Library Philosophy and Practice (e-journal)

Libraries at University of Nebraska-Lincoln

Fall 10-8-2020

A Bibliometric Survey on Cognitive Document Processing

Dipali Baviskar
Symbiosis Institute of Technology ,Pune, dipali.baviskar.phd2019@sitpune.edu.in

Swati Ahirrao
Symbiosis Institute of Technology ,Pune, swatia@sitpune.edu.in

Ketan Kotecha
Symbiosis Institute of Technology,Pune, director@sitpune.edu.in

Baviskar, Dipali; Ahirrao, Swati; and Kotecha, Ketan, "A Bibliometric Survey on Cognitive Document Processing" (2020). *Library Philosophy and Practice (e-journal)*. 4557.
https://digitalcommons.unl.edu/libphilprac/4557

A Bibliometric Survey on Cognitive Document Processing

Dipali Baviskar[1], Dr.Swati Ahirrao[2], Dr.Ketan Kotecha[3]

[1]*Research Scholar, Symbiosis Institute of Technology (SIT) affiliated to Symbiosis International (Deemed University), Pune, India.*
Email: dipali.baviskar.phd2019@sitpune.edu.in

[2]*Ph.D. Guide and Associate Professor, Symbiosis Institute of Technology (SIT) affiliated to Symbiosis International (Deemed University), Pune, India.*
Email: swatia@sitpune.edu.in

[3]*Head, SCAAI, Symbiosis International (Deemed University), Pune, India.*

Email: head@scaai.siu.edu.in,director@sitpune.edu.in

ABSTRACT

Heterogenous and voluminous unstructured data is produced from various sources like emails, social media tweets, reviews, videos, audio, images, PDFs, scanned documents, etc. Organizations need to store this wide range of unstructured data for more and longer periods so that they can examine information all the more profoundly to make a better decision and extracting useful insights. Manual processing of such unstructured data is always a challenging, time-consuming, and expensive task for any organization. Automating unstructured document processing using Optical Character Recognition (OCR) and Robotics Process Automation (RPA), seems to have limitations, as those techniques are driven by rules or templates. It needs to define the template or rules for every new input, which limits the use of rule or templates based techniques for unstructured document processing. These limitation demands to develop a tool which can be able to process the unstructured documents using Artificial Intelligence techniques. This bibliometric survey on Cognitive Document Processing reveals the mentioned facts about unstructured data processing challenges. This survey is performed on the Scopus database's scientific documents. Various tools such as Microsoft Excel, Sciencescape, VOSviewer, Leximancer, and Gephi for drawing network data analysis diagrams are used. The study revealed that the largest number of publications on Cognitive Document Processing had been explored very recently. It is observed that universities/institutions in India are leading in the research studies focusing on this research topic.

Keywords : unstructured data,document processing, Artificial Intelligence, Robotics Process Automation(RPA),Optical Character Recognition(OCR), bibliometric analysis.

1. INTRODUCTION

In the modern era of digital communications, various applications such as word documents, spreadsheets, scanned documents, document conversion, PDFs, audio, video, and email applications such as Gmail, Yahoo, etc. have brought about a surge of information. This data generally falls under the category of unstructured data as it does not fit in any predefined schema or format [1],[2]. Organizations are trying to get the benefit from this unstructured data by deriving key insights to take the right decision which will be useful for their growth [3], [1], [4]. To automate the business processes organizations are using Optical Character Recognition (OCR) and Robotics Process Automation (RPA) [5], [6]. OCR extracts the text out of scanned documents, PDFs, the camera captured images. Organizations are utilizing this extracted text for analysis, processing, and editing [7]. Although OCR is a useful tool, it is template-based [8] and does not work well on unstructured documents [9]. OCR is not helpful to understand the contents of the documents [10]. Due to these disadvantages of OCR, organizations are using Robotics Process Automation (RPA) to automate the business processes [11]–[13]. RPAs are built to automate the repetitive, high-volume tasks that are rule-based [14]. RPA cannot process semi-structured or unstructured data, since to get the insights it requires some degree of decision-making [1].

So, there is an urge to develop a tool that can able to process the unstructured documents using Artificial Intelligence techniques and get key insights from the unstructured documents.

There are no known literature references available that have been carried out on bibliometric analysis on Cognitive Document Processing despite vast research on OCR and the importance of RPA to automate the business processes. This bibliometric study aims to fill the literature study gap using bibliometric analysis tools to understand the existing information in OCR and RPA literature and to apply it to further increase the organization's productivity. This will also help in directed research plans for the future development of Cognitive Document Processing with value-added collective research among experts.

The bibliometric analysis provides the visualizations of the trends in research, keywords correlation, author contributions, citations analysis in the research area [15]. This bibliometric study would help researchers to find main authors with whom they can collaborate which will help their future research work, quality journal available in the particular research, scope of research in the different areas of specialization, and the degree of collaboration in the research [16].

1.1 Bibliometric Analysis of Cognitive Document Processing

Researchers urge to perform the bibliometric survey which will help them to understand the recent research trends and scope of research in particular research areas. The last page of any book, article, or report tells the list of sources from which the information to write a book, article, or report is taken. Similarly, the concept "Bibliometrics" can be used as a statistical analysis tool for books, articles, or other publications [17]. It is used for the measurement of research sources used in a particular research study. Bibliometric analysis can be used to evaluate the performance of research happening in universities and organizations. It is a remarkable tool to quantitatively analyze the research data based on articles, citation counts, geographical locations, and various other quantifying parameters [18]. Research gaps and Future work/scope can easily be identifiable with the help of a bibliometric analysis of the particular research field [16], [19].

This bibliometric survey on Cognitive Document Processing can be used to get a better insight into the research area.

The objectives of this bibliometric study are as mentioned below:

- To recognize the different sources of publications in the field of research
- To analyze the different language used for publications
- To observe year-wise publications trends
- To locate the countries contributing more to the research on a geographical map
- To analyze different source types of research publications
- To identify a network of highly contributing authors
- To observe publications based on affiliations (university/organization).
- To check useful publication by seeing the citation count for the publications.

2. GATHERING DATA

There are two ways of collecting the scientific articles, publications, documents, one way is subscription-based access by paying the fees to get access to the required article, and the other way is to access without any fees known as open access articles. Organization or institutes library portal allows researchers to access the publications by registering themselves to the respective websites accessing it via personal membership. Some of the popular research databases are Scopus, Web of Science, Science Direct, Research Gate, and Google Scholar, etc.

Scopus database is a standard repository for scientific documents such as journals, proceedings of conferences, reviews, and book chapters. Scopus has a database of abstracts and citation-based peer-reviewed articles for various disciplines such as science, business and management, bioinformatics, medicine, arts, and engineering. Therefore Scopus is considered as a persistent and accurate source whenever research is to be carried out.

2.1 Important keywords

The essential keywords concerning Cognitive Document Processing were divided into two parts namely: primary and secondary keyword. Table 2. indicates the selection of query keywords used as a search strategy for this research.

Table 2: Selection of search keywords for Cognitive Document Processing

Primary keyword	" document processing"
Secondary keyword using (AND)	“Artificial Intelligence” “Optical Character Recognition”
Secondary keywords using (OR)	“ analysis “, “ unstructured data”, “big data”, “retrieval”, “information extraction”, “named entity”, “recognition”, “cognitive” “classification”, “ character recognition”

Thus the query used to search the documents in Scopus is: "document processing” AND “Artificial Intelligence” AND “OCR” OR “analysis “OR “information extraction” OR “retrieval” OR “big-data” OR “large-scale-data” OR “unstructured data” OR “classification” OR “character recognition”

2.2 Results based on the initial query string

The Scopus database is an important unit of this bibliometric study. A query with selected keywords used as a query string and was executed on the Scopus database which returned 312 publications. Table 3. shows different types of publication sources available for Cognitive Document Processing. From the statistics shown in table 3, it is observed that 61.21% of the researchers have published their work in conference paper followed by an article that contributes 33.01%. It is observed that publication in “conference review” is contributed least.

Table 3: Type of Publications in the Cognitive Document Processing

Type of Publications	**Number of Publications**	**Percentage**
Conference Paper	191	61.2179%
Article	103	33.01%
Book Chapter	10	3.20%
Review	5	1.60%
Book	2	0.64%
Conference Review	1	0.32%
Total		100%

Data access information source: http://www.scopus.com (accessed on Oct.02 2020)

The search result was analyzed based on the type of language used in published documents. Table 4. shows the trends in the language used for publishing documents in Cognitive Document Processing. The English language is majorly used by the researchers to publicize their publications while the other languages are least used as shown. The search shows only 02 Chinese language publications.

Table 4: Languages trends used for publishing Cognitive Document Processing

Sr.No.	Language used for Publishing	Count of Publications
1	English	310
2	Chinese	02
Total		312

Data access information source: http://www.scopus.com (accessed on Oct.02 2020)

2.3 Statistical data analysis

The research documents related to Cognitive Document Processing were retrieved between 2010 and 2021. Table 5. shows publication counts per year in the research area of Cognitive Document Processing. It can be easily analyzed that the research area has contributed more in the year 2019 and 2017. However, very few researches were carried out in the years 2010 and 2012.

Table 5: Yearly publishing trends in Cognitive Document Processing

Year	Publication Count
2021	3
2020	18
2019	47
2018	26
2017	41
2016	34
2015	33
2014	26
2013	21
2012	19
2011	26
2010	18

Data access information source: http://www.scopus.com (accessed on Oct.02 2020)

Figure 1. shows the visual representation of Table 5. A line chart represents a prominent year for research document publication. The year 2019 has the highest publication count of a total of 47 documents and 2017 with publication count of 41 documents. The X-axis denotes the year of publication and Y-axis denotes the number of documents.

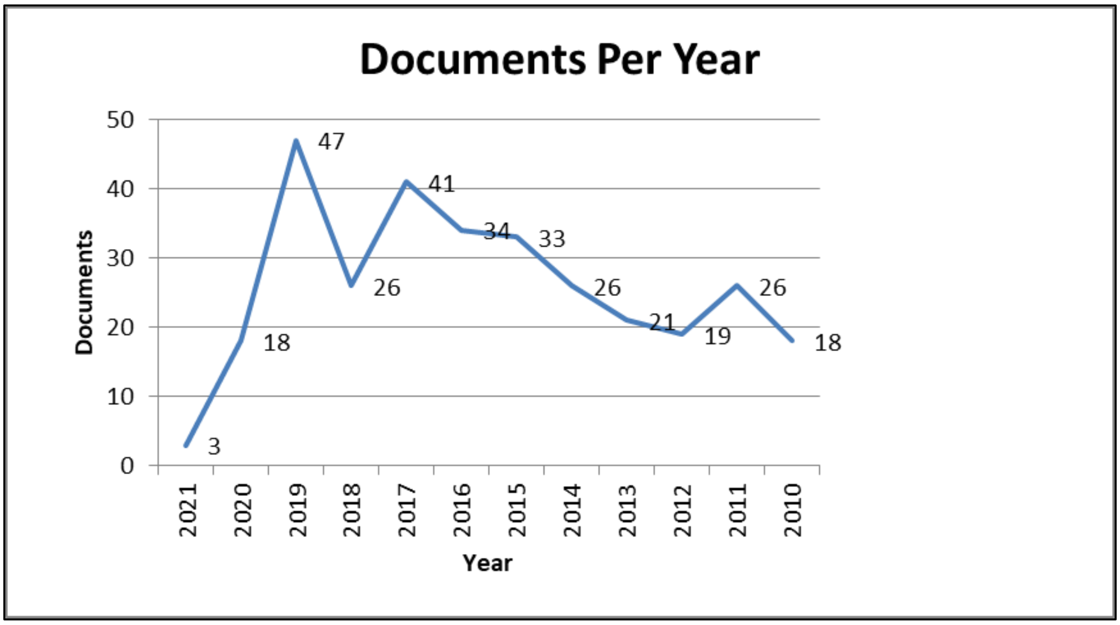


Figure 1: Yearly publishing trend in Cognitive Document Processing

Data access information source: http://www.scopus.com (accessed on Oct.02 2020)

2.4 Data assessment

Section 3 provides a detailed bibliometric study to understand the diversity in the extracted literature, to know the significant research contributors, and the recent research problem statements in the area of Cognitive Document Processing. It shows uniqueness in the research area via geographical locations where the research is conducted, via affiliations of the authors, co-author relationship, and journal titles where the research papers were published. The analysis is done based on keywords used, number of citations for research publication, and collaborative research work.

