4-2021

BIBLIOMETRIC ANALYSIS OF NAMED ENTITY RECOGNITION FOR CHEMOINFORMATICS AND BIOMEDICAL INFORMATION EXTRACTION OF OVARIAN CANCER

Vijayshri Khedkar
Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India,
vijayshri.khedkar@sitpune.edu.in

Charlotte Fernandes
Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India,
fernandes.charlotte.btech2018@sitpune.edu.in

Devshi Desai
Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India,
devshi.desai.btech2018@sitpune.edu.in

Mansi R
Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India,
mansi.r.btech2018@sitpune.edu.in

Gurunath Chavan Dr
Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, India,
gurunath.chavan@sitpune.edu.in

Authors

Vijayshri Khedkar, Charlotte Fernandes, Devshi Desai, Mansi R, Gurunath Chavan Dr, Sonali Tidke Dr., and M. Karthikeyan Dr.

BIBLIOMETRIC ANALYSIS OF NAMED ENTITY RECOGNITION FOR CHEMOINFORMATICS AND BIOMEDICAL INFORMATION EXTRACTION OF OVARIAN CANCER

Vijayshri N. Khedkar[a]*, Charlotte Fernandes[a], Devshi Desai[a], Mansi R[a]. Dr. Gurunath Chavan[a] , Dr. Sonali (Kothari) Tidke[a], Dr. M. Karthikeyan[b]

[a] Department of Computer Science & IT, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune, Maharashtra.

[b]Principal Scientist, National Chemical Laboratory, (CSIR) Pune, India

Abstract

With the massive amount of data that has been generated in the form of unstructured text documents, Biomedical Named Entity Recognition (BioNER) is becoming increasingly important in the field of biomedical research. Since currently there does not exist any automatic archiving of the obtained results, a lot of this information remains hidden in the textual details and is not easily accessible for further analysis. Hence, text mining methods and natural language processing techniques are used for the extraction of information from such publications.Named entity recognition, is a subtask that comes under information extraction that focuses on finding and categorizing specific entities in text.

In this paper, bibliometric analysis of named entity recognition of ovarian cancer is carried out using information about publications from Scopus. The most productive journals, countries and authors are determined. The most frequently cited article and its citation history has been described. Also bibliometric maps based on citation network among countries are constructed. This study can assist people in the medical field to get a comprehensive understanding of the study of BioNER. It can also be utilized for reference works, for the research and application of the BioNER visualization methods.

Keywords— text mining; deep learning ;natural language processing; named entity recognition ; Ovarian cancer; Cancer ; bibliometric map;information extraction.

I. INTRODUCTION

The rapid rise in theamount of bioinformatics articles and resources, has led to increased challenges in searching for and extracting valuable information. Researchers take into consideration numerous sources of information and then proceed to filtering out the important information by transforming the text data to help in research productivity [2]. However, the manual annotation and feature generation by biomedical experts has proved to be inefficient because it involves a complex process and requires costly and time-consuming labour. Natural language processing plays a vital role in modern medicine. More specifically, one of the prime components in biomedical text mining is Biomedical Natural Language Processing (BioNLP), allowing computers to extract information related to biomedical topics from unstructured texts.In BioNLP, the name of biological entities mentioned in texts need to be identified automatically.Hence Named Entity Recognition is used.

NER is used to locate references to entities in texts and classifies them into predefined categories. NER is an important component of any text mining application and has proven relevant in biomedical applications , where the categories of entities to be extracted include pharmacological substances, chemical terms, drug-related information such as dosages or adverse events , diseases, problems, tests and treatments, or genes and proteins, among others.

Biomedical NER is complex due to the fact that biological entities: (a) comprise of characters, symbols, and punctuations (b) have a verbose phrase description, (c) increase continuouslyas new discoveries are made, (d) referred mostly to by their

abbreviations, and(e) have a huge number of equivalent words [2].

II. METHODS

An extensive search was carried out using Scopus for articles written in English language. We recognized search keywords from The British Standards Institution glossaries. The list of keywords used is as follows:

Scopus (articles, reviews, all years):

TITLE-ABS-KEY ("Named Entity Recognition" OR "Ovarian Cancer") AND (LIMIT-TO (OA , "all")) AND (LIMIT-TO (SUBJAREA , "COMP")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "re"))

The analysis focused on three aspects. Firstly, using a statistics analysis method, we drew the literature distribution characteristics. Secondly, for constructing and visualizing bibliometric networks, a software tool, VOSviewer was used along with the "Analyze search results" feature provided by Scopus. Finally, the information extracted from Scopus database was sorted using Excel sheet to recognize the most influential author citations, index keywords and organization citations.

