусственный интеллект.

Для цитирования:
Прокофьева М. С. Исследование наукометрических данных в области искусственного интеллекта // Системный анализ и логистика: журнал.: выпуск №4(30), ISSN 2077-5687. – СПб.: ГУАП., 2021 – с 57-67. РИНЦ. DOI: 10.31799/2077-5687-2021-4-57-67.

RESEARCH OF SCIENTIFIC DATA IN THE FIELD OF ARTIFICIAL INTELLIGENCE

M. S. Prokofieva
St. Petersburg State University of Aerospace Instrumentation

Based on current trends, you can see the need for a deeper development of the field related to artificial intelligence. Mainly for information to minimize the occurrence of human error. Today there are many publications that confirm what was said in the above thesis. Unfortunately, it is extremely difficult to imagine the scale of interest in this topic, due to the almost complete absence of visual representations. This fact, combined with a huge amount of information, can cause difficulties not only in the study, but also in the search for materials. In this regard, there is a need for data analysis using the VOSVIEWER program, which is able to create bibliographic networks, representing the visualization of bibliographic lists, and the BIBLIOMETRIX package, which allows for quantitative analysis and statistics of articles in order to find their citation and publication activity. The analyzed materials will be downloaded from the largest bibliographic and abstract database Scopus.

Keywords: Deep learning, machine learning, fuzzy logic, optimization method, surface learning, neural networks, artificial intelligence.

For citation:
Prokofieva M. S. Research of scientometric data in the field of artificial intelligence // Systems analysis and logistics: №4(30), ISSN 2077-5687. – Russia, Saint-Petersburg.: SUAI., 2021 – p. 57– 67. DOI: 10.31799/2077-5687-2021-4-57-67.

Введение

Наблюдая за бурным технологическим развитием общества, можно заметить потребность в освоении областей, связанных с искусственным интеллектом (далее ИИ). Такой интерес объясняется крайней необходимостью в быстром, эффективном и, исключающем фатальные ошибки, поиске решений для разнообразного рода задач. Однако из-за обширности направлений ИИ, возникают трудности в нахождении актуальных, современных и самое главное достоверных материалов по данному направлению. В связи с этим отмечается спрос в структурировании рассеянной по сети INTERNET информации.

Для произведения анализа будет применяться программа VOSVIEWER, которая способна создавать библиографические сети, представляющие визуализацию

библиографических списков, и пакета BIBLIOMETRIX, позволяющего произвести количественный и статистический анализ статьей, с целью определения их цитируемости и публикационной активности.

Материалы, подвергнутые анализу, выгружаются из крупнейшей библиографической и реферативной базы Scopus в виде файлов с расширениями .csv – для работы с VOSVIEWER и .bib – для работы с BIBLIOMETRIX.

Составление исходных данных и подбор инструментария

Как говорилось выше, достоверность и актуальность публикаций лучше всего определяется путем использования базы Scopus. Однако, полученные из нее файлы практически не восприимчивы для анализа данных, следовательно, производить проверку вручную или крайне сложно, или вовсе невозможно. Именно поэтому необходимо использовать дополнительный инструментарий – VOSVIEWER и BIBLIOMETRIX.

Для корректного составления файлов необходимо определиться с набором ключевых слов, так как это самый простой способ нахождения публикаций:

1. Deep learning (глубокое обучение);
2. Machine learning (машинное обучение);
3. Fuzzy logic (нечеткая логика);
4. Optimization method (метод оптимизации);
5. Surface learning (поверхностное обучение);
6. Neural networks (нейросети);
7. Artificial intelligence (искусственный интеллект).

Данный перечень состоит из самых часто встречаемых поисковых интернет запросов по тематике ИИ.

VOSVIEWER и BIBLIOMETRIX имеют возможность структурирования данных как и по общей выборке, так и отдельно по каждому ключевому слову. Для удобства, в данной статье будет использоваться общая выборка по всем ключевым словам.

Так же исследованию подвергнуться, самые часто печатающиеся, авторы.

Для получения необходимых результатов следует разделить анализ на количественную (оценочную) и визуальную составляющую.

К количественной составляющей будут относиться параметры:

1. Индекс Хирша (характеристика продуктивности учёного, основанная на количестве его публикаций и количестве цитирований этих публикаций);
2. Количество публикаций в год и за все время;
3. Среднее число цитирований за документ;
4. Таблицы выборок по журналам, статьям и авторам.
5. К визуальной составляющей будут относиться такие представления как:
6. Графики скорости роста публикаций;
7. Библиографические сети;
8. Дерево публикационной активности;

Необходимо обратить внимание, что такие параметры и визуальные представления как индекс Хирша, количество публикаций в год и за все время, среднее число цитирований за документ, скорость роста публикаций, дерево публикационной активности, относятся к пакету BIBLIOMETRIX.