3. BIBLIOMETRIC SURVEY

The bibliometric study is performed using two different methods -

• Statistical analysis: Focuses on country-wise contribution to the research area, subject area wise contribution, author's affiliations, authors, source type, and source titles.

• Network analysis: Shows geographical locations, keywords co-occurrence, co-authorship, co-citation count, bibliographic coupling, publication year trends, and collaborative research with other authors.

3.1 Analysis based on geographic locations

Figure 2 shows a map of research papers published from various countries around the world. This map is created using Google sheets. The data from the excel sheet which has two columns country and publication count is copied to Google spreadsheet and by selecting the map option in other charts a map is drawn. To know publication count for a particular country, hover a mouse over a particular map area. The left bottom ruler shows the publication count scale. The region in green is India which has the highest number of publications i.e. 72 for Cognitive Document Processing.

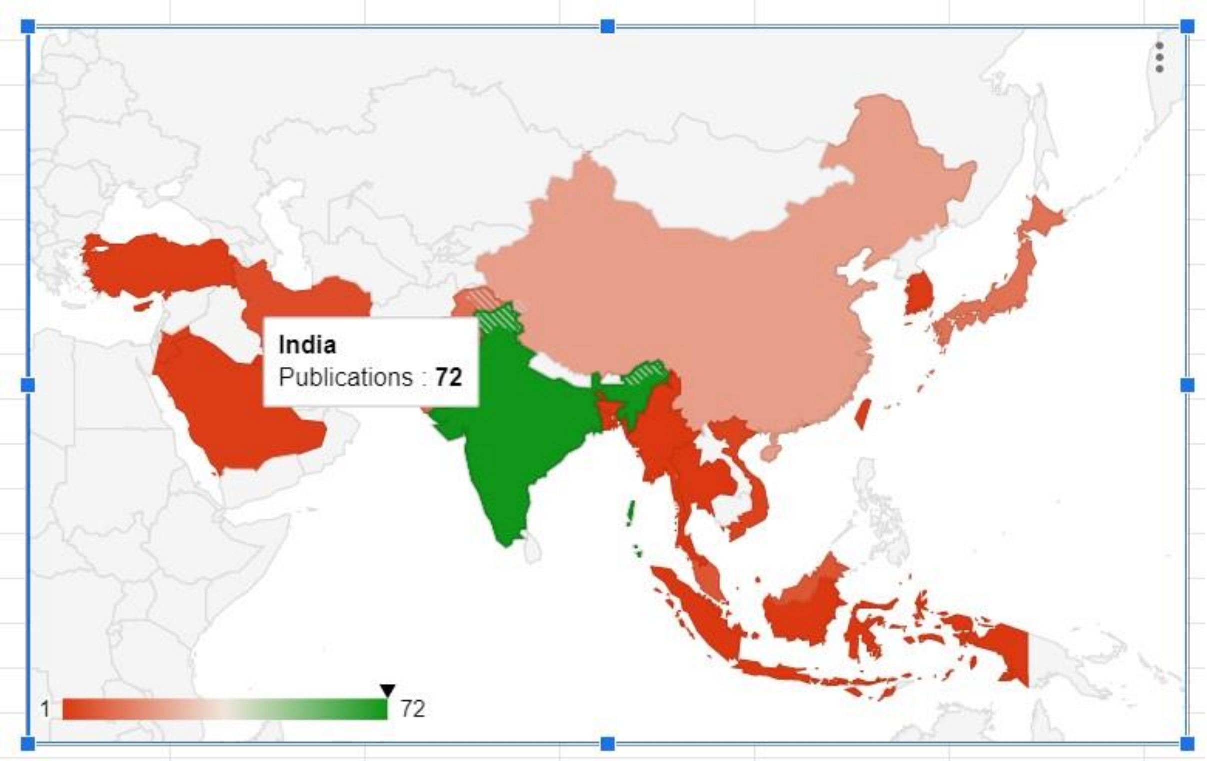


Figure 2: A map of research papers published from various countries around the world

Figure 3 shows publications from the top ten countries for Cognitive Document Processing. The column graph shows that India has 72 published documents in the Cognitive Document Processing followed by the United States with 53 document's contribution. Spain, Australia, Japan, Pakistan, and the United Kingdom are the countries that contribute least to the Cognitive Document Processing.

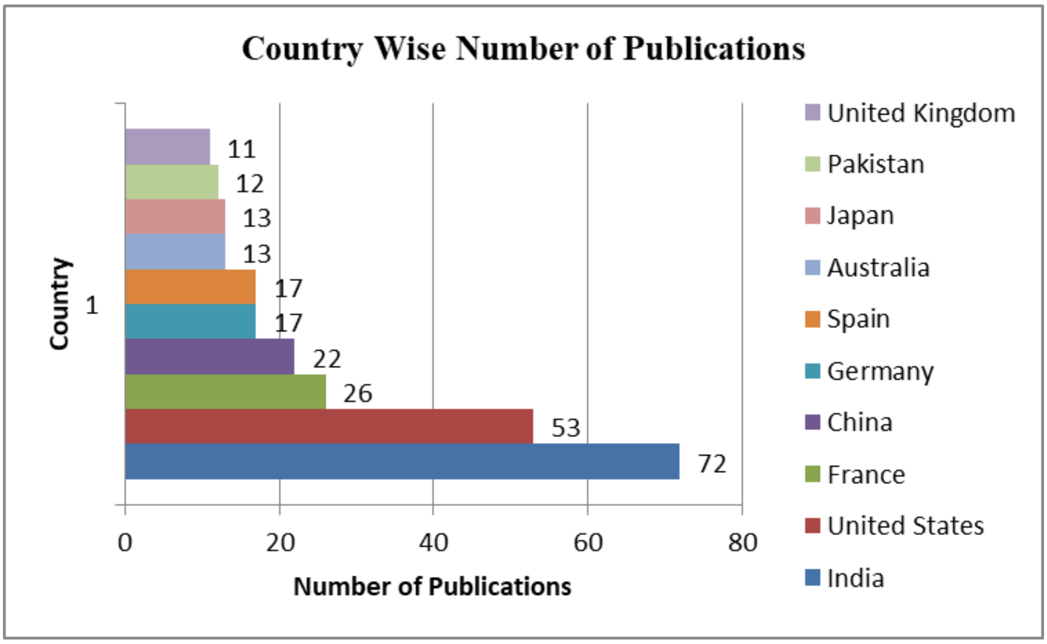


Figure 3: Ten topmost countries publishing papers on Cognitive Document Processing

Data access information source: http://www.scopus.com (accessed on Oct.02 2020)

3.2 Keyword's Statistical analysis

Table 6. shows the top ten keywords searched by the Scopus database for Cognitive Document Processing. A proper combination of keywords helps to identify the publications in the required research area. Keywords indicate the researcher's query string and tell what the researcher wants to search for. The following table shows the count of publications for the top ten important keyword's occurrence in the majority of the Cognitive Document Processing research documents.

Table 6: Count of Publications of the top ten keywords for Cognitive Document Processing

Keywords	Count of Publications
Optical Character Recognition	107
Image Processing	88
Character Recognition	74
Information Retrieval Systems	70
Text Processing	67
Information Retrieval	50
Document Images	48
Natural Language Processing	40
Classification (of Information)	39
Image Segmentation	36

Data access information source: http://www.scopus.com (accessed on Oct.02 2020)

3.3 Network Analysis

Network analysis is used to represent the relationship among different quantifiable attributes using graphical formats. Many tools are available for the same purpose. In this bibliometric study Sciencescape, VOSviewer, Gephi, and Leximancer are used to draw network analysis diagrams. Figure 4-9 shows network analysis diagrams combining different computable parameters on the retrieved Scopus database.

VOSviewer is a freely available tool. The fundamental purpose of the tool is to study the bibliometric network on different quantifying parameters. The .csv extension file from Scopus is given as input to VOSviewer. Three types of visualization analysis are done based on network, overlay, and density. Figure 4 is the network visualization map based on a combination of keywords and source titles extracted from Scopus. Keywords used in the source titles of extracted documents are represented with circles. The high-frequency occurrence of the keyword is shown with the bigger circle in the network diagram. Links between the circles represent the distance between two keywords. If the size of the link is smaller, the association among the keywords is strong. Keywords with the same colors represent clusters of closely related keywords. The diagram has eight different clusters. Each cluster is represented with a different color. The threshold value for the minimum number of occurrence of a keyword is limited to 5 keywords. Out of the total 2408 keywords, 144 keywords met the threshold value. The relevance score is calculated for 144 keywords.VOSviewer has calculated the total strength of the co-occurrence links with other keywords for these 144 keywords and for all the selected 144 keywords the relation between keywords and source titles is drawn.