Scopus:

Method we used to obtain appropriate papers from Scopus to conduct our bibliometric analysis [4]:

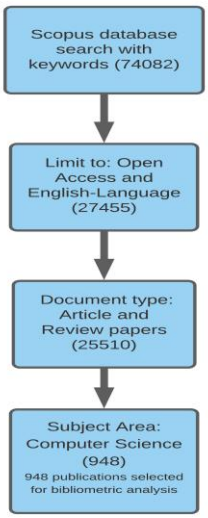

Fig 1.Structured method for literature selection

2.a. Data Collection

The process shown in Fig 1 was used for acquiring the publications from the Scopus database which were relevant to our study. The prime reason behind using the Scopus databaseis being a renowned index that covers a large number of peer-reviewed publications and provides authentic data which can be utilized for bibliometric analysis [1]. A search query with pertinent keywords was utilized to find the title, abstract, and keywords of the journals in the database from 2000 to 2021. Initial query included two keywords, Named Entity Recognition and Ovarian Cancer, which resulted in 74,082 number of documents. We limited our search to open access only, in English language, resulting in 27,455 publications. Further, the search was limited to document type(articles and review papers), which generated 25,510 number of papers. With further refinement, a limit on the subject area to only Computer Science papers, led to generation of papers which were more domain specific, i.e. 948 documents.

2.b. Data Analysis

Co-citation analysis was performed in order to obtain the base literatures of the respective subject area. Co-word analysis was done on the keywords in order to determine the theoretical structure and research topics of the subject area. VOSviewer software (version 1.6.16) was used to carry out the bibliometric analysis [6]. An important preprocessing step .i.e. Data cleaning using a thesaurus file was performed before data analysis. The thesaurus file was used to mergedifferent versions of the same word or concepts like "NER", "Named Entity Recognition", "named entity classification", "information extraction",etc which were all merged into "Named Entity Recognition." The initial step of the study produced a chart displaying the number of articles published every year to show the rise in the literature. The co-citation analysis to extract the core documents was performed after a study of important journals was done. Lastly, in order to recognize the vital research areas, eliminate duplicates through a crucial study of keywords and evolution trends in the research area a co-word analysis was performed. The documents utilized for analysis were of two types,

out of which. 12% are review papers and the rest 88% are Articles on NER, Ovarian Cancer or both.

2.c. General Statistical Literature Trends

As shown in Figure2,there is an increase in the number of documents in the study span. Only 1 paper published in the year 2000, followed by 0 published documents in 2001. Not a lot of papers were published till 2005. In 2005, 17 documents were published. Starting in the year 2006, there has been a stable progress till 2013, and as visible from the figure 2, a minutedip from 2013-2014. A speedy growth can be seen from the year 2015.

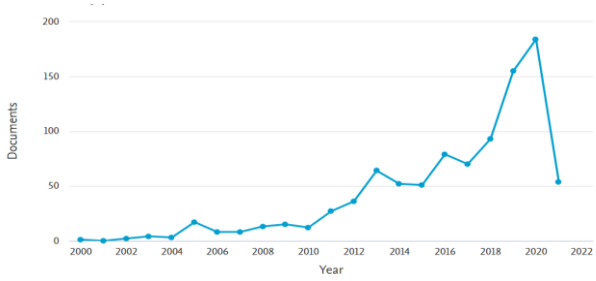

Fig 2.Document by year.