Однако хоть больше половины указанных параметров находятся при помощи BIBLIOMETRIX, библиографические сети и таблицы выборок лучше всего составлять в VOSVIEWER, так как данная программа способна значительно упросить и ускорить процесс построения.

Работа с данными в программе VOSVIEWER

После загрузки данных можно заметить, что программа импортировала из файла гораздо больше ключевых слов, чем было представлено ранее. Это можно объяснить тем, что при выгрузке из Scopus берутся все перечисленные в статьях термины. Следовательно, необходимо исключить из статистики то, что встречается реже всего. Данная проблема решается в самой программе. VOSVIEWER анализирует файл и предлагает ввести порог встречаемости ключевых слов для удаления из анализа совсем редких терминов. Ограничимся анализом тех ключевых слов, которые встречаются минимум 20 раз (см. Рис. 1).

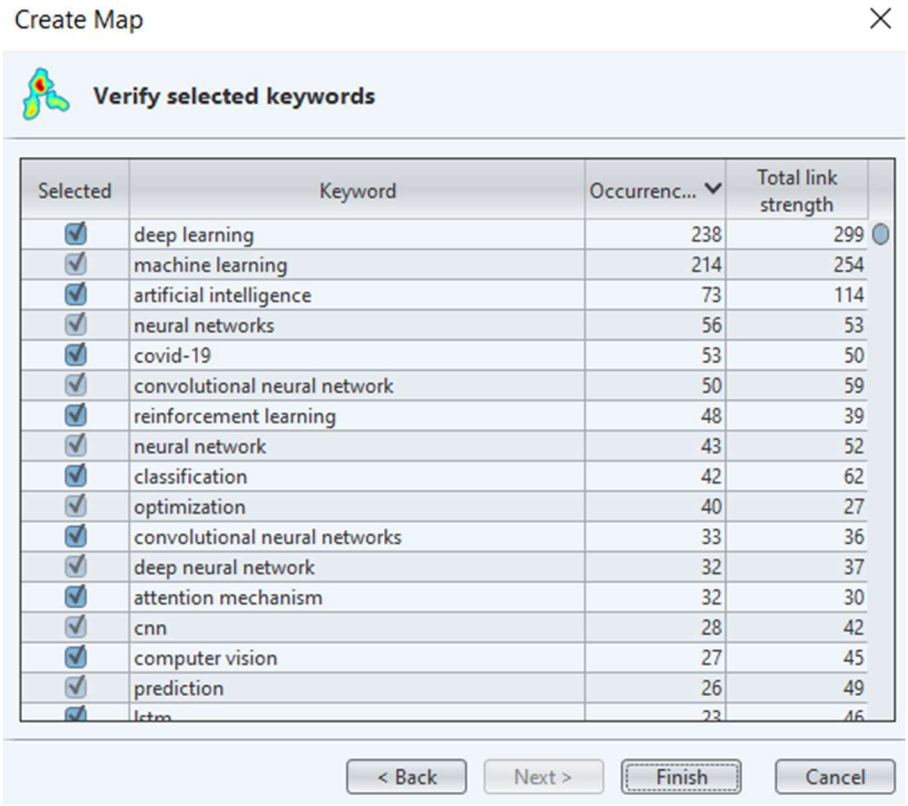

Рисунок 1 – выборка по ключевым словам

VOSVIEWER может построить наукометрические сети и карты в различных режимах:

1. Network Visualization (визуализация сети):
2. Overlay Visualization (визуализация наложения);
3. Density Visualization (визуализация плотности) [1, 2, 3].

В данной статье будет использоваться только представление Network Visualization, так как оставшиеся режимы демонстрируют те же самые результаты, только отображают их несколько иначе.