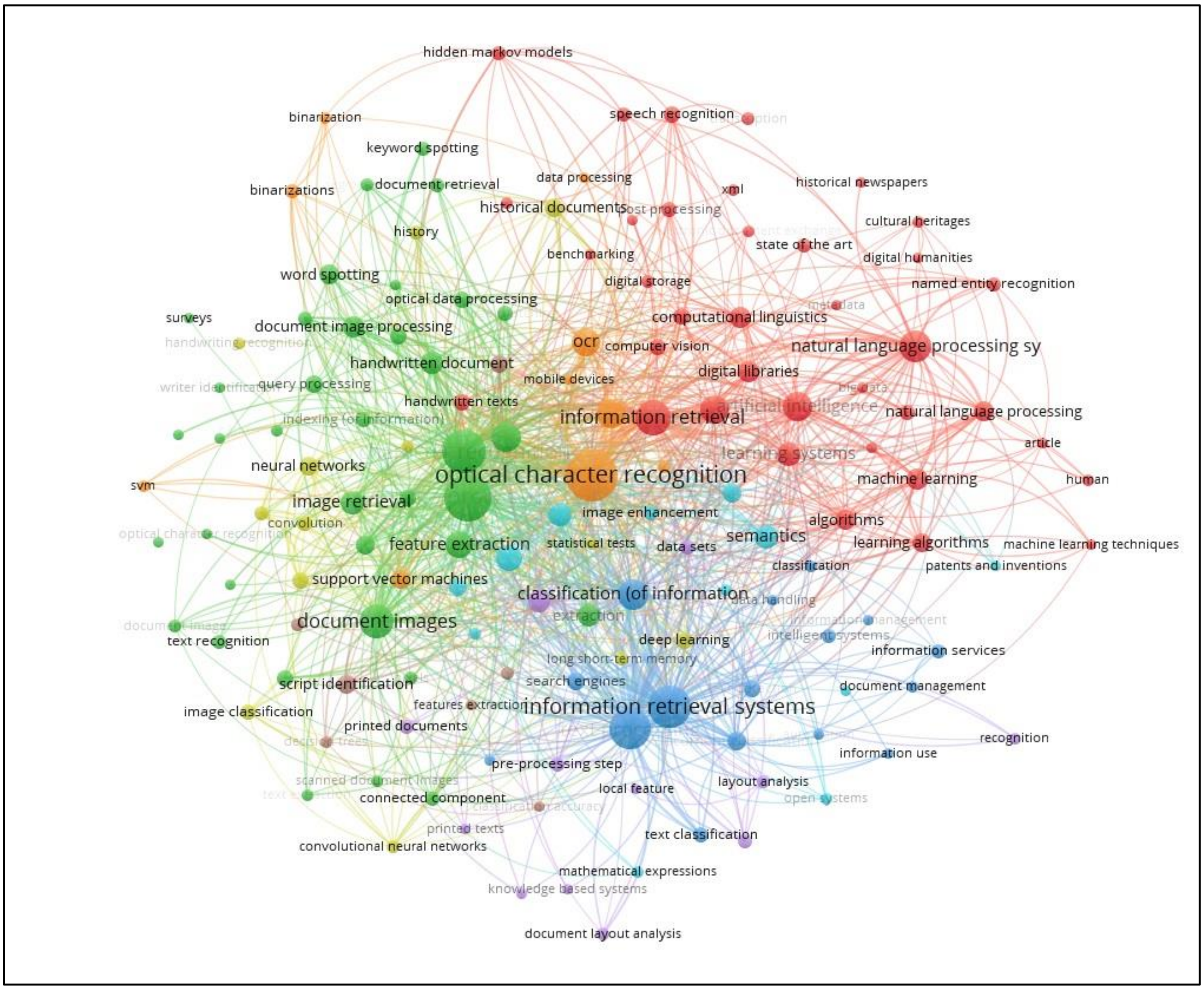


Figure 4: Network visualization diagram based on keywords and source title.

(Source: https://www.vosviewer.com)

Figure 5 shows the Reference scape network diagram to show the networks of bibliographical references. The exported network consists of bibliographic references connected and with Authors, Author Keywords, and Sources (journals) having Nodes: 228 Edges: 4676. Nodes represent the entities Author or a Keyword and Edges represent a collaboration between them. The following filters are applied to draw the network diagram in Gephi.

- References cited in less than 2 papers are filtered out
- References connected to less than 107 other references are filtered out
- Authors present in less than 3 papers are filtered out
- Keywords present in less than 3 papers are filtered out
- Journals present in less than 5 papers are filtered out

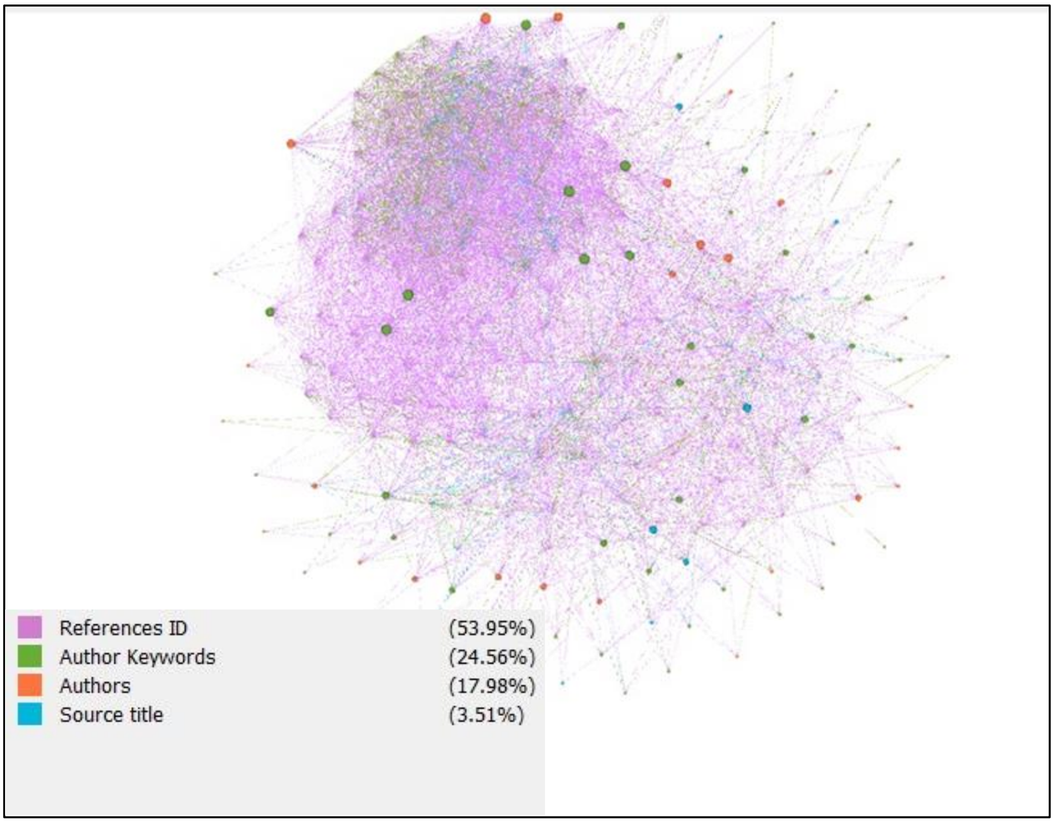


Figure 5: The Reference scape network diagram to show the networks of bibliographical references (Source: https://gephi.org/)

Figure 6 shows a "cluster of co-authors and authors co-appearing among the same papers". The collaborative work is shown among the authors. The link represents the collaborative work of authors on the documents published. The threshold value for "the minimum number of documents per author" was set manually to 2 along with the "minimum number of citations per author" was set manually to 2, which resulted in 110 authors. Out of 866 authors, 110 met the threshold. For every 110 authors, 12 authors have a strong connecting link and are shown with the colour link lines between them. For selected 12 authors, VOSviewer calculated the “total strength of the co-authorship links with other authors” as shown in figure 6.

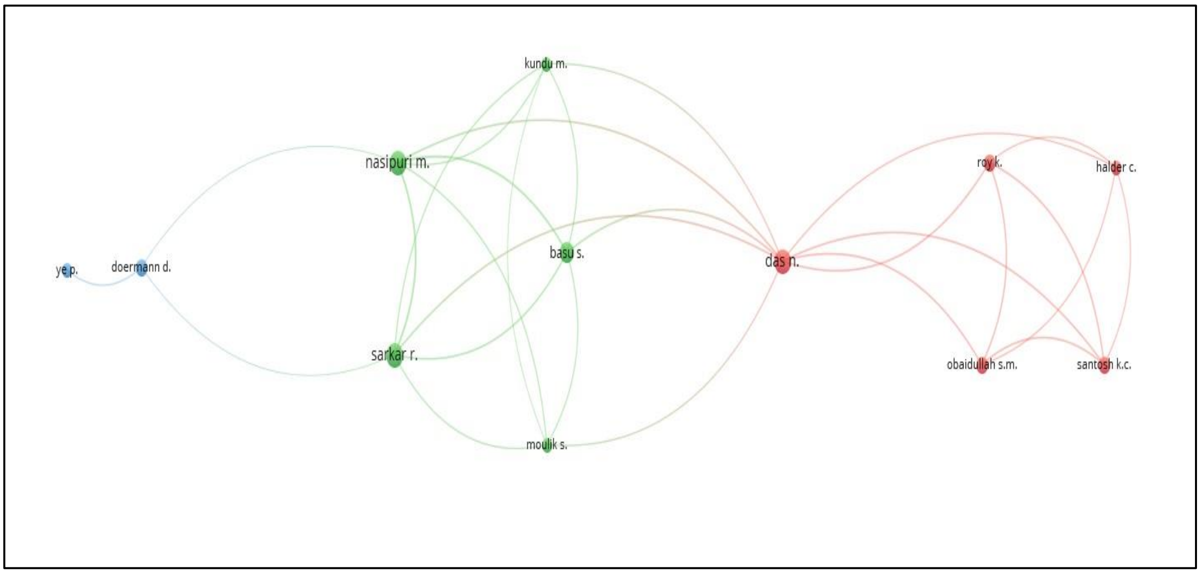


Figure 6: Network analysis diagram of co-authors and authors based on co-appearance among the same papers (Source: https://www.vosviewer.com)

The network diagram is also drawn by changing the threshold value for “the minimum number of documents per author” to 2 and “the minimum number of citations per author” to 0 which results in 121 authors. VOSviewer has calculated the co-authorship relationship with a minimum of 2 authors with default zero citation count for their publications as shown in figure 7.

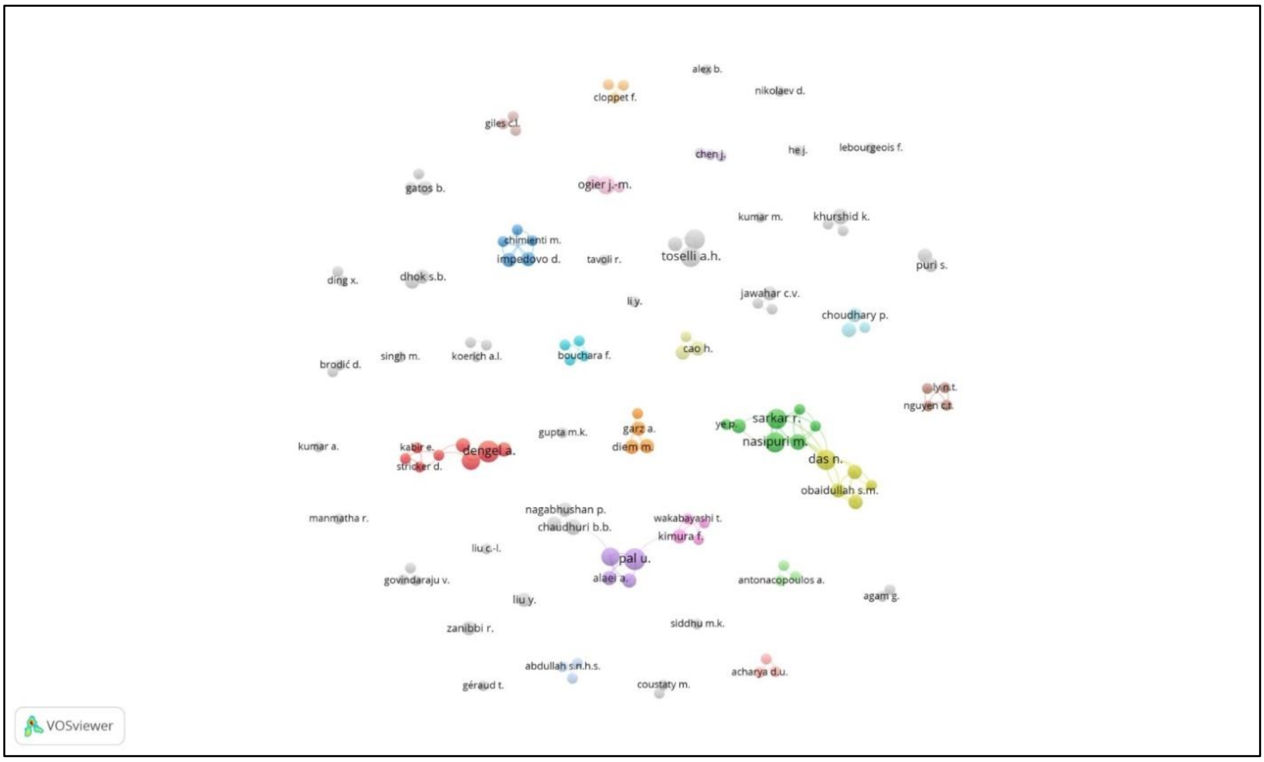


Figure 7: Network analysis diagram of co-authors and authors based on co-appearance among the same papers with zero citations (Source: https://www.vosviewer.com)

Figure 8 shows the network map of the publication title and the citations received by publications. Gephi open-source software is used to draw this diagram. The Fruchterman-Reingold layout is used to plot diagrams. The layout shows a 104 number of nodes as the publication title is a collaborative work of the authors and 5058 edges. The edges were set to "in-degree" property which means that the arrows coming towards a specific node have cited that particular paper. The dark red color dot represents the publication title that received the highest number of citations.