Figure 3 shows the number of documents, related to our research domain, published by various sources.

i) Bioinformatics source:

It is represented by the green line. The highest number of papers published by this source is 5, in the years: 2014 and 2018. The oldest paper from analysis of Scopus is from the year 2000.

ii) BMC Informatics source:

It is represented by the brown line. They started publishing papers specific to our research domain from the year 2003. The highest number of documents published by this source is 9 in the year 2019. A decrease in the documents published is visible in the figure3, from the year 2006-2007, 6 documents to 1 document.

iii) International Journal of Molecular Sciences source:

It is represented by the blue line. The first paper, related to our research domain, was published by them in the year 2009. There was a rapid growth from the year 2015 to 2020, as shown in figure 3. They have published the highest number of papers, i.e. 71 documents (in the year 2020), compared to other four sources present in the graph. The second highest peak shown in the graph, 31 documents, is also by the same source.

iv) Journal of Biomedical Informatics source:

It is represented by the orange line. The maximum number of papers published by them was in the year 2015, i.e. 7 documents. In 2009, they started publishing papers related to the domain.

v) IEEE access source:

It is represented by the purple line. The third highest peak, i.e. 23 documents, was published in 2020 by this source. IEEE access is relatively new in the domain, as the first paper was published by them in the year 2016.

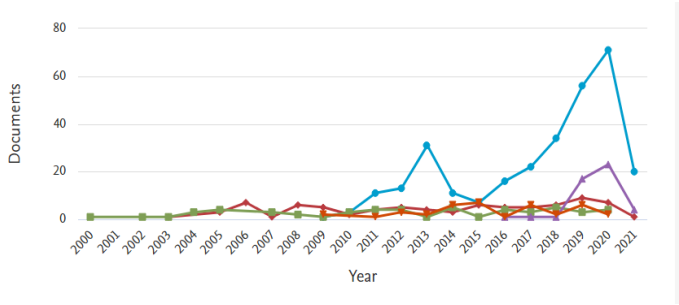

Fig 3.Documents by Sources

Figure 4 shows the 10 most influential countries that published papers in our domain. In the field of NER or Ovarian Cancer, the study carried out an analysis of the number of publications and contributions over various regions/ countries. From fig4, we can analyze that the United States has the highest publication,i.e. 269 documents. The number of publications from India are 43 documents.

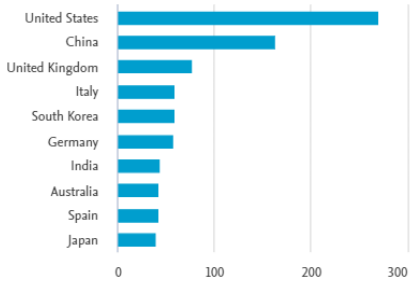

Fig 4. Documents by country/territory

All documents were published in 11 Scopus subject areas. There are 948 (32.3%) documents whose major focus is Computer Sciences. The list of number of documents, along with subject area name is- Biochemistry Genetics and molecular biology: 550, Chemistry: 334, Chemical Engineering: 327, Mathematics: 224, Engineering: 131, Medicine: 78, Material Science: 74, Social Science: 71, Agriculture and Biological Sciences: 59, Other: 140.

The 'Other' field includes area subjects like Physics and Astronomy, Neuroscience, Decision Sciences, Environmental Science, Arts and humanities, Health professions, Business, Management and Accounting, Pharmacology Toxicology and Pharmaceutics, Multidisciplinary, and Psychology.

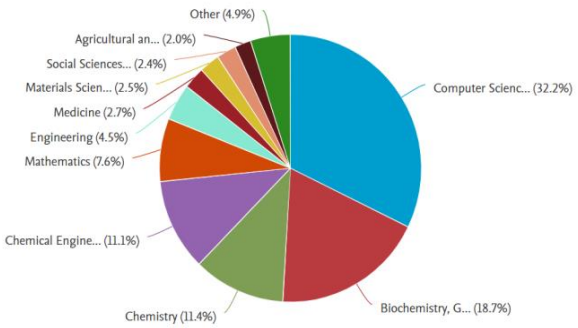

Fig 5.Documents by subject area

III. RESULTS AND DISCUSSION

The relation between entities can be represented as a network or graph, where the nodes of circles represent the units and the relations among them represent a link between two nodes.

To recognize the vital research topics, a co-word network map was created. For obtaining the said map,

i)the type of analysis: co-occurrence,

ii)unit of study: author keywords,

iii) counting method: full counting,

were set as the parameters [4]. The finalcount of keywords was 2148, which were too many to fit on a chart. So, we reduced the key words to 80, by limiting the number of occurrences to minimum five instances. Association strength is the technique used for normalization. The linkstrengthis based on the number of occurrences.