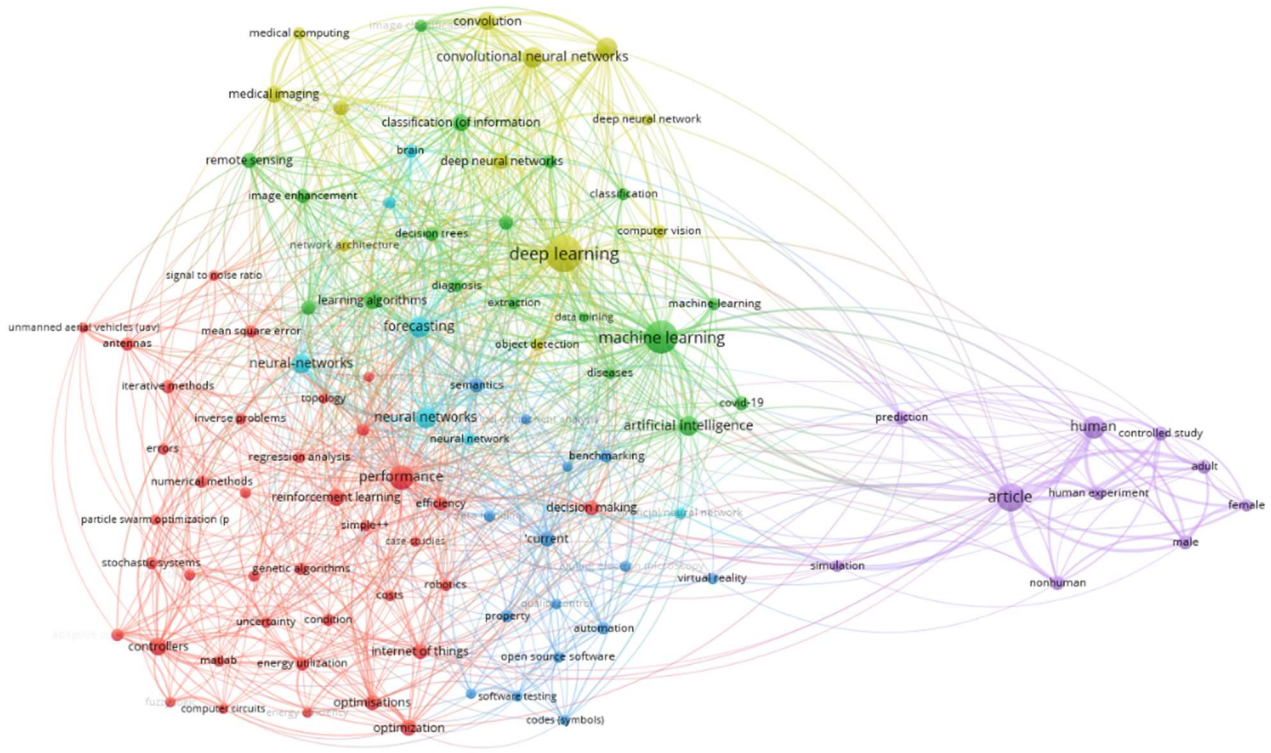

Рисунок 2 – Network Visualization (визуализация сети)

Каждый элемент в сетевой визуализации представлен своей меткой и кружком (см. Рис. 2). Размер метки и окружности элемента определяется его весом. Чем выше вес элемента, тем крупнее этикетка и кружок [1, 2].

Для некоторых элементов метка может не отображаться. Это сделано для того, чтобы метки не накладывались друг на друга. Цвет элемента определяется кластером, к которому он принадлежит [3].

В данном случае можно заметить принадлежность к разным направлениям в области изучения и применения искусственного интеллекта:

1. зеленый цвет – разновидности методов машинного обучения для решения задач в сфере здравоохранения;
2. красный цвет – чрезвычайные ситуации, безопасность;
3. голубой цвет – область прогнозирования и автоматизации;
4. фиолетовый – различные социальные сферы;
5. желтый цвет –применение методов глубокого обучения в области здравоохранения и изучения нейронных сетей.

Строки между элементами представляют собой ссылки. По умолчанию отображается не более 1000 строк, представляющих 1000 самых сильных связей между элементами, в данном случае количество строк было сокращено до 100 [1]. Расстояние между двумя журналами в визуализации приблизительно указывает на родство журналов с точки зрения ссылок совместного цитирования.

Дополнительно следует составить наукометрическую картину, по самым часто публикующимся, авторам и журналам (см. Рис. 3 – 5).