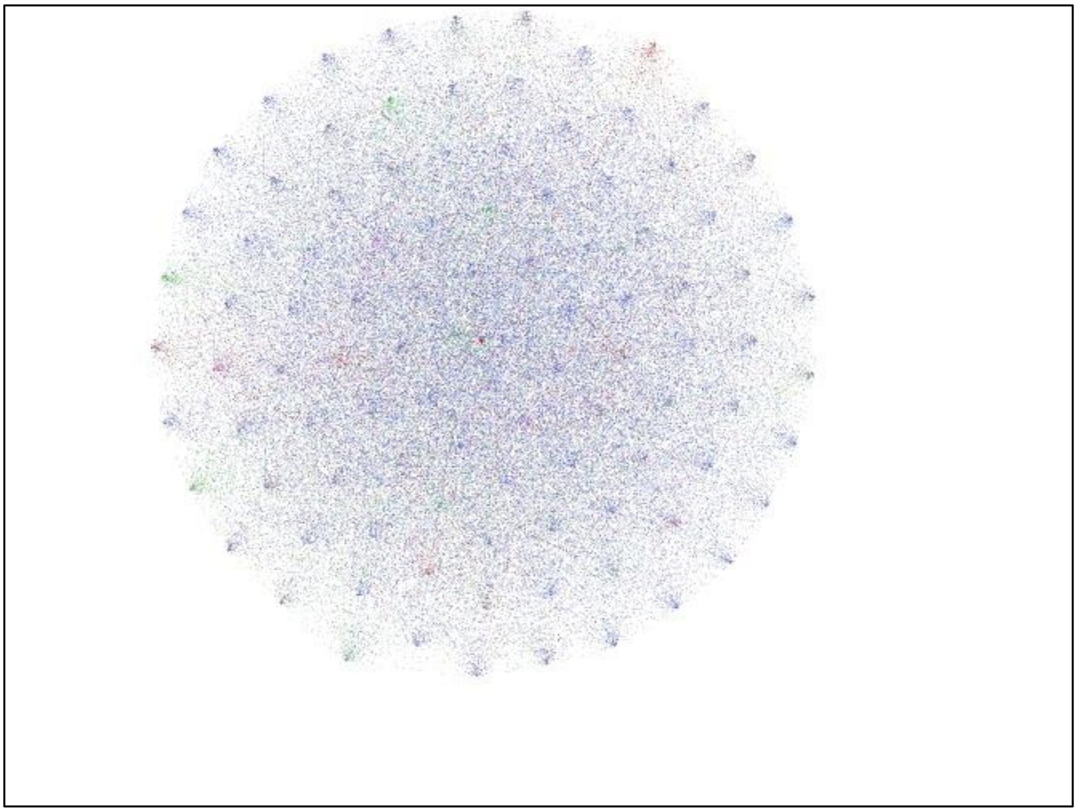


Figure 8: Network map of publication title and the citations received by publications (Source: https://gephi.org/)

3.4 Discipline/Subject-area Statistical analysis

Figure 9 shows the division of publications in different disciplines. The pie-chart can be easily seen that the majority of the research is carried out in the Computer Science area followed by the engineering and Mathematics area. It is observed that Cognitive Document Processing research has been also carried out in the area of Business, Management, and Accounting with fewer publications.

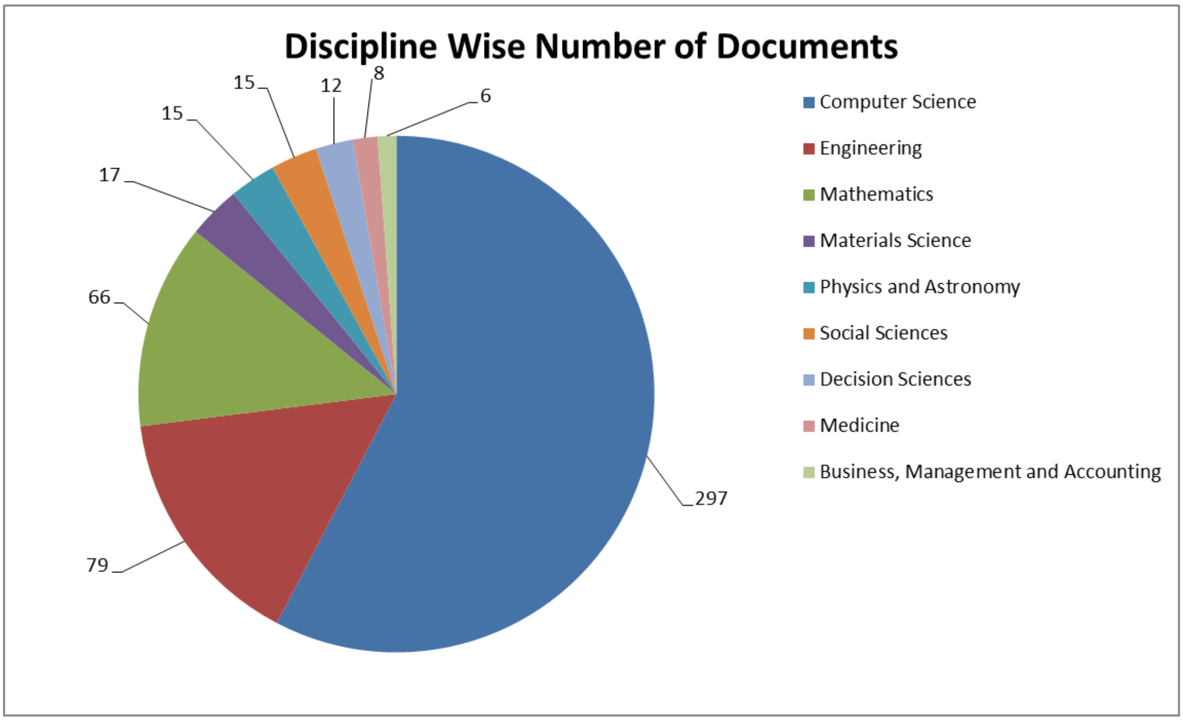


Figure 9: Discipline wise analysis of extracted literature for Cognitive Document Processing

Data access information source: http://www.scopus.com (accessed on Oct.02 2020)

3.5 Affiliations Statistical analysis

Figure 10 shows the top ten universities and organizational affiliations contributing to the research.“The German Research Centre for Artificial Intelligence DFKI” shows the maximum contribution to Cognitive Document Processing research followed by the Indian Statistical Institute, Kolkata. It can be observed that many universities are contributing to this research with a different number of their publications.

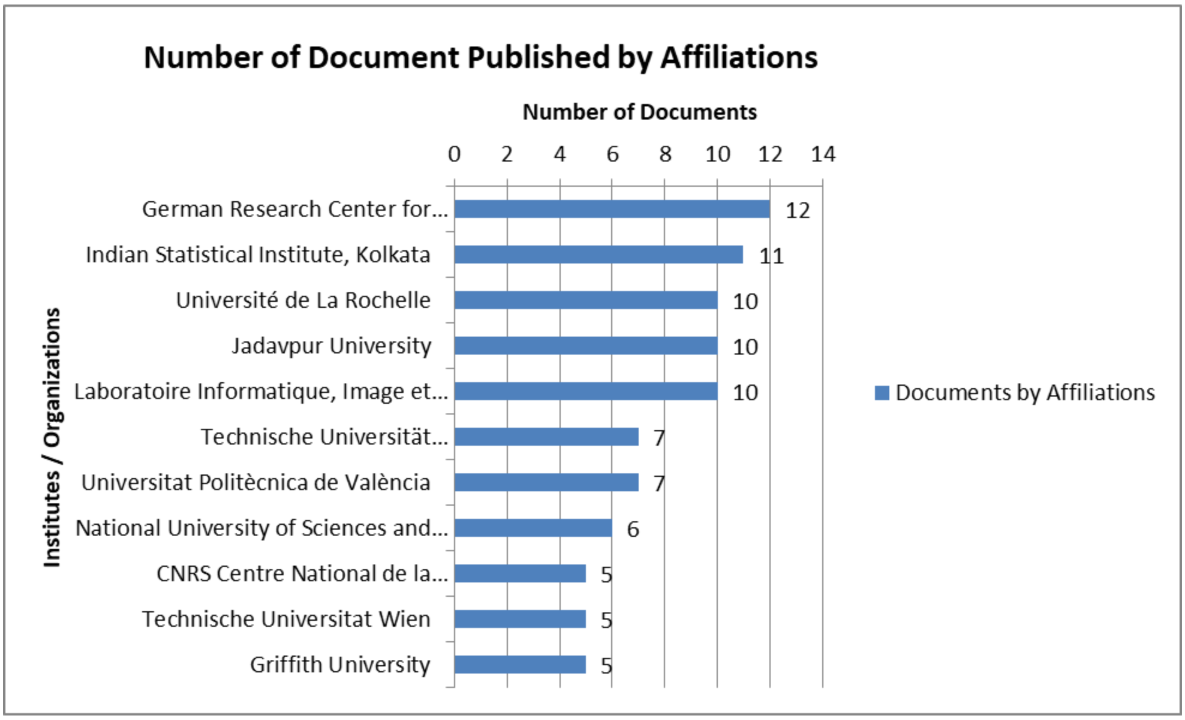


Figure 10: Affiliation statistics for Cognitive Document Processing

Data access information source: http://www.scopus.com (accessed on Oct.02 2020)

3.6 Authors Statistical analysis

Figure 11 shows the top ten authors contributing to the Cognitive Document Processing to understand the impact of a particular author.