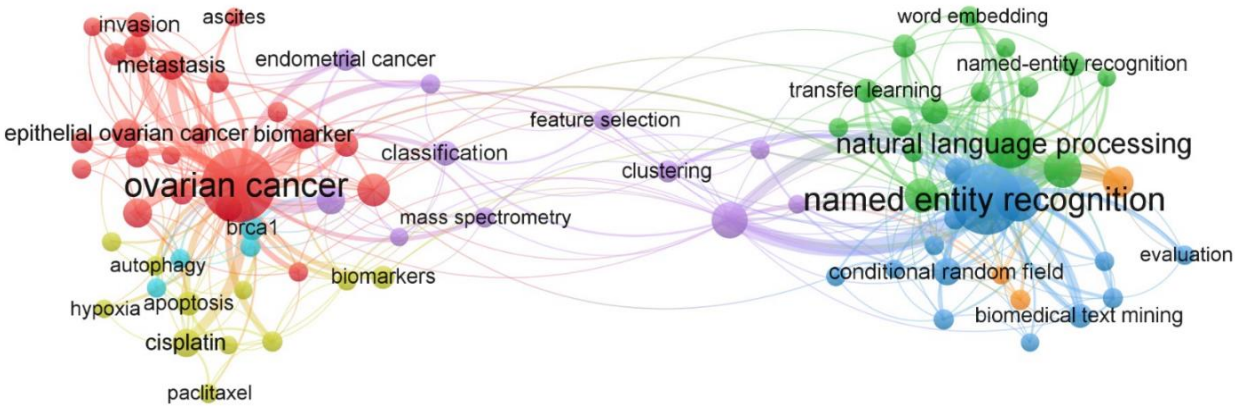

Fig 6. Bibliometric map network visualization for co-occurrence of Author keywords

expressions,plurals,etc.The top 10 index keywords based on the study are presented in Table 1.

Keywords provide details about the primary content of a document, and can also be used to recognize research trends in a specific domain [3]. Author keywords are the keywords provided by the author to describe the documents. Fig 6 shows the main author keywords that have been used in the online literature reviews. The co-occurrences of two terms as key words displays the vicinage by the thickness of lines joining them. The node size is directly proportional to the number of times a key word appears.

The largest nodes, like “ovarian cancer,” “named entity recognition,” “natural language processing”, have maximum number of occurrences.Not all labels are visible in a photographicimage of VOSviewer[6] map, in order to avert crisscrossing of labels [4]. Based on the information providers, index keywords are chosen, further regularized depending onplainly accessiblethesaurus. In contrast to Author keywords, the Indexed keywords also consider words having same meaning, different

Table 1. Index keywords

Keyword	Occurrences
Female	351
Ovarian Neoplasms	294
Metabolism	236
Ovary Cancer	230
Genetics	229
Ovary Tumor	226
Controlled Study	210
Named Entity Recognition	209
Natural Language Processing Systems	164
Protein Expression	139