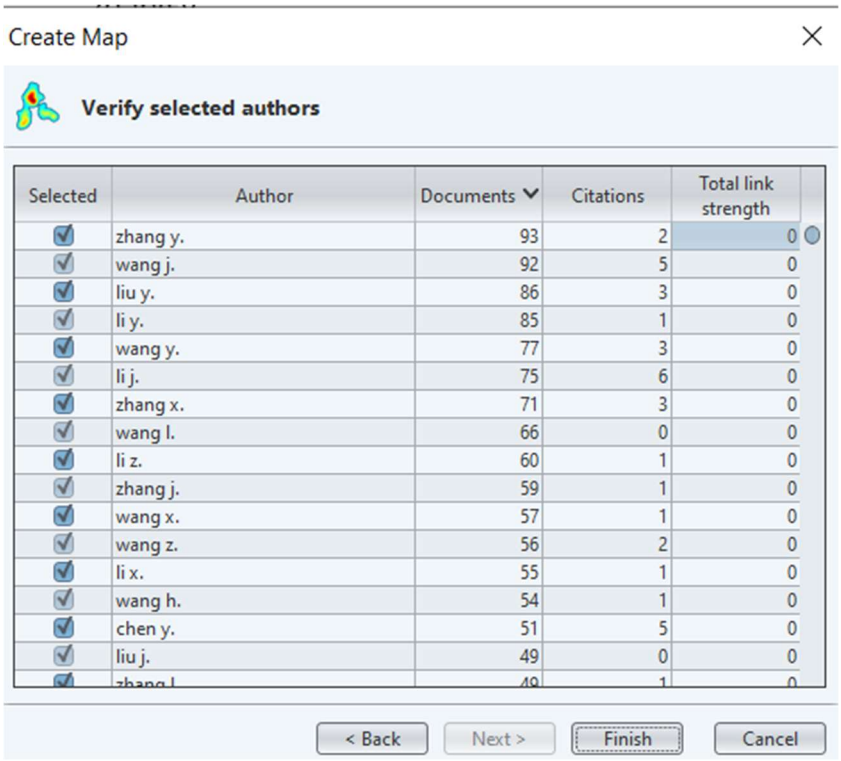

Рисунок 3 – выборка по авторам

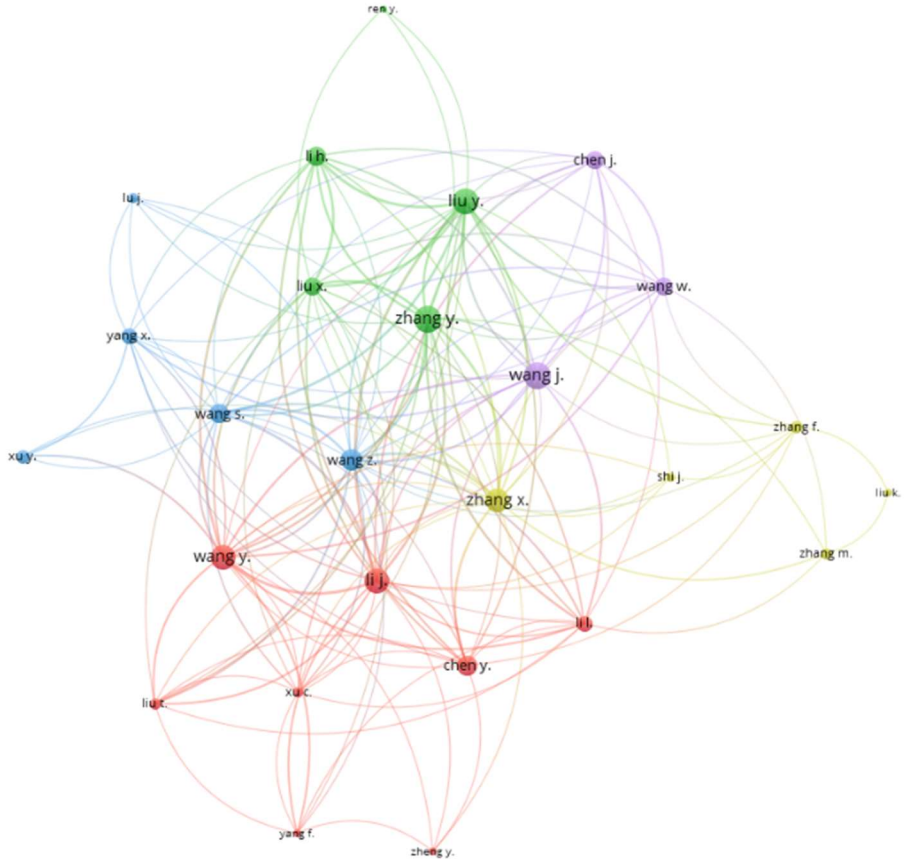

Рисунок 4 – Network Visualization (визуализация сети) для общей выборки по авторам