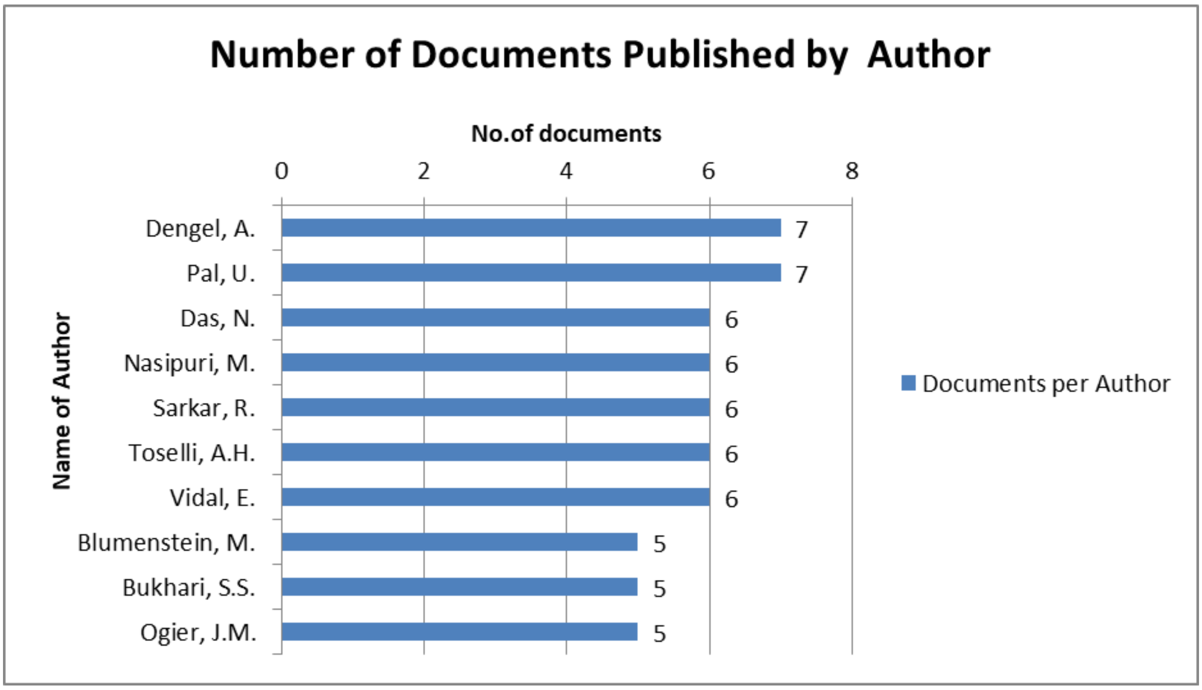


Figure 11: Top ten authors contributing for Cognitive Document Processing

Data access information source: http://www.scopus.com (accessed on Oct.02 2020)

3.7 Statistical analysis based on Source Types

The meaning of "source types" of scholarly articles is the place where the original research work is published. From the Scopus extracted documents for Cognitive Document Processing, it can be clearly stated that 61% of the publications are from "conference proceedings" followed by 33% publications in the "journal". It has been observed that "conference review" publications for the Cognitive Document Processing are not yet published.

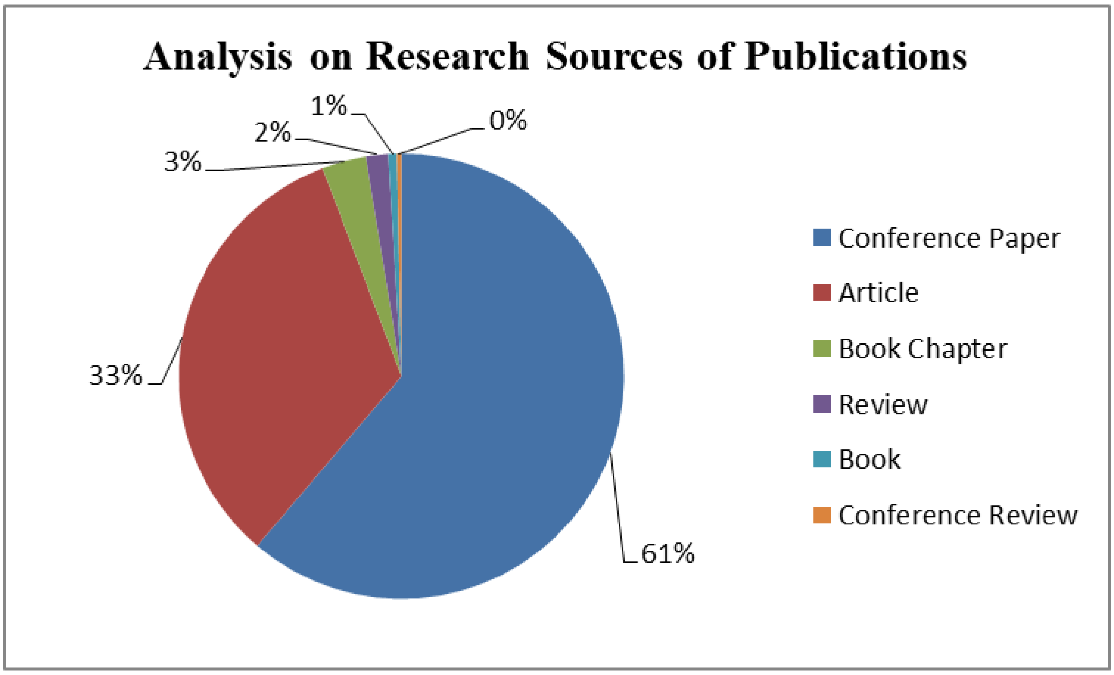


Figure 12: Source types for publications for Cognitive Document Processing

Source data access: http://www.scopus.com (accessed on Oct.02 2020)

3.8 Citations statistical analysis

Table 7 shows the statistics for the number of citations per year for the Cognitive Document Processing publications. So far the total citation count of Scopus scientific 312 publications is 2193. Very few documents are cited from 2010 to 2012 summing up to 86 citations. The publications have received the maximum number of citations in 2019 followed by 2020.

Table 7: Analysis based on citations for publications in the Cognitive Document Processing

Year	**No.of Citations Per Year**
<2010	0
2010	5
2011	34
2012	47
2013	83
2014	137
2015	146
2016	219
2017	292
2018	384
2019	492
2020	343
Sub-Total	**2182**
>2020	11
Total	**2193**

Data access information source: http://www.scopus.com (accessed on Oct.02 2020)

The top ten publication titles extracted from the Scopus database that received the maximum number of citations so far are represented in Table 8. It can be seen that the research work with the title “Image Processing and Pattern Recognition: Fundamentals and Techniques” received the maximum number of citations in Cognitive Document Processing research.

Table 8: An analysis of the top ten publications based on citations in the Cognitive Document Processing

Publication Title	Citations received by the Publications yearly									
	<2014	2014	2015	2016	2017	2018	2019	2020	2021	Total
Image Processing and Pattern Recognition: Fundamentals and Techniques	55	24	11	17	20	25	20	12	1	185
A binarization method with learning-built rules for document images produced by cameras	25	8	7	11	7	6	11	6	0	81
An ensemble heterogeneous classification methodology for discovering health-related knowledge in social media messages	0	2	11	12	15	8	12	9	0	69
Document image quality assessment: A brief survey	0	3	11	8	5	12	11	6	0	56
ICDAR 2013 competition on writer identification	0	6	11	8	6	7	5	6	0	49

Patent retrieval	2	16	4	7	8	5	6	0	0	48
De-identification for privacy protection in multimedia content: A survey	0	0	0	1	9	10	15	9	0	44
Arabic information retrieval	0	2	3	11	6	9	8	4	0	43
Natural language processing for historical texts	0	3	2	11	3	7	11	5	0	42
Large-scale extraction of gene interactions from full-text literature using DeepDive	0	0	0	6	6	15	9	2	0	38

Data access information source: http://www.scopus.com (accessed on Oct.02 2020)

3.9 Source titles statistical analysis

Figure 13 shows the top ten source title’s statistic data for Scopus scientific documents for Cognitive Document Processing. It is observed that maximum numbers of publications are published in a source titled “ Lecture Notes In Computer Science Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics”.The sources such as “IEEE Access”, “International Journal Of Pattern Recognition And Artificial Intelligence Proceedings” are less used by the researchers to publish their scientific documents.

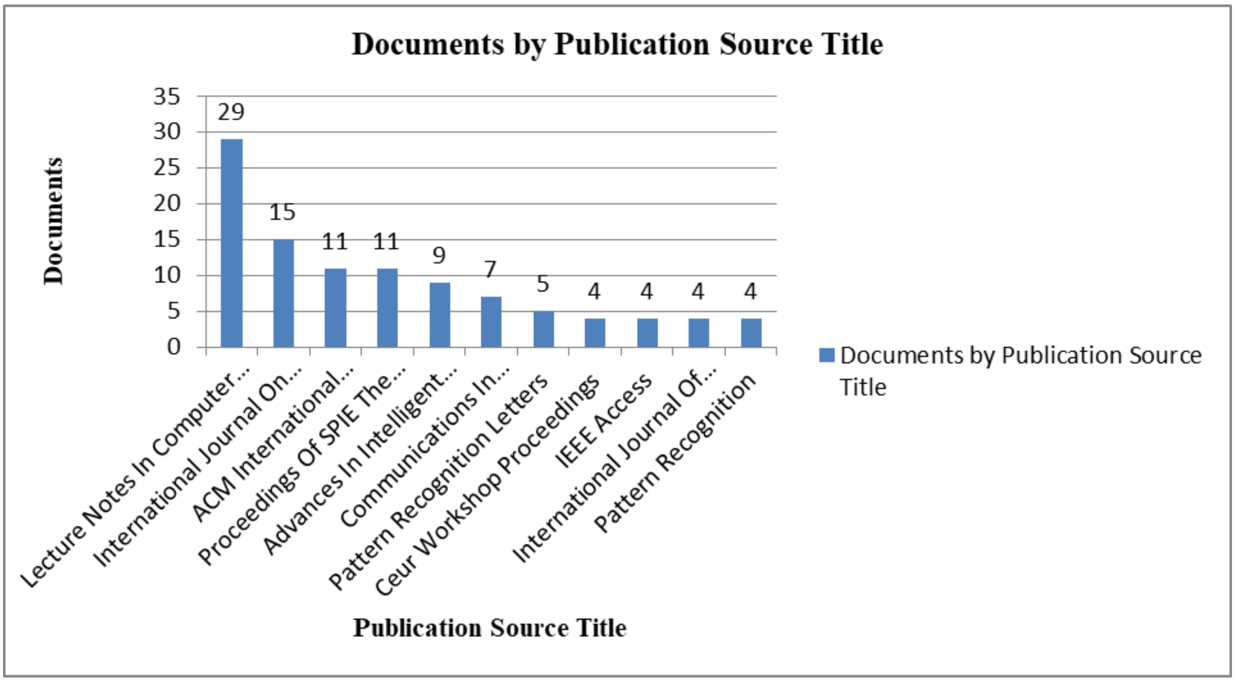


Figure 13: Source statistics for publications in Cognitive Document Processing

Data access information source: http://www.scopus.com (accessed on Oct.02 2020)

3.10 Funding Sponsors statistical analysis

Figure 14 shows a statistical analysis based on funding sponsors in the Cognitive Document Processing research. The top 10 funding sponsors are considered and it can be seen that the "National Science Foundation" is the highest funding foundation followed by the "National Natural Science Foundation of China".