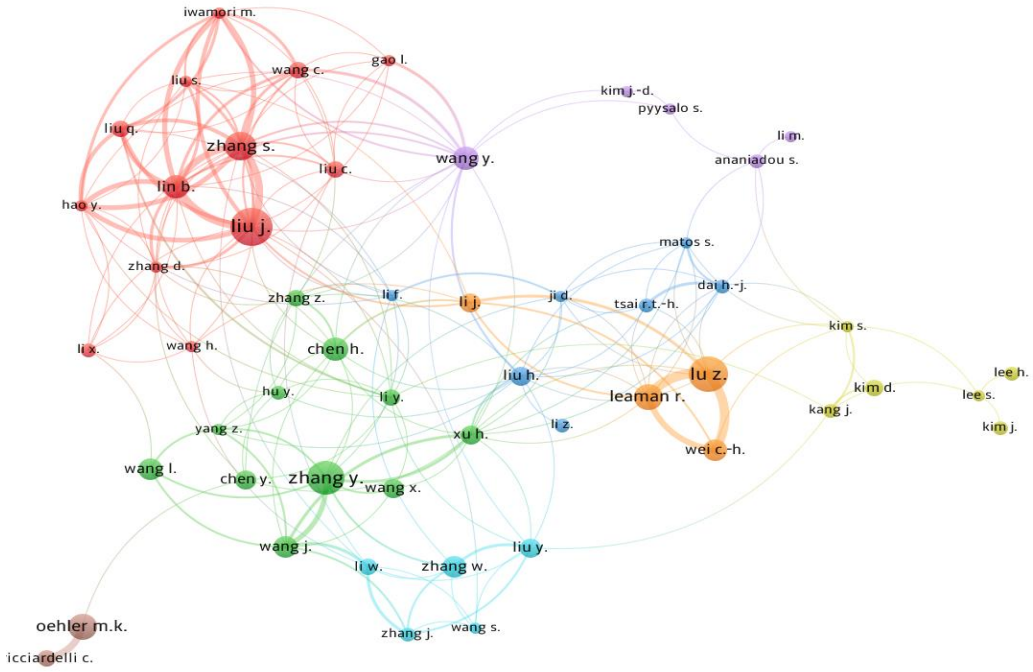

Fig 7. Network visualization of co-authorship of authors

Fig. 7 shows a co-authorship network map generated based on the number of published documents by the authors. For obtaining the said map,
i)the type of analysis: co-authorship,
ii)unit of study: author,
iii) counting method: full counting,
were set as the parameters. The total number of keywords was 3903, which were too many to fit on a chart. Therefore, a threshold of five documents and five citations per author was set, which was fulfilled by 64 authors. Association strength is the technique used for normalization. The visualization strength is based on the number of documents published by the authors [7].

Various components like the author nodes, co-occurrence weight, name of authors and networked relationship clustering are included in this map. Co-authorship relations represents the collaboration between two authors if they have worked together.Analysis of co-authorship details helped identify groups of people who worked together closely. The maximum number of the co-occurrences is attributed to four pairs of authors: Lu Z. &Leaman R., Lu Z. & Wei c h., Liu J. & Lin B. and Liu J. & Zhang S.

As seen in Fig 7.the authors;Lin B, Liu J, Zhang S, have the highest number of co-occurrences. Liu J. with 47 co-occurrence weights published 16 and Lu Z. with 29 co-occurrence weights published 15 papers .Zhang Y. with 19 link strength published 14 research papers .These authors can be considered as the most influential authors in this field.

Table 2.Author citations

Author	Documents	Citations
Furey T.S.	2	1801
Bednarski D.W.	1	1780
Cristianini N.	1	1780
Duffy N.	1	1780
Haussler D.	1	1780
Schummer M.	1	1780
Koestler D.C.	2	1387
Kelsey K.T.	1	1386
Marsit C.J.	1	1386
Nelson H.H.	1	1386

Table2.shows the top 10 of most-productive authors in the area with the number of published materials and number of citations in the analyzed area.
Furey T.S. is seen to have the maximum number of citations but the number of documents published are only 2. Similarly the rest of the authors in the table though they have a large number of citations but as the number of published documents are fairly low they cannot be seen in the visualization plot in Fig 7 because of the threshold set.

Fig 8 shows the network among 91 countries of international citation. It projects the author's countries of origin based on the global citations. In the map, a label along with a circle denotes a country. Based onlevel of importance the label and size of the circle grows larger. The circle size is proportionate to the published document count by all authors belonging to a country. The lineconnecting two nodesrepresentingdistinct countries stipulates a co-citation among the organizations in these countries. Out of 370 citation relations of countries, the 3 topmostintercontinentaltie-ups are:
Australia - USA (5.79%),
United States-South Korea (3.14%)
United Kingdom- United States (2.82%).
Among the European countries, Germany ranks first in articles production on the application of named entity recognition in biomedical research. Among the Asian countries, China ranks first followed by South Korea.

The United States has stronger collaborations with countries from Eastern Asia (such as China, South Korea, Japan and India), Germany, Italy, Portugal. Fewer collaborations have been observed with countries from the Middle East and Africa. The four highly cited and most productive regions are: USA, UK, China and South Korea.

India has a total of 222 citations for a total of 43 documents published. India has maximum

collaboration with the United States (18.18%) followed by the United Kingdom (16.36%).