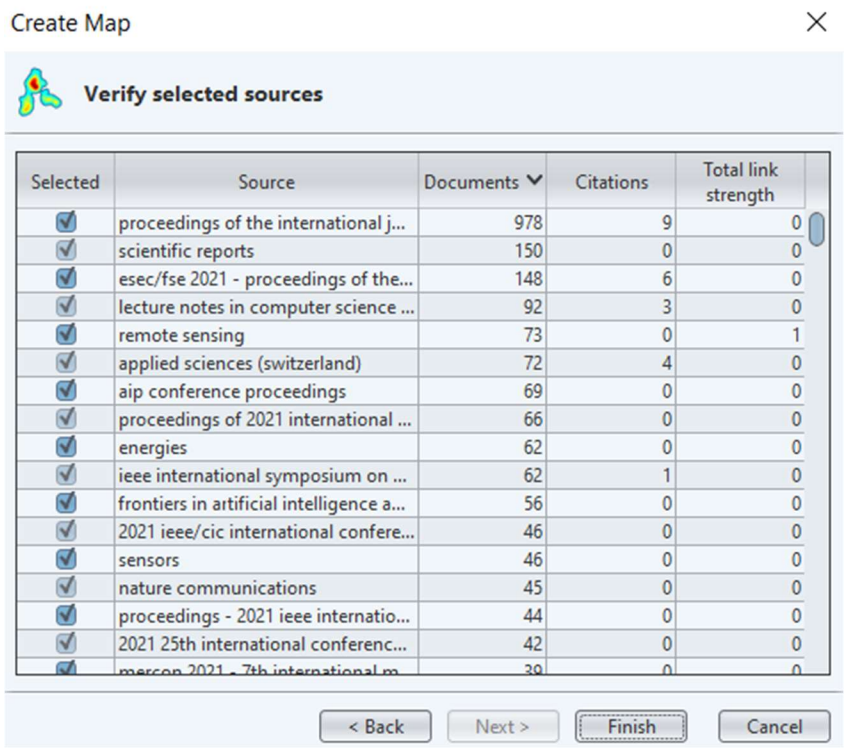

Рисунок 5 – выборка самых часто публикующихся, в данной области, журналов и конференций

Работа с данными при помощи пакета BIBLIOMETRIX

Для работы с пакетом BIBLIOMETRIX сначала следует создать проект в системе с открытым исходным кодом rStudio. Затем ввести скрипт:

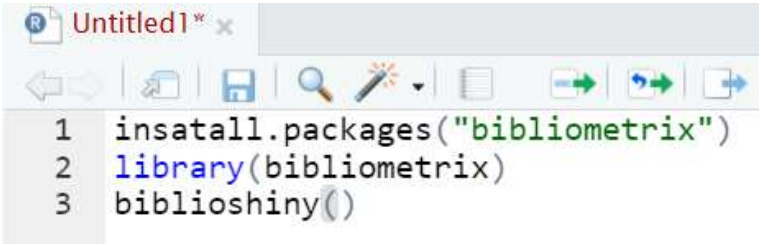

Рисунок 6 - Скрипт bibliometrix

После чего открывается инструмент для анализа – BIBLIOSHINY, который схож с программой VOSVIEWER с тем лишь отличием, что BIBLIOSHINY имеет больше возможностей, но при этом их выполнение отнимает слишком много времени. Тем не менее данный инструмент лучше себя показывает в определении количественных характеристик.

Для получения требуемых результатов следует импортировать файл Scopus, в установившийся BIBLIOSHINY.

Определим значение среднего цитирования на документ (см. Рис 7)

Description	Results
MAIN INFORMATION ABOUT DATA	
Timespan	2018:2021
Sources (Journals, Books, etc)	520
Documents	1265
Average years from publication	0.0561
Average citations per documents	0.03794
Average citations per year per doc	0.028
References	46436
DOCUMENT TYPES	
article	791
book	6
book chapter	14
conference paper	383
conference review	35
erratum	2
note	2
review	32

Рисунок 7 – Значение среднего числа цитирования за документ

Значения для среднего цитирования в течение года, в данной области, = 0,0561, такое возможно объяснить всплеском недавней активности.