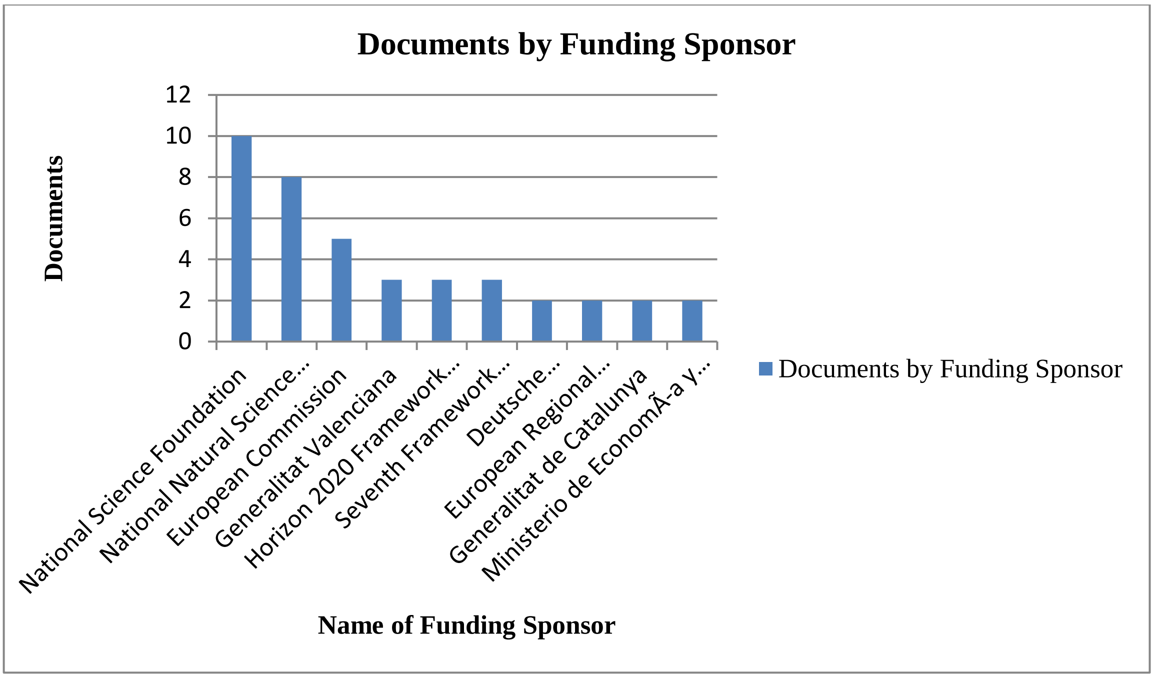


Figure 14: Funding Sponsors statistics in Cognitive Document Processing

Data access information Source: http://www.scopus.com (accessed on Oct.02 2020)

3.11 A-K-J Sankey Analysis

Figure 15 shows the associations of top Author-main Keywords-main Journal associations for the retrieved Scopus research publications. To visualize, explore, and download networks of keywords and/or authors and/or journals this network graph can be used. For creating this A-K-J Sankey network graph, a Scopus CSV file is uploaded and a network map of the top authors, keywords, and journals and their association is generated.

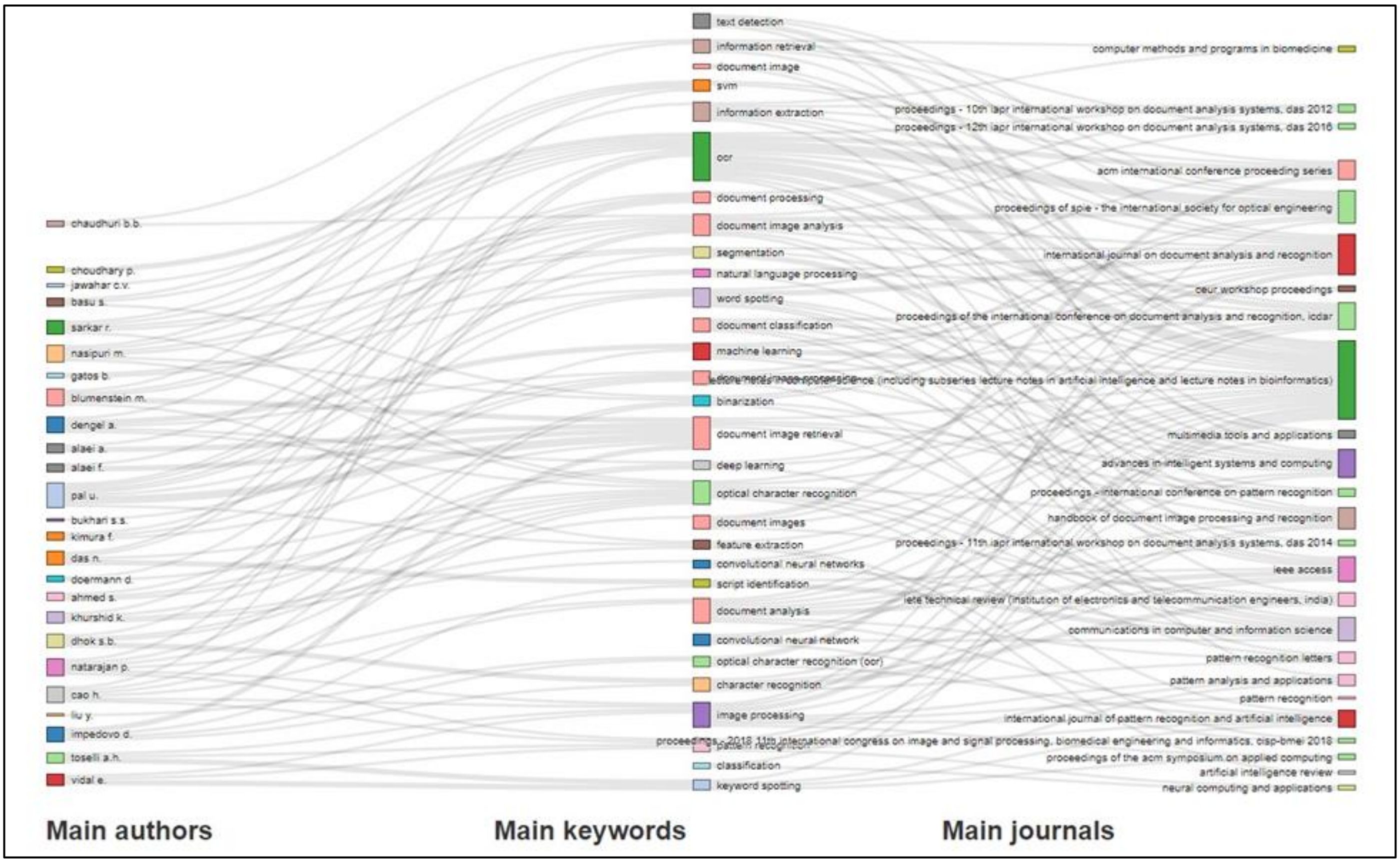

chaudhuri b.b.
choudhary p.
jawahar c.v.
basu s.
sarkar r.
nasipuri m.
gatos b.
blumenstein m.
dengel a.
alaei a.
alaei f.
pal u.
bukhari s.s.
kimura f.
das n.
doermann d.
ahmed s.
khurshid k.
dhok s.b.
natarajan p.
cao h.
liu y.
impedovo d.
toselli a.h.
vidal e.
text detection
information retrieval
document image
svm
information extraction
ocr
document processing
document image analysis
segmentation
natural language processing
word spotting
document classification
machine learning
document image processing
binarization
document image retrieval
deep learning
optical character recognition
document images
feature extraction
convolutional neural networks
script identification
document analysis
convolutional neural network
optical character recognition (ocr)
character recognition
image processing
pattern recognition
classification
keyword spotting
computer methods and programs in biomedicine
proceedings - 10th iapr international workshop on document analysis systems, das 2012
proceedings - 12th iapr international workshop on document analysis systems, das 2016
acm international conference proceeding series
proceedings of spie - the international society for optical engineering
international journal on document analysis and recognition
ceur workshop proceedings
proceedings of the international conference on document analysis and recognition, icdar
lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)
multimedia tools and applications
advances in intelligent systems and computing
proceedings - international conference on pattern recognition
handbook of document image processing and recognition
proceedings - 11th iapr international workshop on document analysis systems, das 2014
ieee access
iete technical review (institution of electronics and telecommunication engineers, india)
communications in computer and information science
pattern recognition letters
pattern analysis and applications
pattern recognition
international journal of pattern recognition and artificial intelligence
proceedings - 2018 11th international congress on image and signal processing, biomedical engineering and informatics, cisp-bmei 2018
proceedings of the acm symposium on applied computing
artificial intelligence review
neural computing and applications
Main authors
Main keywords
Main journals

Figure 15:The associations of top Author-main Keywords-main Journal associations

Source: http://www.scopus.com (accessed on Oct.02 2020)

3.12 Term Co-occurrence network analysis

Figure 16 shows the “term co-occurrence map based on text data” of 321 retrieved Scopus publications. Terms will be extracted from the title and abstract of papers. Binary counting which is the default in VOSviewer is used. The minimum number of term occurrences was set as 5 in VOSviewer, out of 6843 terms extracted,344 met the threshold. The most relevant terms with a score of more than 60 % relevance are selected by VOSviewer and a map is drawn.

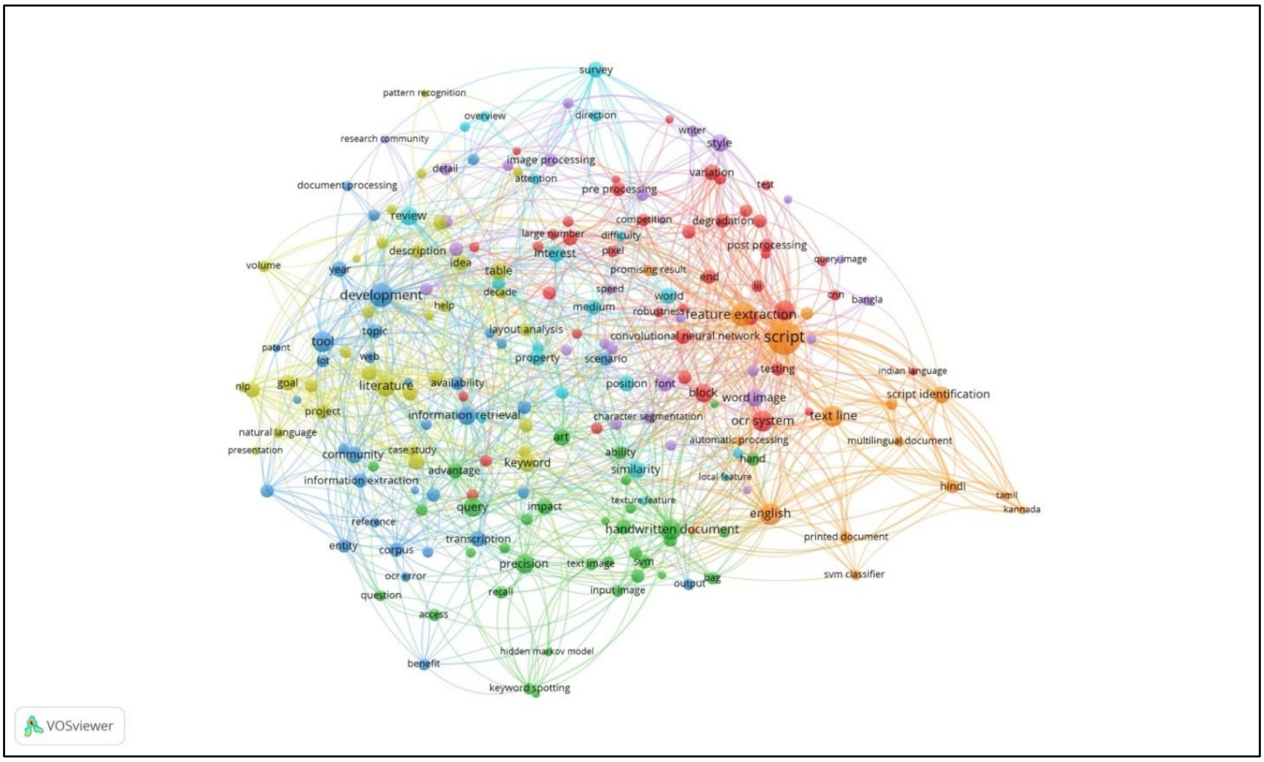


Figure 16:The term co-occurrence map based on text data

(Source: https://www.vosviewer.com)

3.13 Analysis of Bibliographic Coupling for research documents

Alike to co-citation a bibliographic coupling is a similarity degree that uses document citation analysis to establish a similarity relationship between documents [17]. Bibliographic coupling methods can be utilized to identify 'hot' research points [19]. The analysis depends on fitting limits for both numbers of related documents and the quality of bibliographic connections. Those papers are called “core documents”. The “minimum number of citations per document” was selected as two. Out of 312 documents,192 documents met the threshold. For these 192 documents, “the total strength of the bibliographic coupling links with other documents” is calculated and shown in the analysis map.