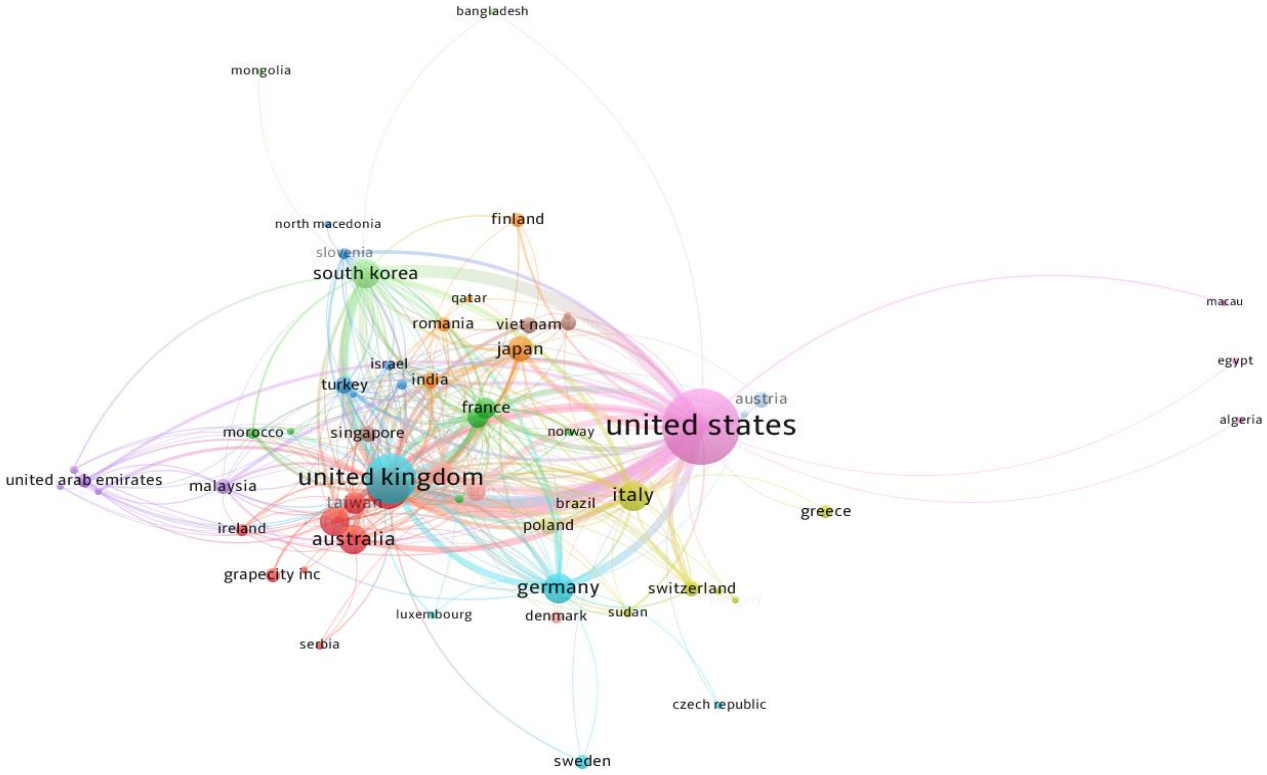

Fig 8. Citation network analysis of countries

The organization's co-authorship network is shown in Figure 9. The Department of Computer Science from the National Central University Taiwan, the Graduate Institute of Biomedical Informatics from Taiwan, and Database/Bioinformatics Laboratory, South Korea are the top three influential organizations for NER and biomedical based publications. The top universities from India include IIT Patna and IIT Madras.

In total, there were 948 publications by 2638 different organizations out of which only 36 have at least one link with another organization. Only these 36 organizations are displayed. The link strength is based on the number of documents that have been co authored by the node organization with other organizations. As it can be observed from the figure, the density near a node represents the number of publications by that organization in co-authorship with other organizations.

Table 3 displays the top 10 most productive organizations based on the number of documents published by that organization. We see that the organizations in the table are not displayed in the density graph in Fig 9 since these organizations have published documents without any collaboration with other organizations. From the above table we observe that Shengjing Hospital and School Of Computer Science, University Of Manchester, UK have published a maximum number of documents which is understandable since our area of interest involves “ovarian cancer” and “named entity recognition”.