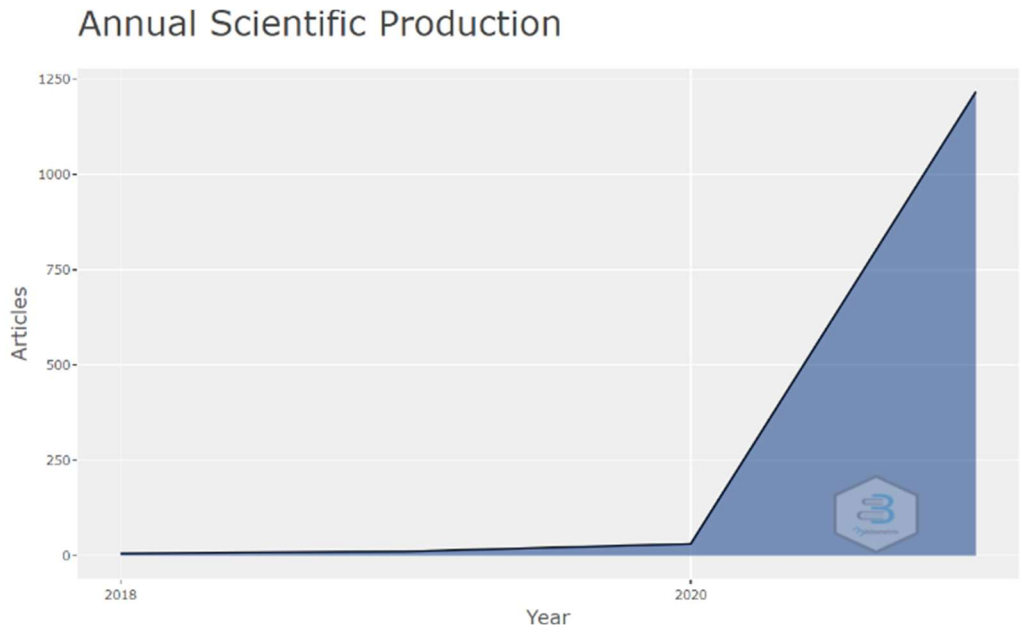

Рисунок 8 – Визуализация значение среднего числа цитирований по годам

Plot Table

Show 20 rows Copy CSV Excel PDF Print Search:

Year	Articles
2018	6
2019	11
2020	31
2021	1217

Showing 1 to 4 of 4 entries Previous 1 Next

Рисунок 9 – Табличный вид значение среднего числа цитирований по годам scopus

Исходя из графика, представленного на рисунке 8, можно заметить, что интерес к теме ИИ находится на пике, так как количество исследований от года к году растёт. Особенный всплеск проявляется в 2021 г. Это лучше всего отображено на рисунке 9, где видно, что количество цитирований в сравнении с 2020 г. увеличилось примерно в 39 раз.

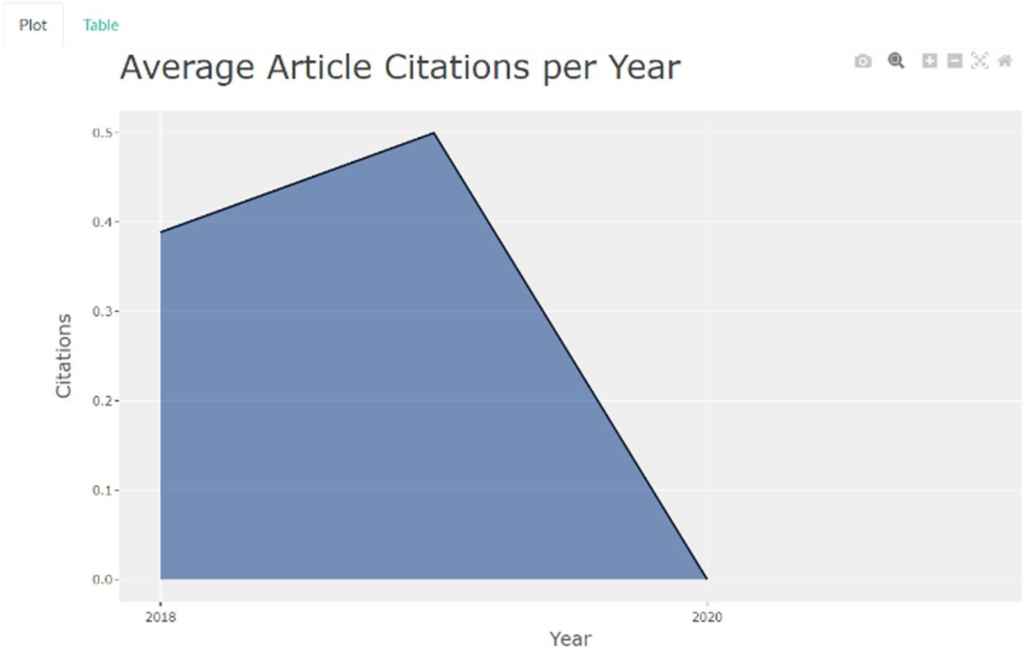

Рисунок 10 – Визуализация значение среднего числа цитирований в год

Plot Table

Show 20 rows Copy CSV Excel PDF Print Search:

Year	N	MeanTCperArt	MeanTCperYear	CitableYears
2018	6	1.16666666666667	0.388888888888889	3
2019	11	1	0.5	2
2020	31	0	0	1
2021	1217	0.0246507806080526		0

Showing 1 to 4 of 4 entries Previous 1 Next

Рисунок 11 – Табличный вид значение среднего числа цитирований в год

При рассмотрении графика на рисунке 10, наблюдается спад области исследования ИИ, который, опять же, можно объяснить большим количеством недавно выпущенных публикаций.

Рассмотрим данные по индексу Хирша, из рисунка ниже видно, что, с учетом выборки (20 первых авторов в топе), значения у всех авторов = 1, это может свидетельствовать о большем интересе к данной области среди молодых специалистов.