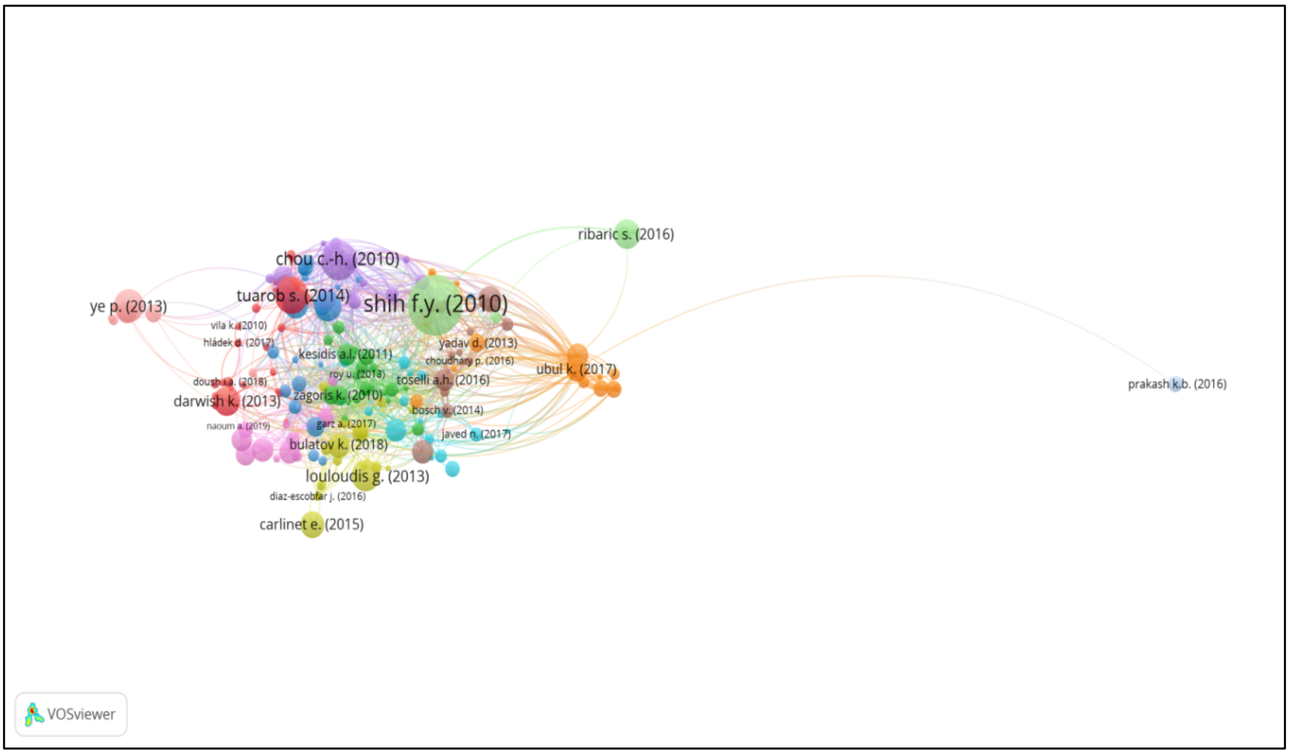


Figure 17: The bibliographic Coupling analysis of documents

(Source: https://www.vosviewer.com)

3.14 Theme diagram using Leximancer

A Leximancer 'Theme' is a cluster of Concepts that have some shared trait or connectedness as observed from their nearness on the Concept Map. The size of the “Theme circle” has no relation concerning its predominance or significance in the content; the circles are only limits. Commonness is dictated by the number of Concepts present in the theme and this is demonstrated in the “Thematic Report”. Figure 18 shows the theme diagram for 312 research publications derived from its relevance of concepts and their connectedness.

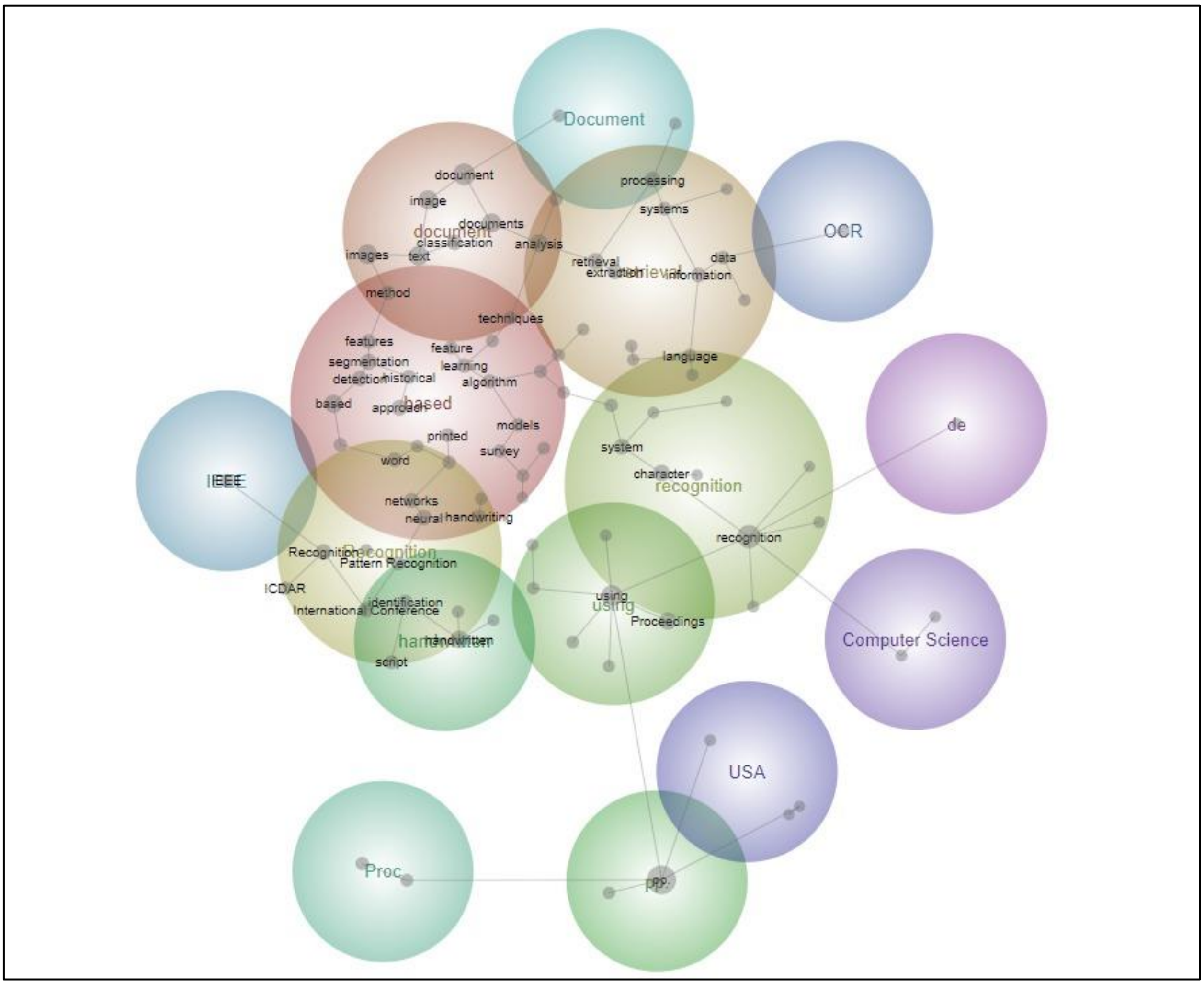
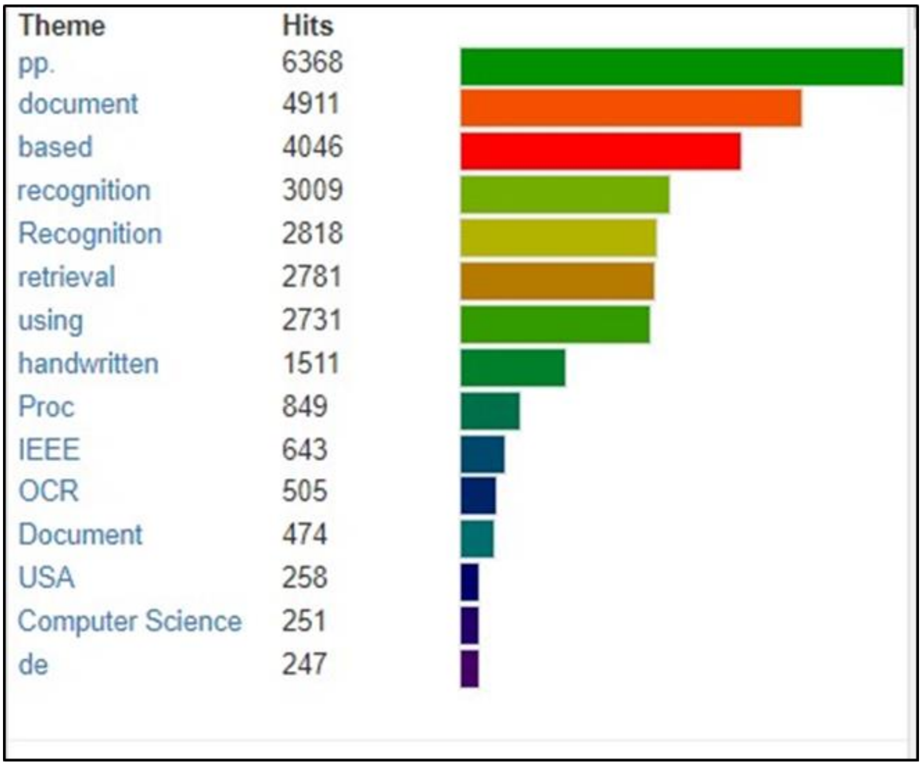


Figure 18: The theme diagram (Source: https://lexiportal-app.leximancer.com/)

Figure 19 shows the theme relevance diagram used as a thematic report for drawing the theme diagram shown above in figure 18.