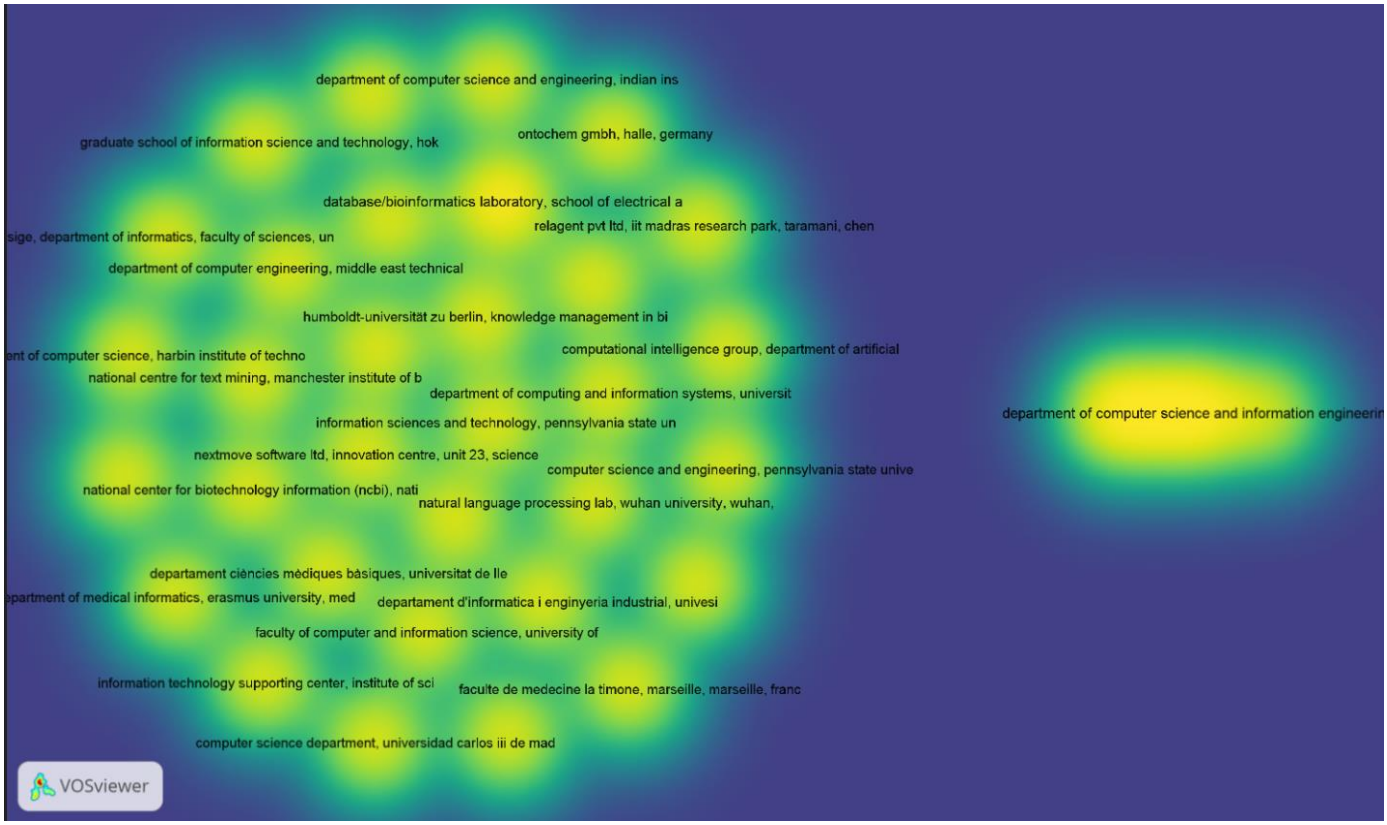

Fig 9. Co-authorship between organizations

Table 3. Organization citations

Organization	Documents	Citations
Department Of Obstetrics And Gynecology, Shengjing Hospital Affiliated To China Medical University, Shenyang 110004, China	5	105
School Of Computer Science, University Of Manchester, Manchester, United Kingdom	5	68
Department Of Biochemistry, Faculty Of Science And Technology, Kinki University, 3-4-1 Kowaka , Higashiosaka, Osaka 577-8502, Japan	4	78
Department Of Biomedical Sciences, Joan C. Edwards School Of Medicine, Marshall University, Huntington, Wv 25755, United	3	8