Plot Table

Show 20 rows Copy CSV Excel PDF Print Search:

Element	h_index	g_index	m_index	TC	NP	PY_start
ATKINSON TM	1	1	1.000	6	1	2021
BELLAVIA S	1	1	1.000	5	1	2021
BENVENGO S	1	1	1.000	6	1	2021
CAHLON O	1	1	1.000	6	1	2021
CARTER J	1	1	1.000	1	1	2021
CHA E	1	1	1.000	6	1	2021
CHINO F	1	1	1.000	6	1	2021
CHOI J	1	1	1.000	1	1	2021
CHOI W	1	1	1.000	1	1	2021
FRASER G	1	1	1.000	1	1	2021
GARRIDO S	1	1	1.000	1	1	2021
GILLESPIE EF	1	1	1.000	6	1	2021
GOMEZ DR	1	1	1.000	6	1	2021
GURIOLI G	1	1	1.000	5	1	2021

Рисунок 12 – Табличный вид значения индекса Хирша

Далее для лучшей визуализации связи журналов, авторов и ключевых слов, было построено дерево публикационной активности, которое отображает связи между данными параметрами.

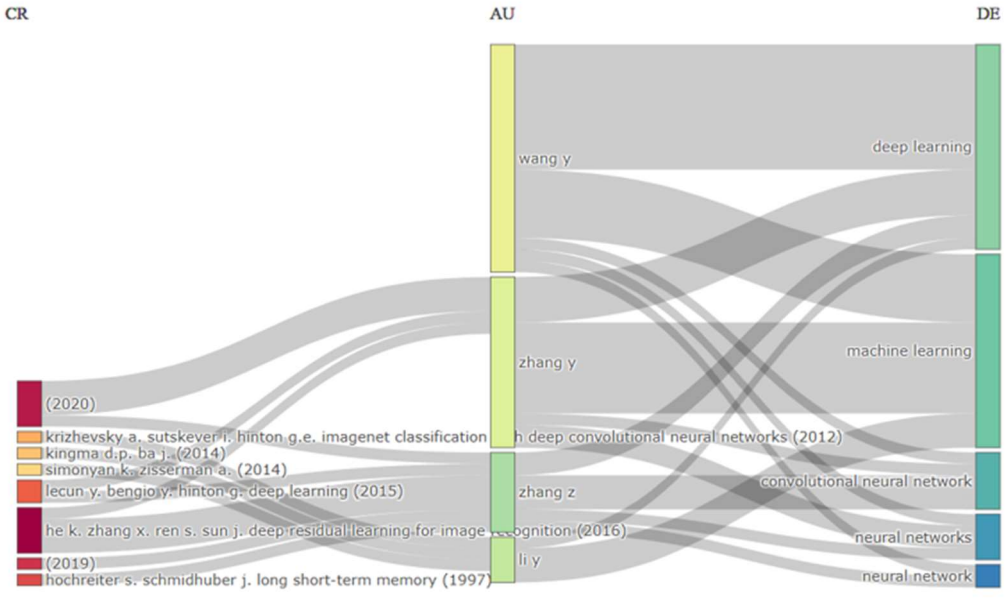

Рисунок 13 – Дерево публикационной активности авторов

На рис. 13.: CR – использованная литература, AU – авторы, DE – ключевые слова.

Результаты, показанные на рисунке 13, коррелируются с выборкой из программы VOSVIEWER (см. Рис. 1, 3, 5), с теми лишь отличиями, что в дереве публикационной активности отображаются связи между журналами, авторами и ключевыми словами, а в программе VOSVIEWER такие связи не учитываются и данные строятся по каждому критерию без какого либо влияния из вне.

Результаты анализа

Таким образом, при помощи программы VOSVIEWER и пакета BIBLIOMETRIX были:

1. построены библиографические сети, с помощью которых были ключевые слова, имеющие самый большой вес в направлении ИИ:

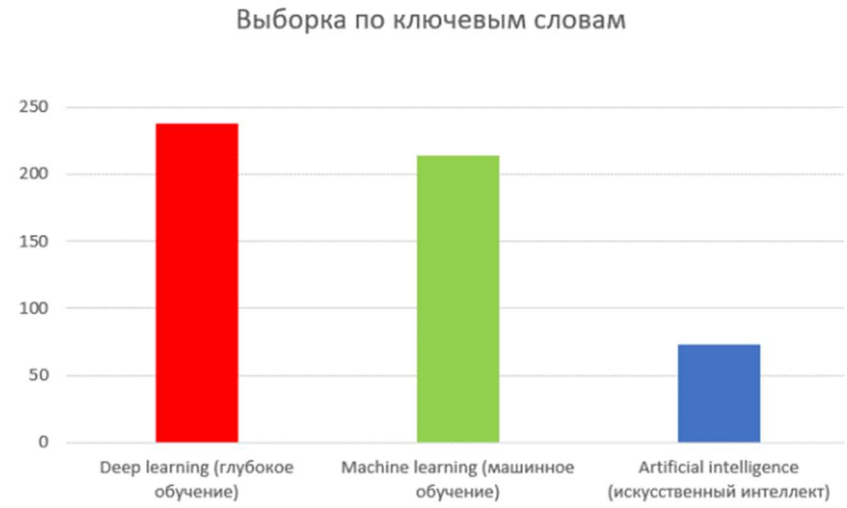

Рисунок 14 – выборка по самым часто встречающимся ключевым словам

2. определены, самые часто публикующиеся в данной тематике, авторы:

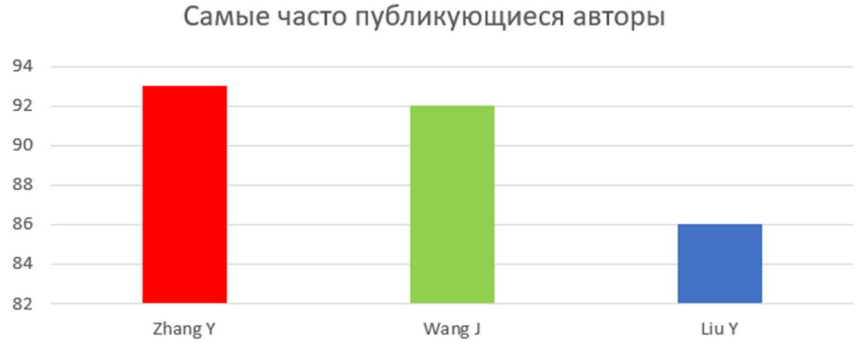

Рисунок 15 – самые часто публикующиеся авторы

3. найдены научных журналов и конференций, специализирующиеся в изучение исследуемой области:

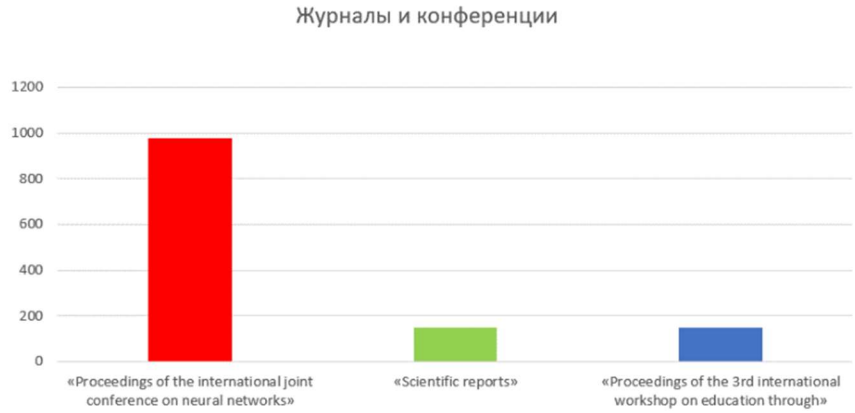

Рисунок 16 – научные журналы и конференции, публикующие статьи в области ИИ

Заключение

Выявлены наиболее важные сферы исследования:

- Разновидности методов машинного обучения для решения задач в сфере здравоохранения;
- Чрезвычайные ситуации, безопасность;
- Применение методов глубокого обучения в области здравоохранения и изучения нейронных сетей.

Найдено значение среднего цитирования за документ = 0,0561 (Рис. 7). Показаны визуализации среднего числа цитирований в год и за все время (см. Рис. 8, 9). Определен индекс Хирша, равный во всех случает единице (см. Рис. 12). Построено дерево публикационной активности авторов, где отслеживается параллель с полученными результатами в программе VOSVIEWER (см. Рис. 13).

СПИСОК ЛИТЕРАТУРЫ

1. De Nooy, W., Mrvar, A., & Batagelj, V. (2011). Exploratory social network analysis with Pajek (2nd ed.). Cambridge University Press.
2. Newman, M.E.J. (2004). Fast algorithm for detecting community structure in networks. Physical Review E, 69, 066133.
3. Noack, A. (2007). Energy models for graph clustering. Journal of Graph Algorithms and Applications, 11(2), 453–480.

ИНФОРМАЦИЯ ОБ АВТОРЕ

Прокофьева Марина Сергеевна –
магистр
Санкт-Петербургский государственный университет аэрокосмического приборостроения
190000, Россия, Санкт-Петербург, ул. Большая Морская, д. 67, лит. А
E-mail: m4riprokofjeva@yandex.ru

INFORMATION ABOUT THE AUTHOR

Prokofeva Marina Sergeevna –
master
Saint-Petersburg State University of Aerospace Instrumentation
SUAI, 67, Bolshaya Morskaia str., Saint-Petersburg, 190000, Russia
E-mail: m4riprokofjeva@yandex.ru