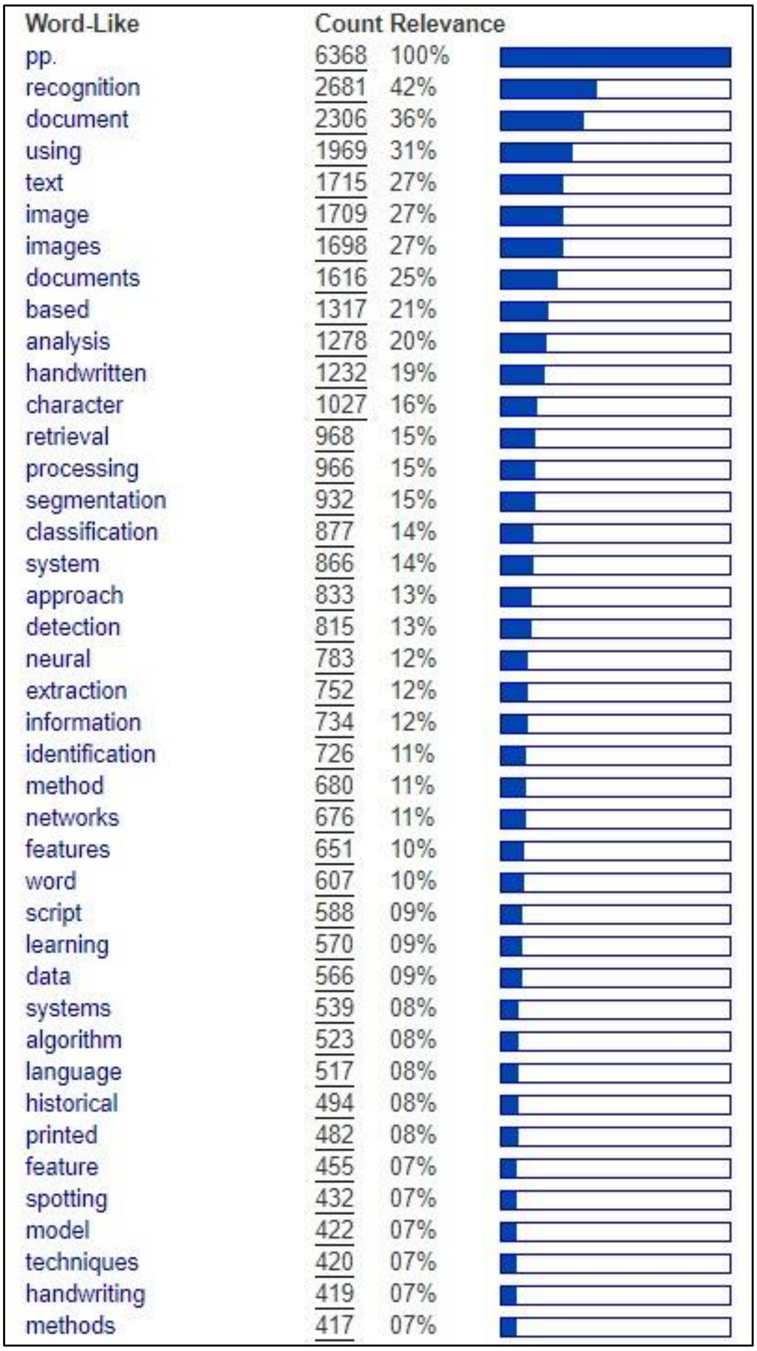


Theme	Hits
pp.	6368
document	4911
based	4046
recognition	3009
Recognition	2818
retrieval	2781
using	2731
handwritten	1511
Proc	849
IEEE	643
OCR	505
Document	474
USA	258
Computer Science	251
de	247

Figure 19: Theme relevance(Source: https://lexiportal-app.leximancer.com/)

3.15 Word relevance analysis using Leximancer

Leximancer is a tool to conducts a quantitative analysis of the principle concepts present in a text and their associations with one another. Leximancer gives word occurrence count present in a text with a percent of relevance with one other. Retrieved 312 Scopus research publications are given as input to Leximancer to get the word relevance analysis diagram as shown in figure 20.



Word-Like	Count	Relevance
pp.	6368	100%
recognition	2681	42%
document	2306	36%
using	1969	31%
text	1715	27%
image	1709	27%
images	1698	27%
documents	1616	25%
based	1317	21%
analysis	1278	20%
handwritten	1232	19%
character	1027	16%
retrieval	968	15%
processing	966	15%
segmentation	932	15%
classification	877	14%
system	866	14%
approach	833	13%
detection	815	13%
neural	783	12%
extraction	752	12%
information	734	12%
identification	726	11%
method	680	11%
networks	676	11%
features	651	10%
word	607	10%
script	588	09%
learning	570	09%
data	566	09%
systems	539	08%
algorithm	523	08%
language	517	08%
historical	494	08%
printed	482	08%
feature	455	07%
spotting	432	07%
model	422	07%
techniques	420	07%
handwriting	419	07%
methods	417	07%

Figure 20: The word relevance analysis (Source: https://lexiportal-app.leximancer.com/)

4. DISCUSSION

Research in Cognitive Document Processing is crossing its boundaries throughout the world and it is spreading enormously. This study throws light on the importance of the contribution to Cognitive Document Processing using the Artificial Intelligence research area.

The types of publications in the respective field of research are majorly conference papers followed by articles. To refine and state the research idea, these statistical and network analysis are effective channels. The English language is used by the majority of the researchers to publish the documents. It is observed that minimal contribution was made towards publishing the documents in the initial years of 2010 to 2012. The graph shows incremental changes every year. The year 2019 and 2017 shows the maximum number of document publications respectively in Cognitive Document Processing.

Countries like India and the United States are contributing more towards the research in this area. It can be observed that conferences are the primary source types used by the researchers to publish their scientific documents. Based on the extracted literature, it is observed that out of 312 papers only five are review papers. Keyword analysis diagrams show that the most significant keywords for Cognitive Document Processing are Optical Character Recognition, Information extraction, document images, Natural Language Processing, Artificial Intelligence, document structure, layout analysis, etc. The network analysis diagram on “Co-occurrence of keywords” inferred that the most significant keywords are present in the retrieved Scopus scientific publications for Cognitive Document Processing. The query string consisting of these co-related keywords will help researchers to get more accurate results in Cognitive Document Processing research. Bibliographic coupling analysis proved that from the retrieved 312 Scopus scientific documents,192 are “core documents” related to Cognitive Document Processing research.

It can be observed that “The German Research Centre For Artificial Intelligence DFKI” shows maximum contribution towards the research in Cognitive Document Processing followed by the “Indian Statistical Institute, Kolkata”. Other universities and countries are also contributing to this research . From the citations analysis table, it can be concluded that the maximum number of citations is seen in the year 2019. Publications got few citations between 2010 and 2012. With the rising need to explore Cognitive Document Processing, researchers can investigate this area that was ignored earlier.

5. LIMITATIONS OF THE PRESENT STUDY

The present study considered the available Scopus database scientific publications only. Scientific documents from the other research databases such as Web of Science, Google Scholar, etc. are not considered for the current research analysis. The selection of the keywords used for querying the Scopus database is the author’s specific viewpoint. Different combinations of the keywords yield different results from the research databases. Keywords used for analysis can be re-arranged, modified, and updated as per the researcher’s perspective and need. The publication year range between 2010 and 2021 is considered for analyzing the results. Researchers may use other year ranges as a filter. English language scientific documents are considered for this research study. The secondary documents and patent data available in the Scopus database are not considered as part of the research study. Further study could be explored by incorporating those sets of documents as well.

6. CONCLUSION

This bibliometric study attempted to show the statistical data analysis on the retrieved Scopus database’s scientific documents that can be used by the upcoming researchers in Cognitive Document Processing. Network analysis diagrams depict the strong relatedness of the retrieved documents based on the co-occurrence of the keywords, author citations, bibliographic coupling, terms relationships from the text data of documents, etc. It is observed that before 2010 the research in Cognitive Document Processing was less explored but recently researchers are understanding the importance of automation in the business processes and thus this research area is gaining popularity ever since 5 years. India is considered as the upcoming country for Cognitive Document Processing research with around 40% contribution between 2010 and 2021 followed by the United States. There is much scope of the Cognitive Document Processing research publications in various journal sources since most of the publications in this area found are in the conference paper. The exposure to the study shows that research must be geared towards the subject areas that benefit society directly or indirectly.

7.REFERENCES

[1] E. Group, “Everest Group - Unstructured Data Process Automation,” 2019.

[2] S. Al-Shiakhli, “Big Data Analytics: A Literature Review Perspective,” *Luleå Univ. Technol.*, vol. 1, no. 1, pp. 1–57, 2019.

[3] W. Cmr and D. T. Ocr, “How and Why CMR is Disrupting Traditional OCR.”

[4] H. Sidhwa, S. Kulshrestha, S. Malhotra, and S. Virmani, “Text Extraction from Bills and Invoices,” *Proc. - IEEE 2018 Int. Conf. Adv. Comput. Commun. Control Networking, ICACCCN 2018*, pp. 564–568, 2018, DOI: 10.1109/ICACCCN.2018.8748309.

[5] C. Zhang, “Intelligent process automation in audit,” *J. Emerg. Technol. Account.*, vol. 16, no. 2, pp. 69–88, Sep. 2019, DOI: 10.2308/jeta-52653.

[6] R. Sindhgatta, A. H. M. Hofstede, and A. Ghose, *Resource-Based Adaptive Robotic Process Automation: Formal/Technical Paper*, vol. 12127 LNCS. 2020.

[7] Y. Ye *et al.*, “A unified scheme of text localization and structured data extraction for joint OCR and data mining,” in *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, 2019, pp. 2373–2382, DOI: 10.1109/BigData.2018.8622129.

[8] S. Desai and A. Singh, “Optical character recognition using template matching and back propagation algorithm,” *Proc. Int. Conf. Inven. Comput. Technol. ICICT 2016*, vol. 2016, 2016, DOI: 10.1109/INVENTIVE.2016.7830161.

[9] “Automating Receipt Digitization with OCR and Deep Learning.” https://nanonets.com/blog/receipt-ocr/ (accessed Sep. 21, 2020).

[10] “Document Digitization: Rethinking OCR with Machine Learning.” https://www.infoq.com/presentations/ocr-ml-doc-digitization/ (accessed Sep. 28, 2020).

[11] O. Doguc, “Robot process automation (RPA) and its future,” in *Handbook of Research on Strategic Fit and Design in Business Ecosystems*, IGI Global, 2019, pp. 469–492.

[12] “A Comprehensive Guide to OCR with RPA and Document Understanding.”

https://nanonets.com/blog/ocr-with-rpa-and-document-understanding-uipath/ (accessed Sep. 28, 2020).

[13] M. Kukreja, “Study of Robotic Process Automation (RPA),” *Int. J. Recent Innov. Trends Comput. Commun.*, no. June, pp. 434–437, 2016.

[14] N. Zhang and B. Liu, “The key factors affecting RPA-business alignment,” Jul. 2018, DOI: 10.1145/3265689.3265699.

[15] R. R. Patil and S. Kumar, “A Bibliometric Survey on the Diagnosis of Plant Leaf Diseases using Artificial Intelligence DigitalCommons @ the University of Nebraska - Lincoln A Bibliometric Survey on the Diagnosis of Plant Leaf Diseases using Artificial Intelligence Rutuja Rajendra Patil,” no. February 2020.

[16] A. Gokhale, P. Mulay, D. Pramod, and R. Kulkarni, “A Bibliometric Analysis of Digital Image Forensics,” *Sci. Technol. Libr.*, vol. 00, no. 00, pp. 1–18, 2020, DOI: 10.1080/0194262X.2020.1714529.

[17] N. S. Harinarayana, “Data Sources and Software Tools for Bibliometric Studies,” 21AD, [Online]. Available: https://epgp.inflibnet.ac.in/epgpdata/uploads/epgp_content/library_and_information_science/informetrics_&_scientometrics/data_sources_and_software_tools_for_bibliometric_studies/et/333_et_m2.pdf.

[18] K. D. Kadam, S. A. Ahirrao, and K. Kotecha, “DigitalCommons @ University of Nebraska - Lincoln Bibliometric Analysis of Passive Image Forgery Detection and Explainable AI Bibliometric Analysis of Passive Image Forgery Detection and,” no. January 2020.

[19] “(PDF) Application of author bibliographic coupling analysis and author keywords ranking in identifying research fronts of Indian Neurosciences research.” https://www.researchgate.net/publication/333104239_Application_of_author_bibliographic_coupling_analysis_and_author_keywords_ranking_in_identifying_research_fronts_of_Indian_Neurosciences_research (accessed Oct. 04, 2020).