States		
Department Of Computer Science And Engineering, Korea University, Seoul, 02841, South Korea	3	229
Department OfGynaecological Oncology, Royal Adelaide Hospital, Adelaide, Sa 5000, Australia	3	256
Department Of Histology And Embryology, Poznan University Of Medical Sciences, Święcickiego 6 St, Poznań, 61-781, Poland	3	26
Department Of Human Genetics, Faculty Of Medicine, University Of Debrecen, Debrecen, H-4032, Hungary	3	3
Department Of Laboratory Medicine And Pathology, University Of Minnesota, Minneapolis, Mn 55455, United States	3	35
Department Of Obstetrics And Gynecology, Taipei Veterans General Hospital, Taipei, 112, Taiwan	3	33

IV. CONCLUSIONS

In named entity recognition for ovarian cancer research dominant countries, organizations and authors were determined. Also the most cited article, the most productive journals were revealed. For identifying the research distribution around the globe, bibliometric maps using VOSviewer mapping software are displayed in this paper[6].

With respect to the results of this research work, over the past twenty years there has been rapid evolution in scientific research dedicated to chemoinformatics [5]. This is the result of an increasing attention that government authorities are paying to the oncology department of the health care sector, which has led to an increased interest in this research area. A total number of 948 papers related to NER or Ovarian Cancer were published in scientific journals indexed to Scopus database in the year span from 2000 to 2021. The most influential countries which have performed research in this field are the United States, China and the United Kingdom, of which the United States occupied the first place in terms of productivity based on the number of publications and number of citations.

In this study, we presented eight criteria including highly-cited articles, subject areas, productivity, keywords frequency, global institutions, most influential authors, leading publishing journals, collaboration between countries. These criteria helped reveal the global trends related to research on named entity recognition in biomedical information extraction. It was noted that 86.7% of the documents, out of 948 shortlisted documents, were published within the last decade. It was noted that in order to ensure high quality of research articles and to increase the number of citations, it is crucial to publish articles in highly ranked journals. This study also helped bring in the forefront the active authors in terms of publications and the top 10 most active authors were projected. From the visualization plot it was found that Liu.J. contributed the most publications followed by Lu Z. Analysis of the keyword frequencies helped describe the trends and research directions for future study in the related field. Various keywords and such as “ovary cancer”,

"genetics", "named entity recognition", "natural language processing systems" served as important words in the study.

V. REFERENCES

1. Jeroen Baas, Michiel Schotten, Andrew Plume, Grégoire Côté, and Reza Karimi,"Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies" , Quantitative Science Studies MIT Press, Volume 1,ISSN: 2641-3337,Issue 1,February 2020, pp.377-386 https://doi.org/10.1162/qss_a_00019
2. Xieling Chen, HaoranXie, Fu Lee Wang, Ziqing Liu, Juan Xu and Tianyong Hao,"A bibliometric analysis of natural language processing in medical research", BMC Medical Informatics and Decision Making, Volume 18, ISSN: 1472-6947, Article 14,March 2018, pp.1-72. https://doi.org/10.1186/s12911-018-0594-x
3. In-SeonLee,HyangsookLee,Yi-Hung Chen,Yongbyoung Chae, "Bibliometric Analysis of Research Assessing the Use of Acupuncture for Pain Treatment Over the Past 20 Years", Dove Press Journal of Pain Research, Volume 13,ISSN:1178-7090, January 2020, pp.367-376. http://doi.org/10.2147/JPR.S235047
4. BabajideAbubakrMuritala , Maria-Victoria Sánchez-Rebull, Ana-Beatriz Hernández-Lara, "A Bibliometric Analysis of Online Reviews Research in Tourism and Hospitality",Online Reputation and Sustainability,Volume 12 , ISSN 2071-1050, Issue 23, November 2020, pp.1-18. http://dx.doi.org/10.3390/su12239977
5. Howard Ramírez-Malule, Diego H. Quinones-Murillo, Diego Manotas-Duque,"Emerging contaminants as global environmental hazards. A bibliometric analysis", Ke Ai Publication:Emerging Contaminants, Volume 6, 2405-6650, pp.179-193. https://doi.org/10.1016/j.emcon.2020.05.001
6. VOSviewer website - https://www.vosviewer.com/
7. Vijayshri Nitin Khedkar, Sonali Kothari Dr, Aarohi Prasad, Arunima Mishra, Varun Saha, Vinay Kumar," Analysis of Recent Trends in Continuous Sign Language Recognition using NLP", Library Philosophy and Practice (e-journal), Volume 5231 , ISSN 1522-0222 , Issue 3, March 2021, pp.1-21.